

HIV Database Workshop

<https://hiv.lanl.gov>

seq-info@lanl.gov

Presenter: Brian Foley

Database PI: Brian Foley

Additional database staff: Werner Abfalterer, Katie Belobrajdic, Will Fischer, Elizabeth-Sharon Fung, Kumkum Ganguly, Jennifer Macke, James Szinger, Hyejin Yoon



Contract Officer Representative: Anjali Singh, NIAID, NIH



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



Workshop Topics

Day 1: HIV Sequence Database

General introduction

Sequence search interface – alignments and basic trees

Geography search interface

Database Alignments

Tool Examples:

- GeneCutter – proteins from nucleotide sequences (HIV, SIV)
- TreeMaker (Neighbor Joining) / IQTree (approximate ML)
- Sequence Locator tool: HIV, SIV, HCV, HFV
- QuickAlign: HIV, SIV, HCV, HFV
- Alignment Multitool
- Highlighter
- Hypermut
- Quality Control (HIV)

HIV Immunology Database Workshop

- **Day 2, Keystone 2022**

- **HIV Immunology Database**

- **Part 1:**

- HIV Immunology Database overview
- T cell epitopes – entries and searches
- Antibody Database – entries and searches
- Neutralizing Antibody Resources

- **Part 2:**

- Antibody Features Database
- Genome Browser
- CATNAP, both tailored for HIV and applicable to any pathogen

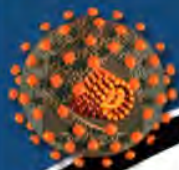
- **Part 3:**

- CombiNAber, applicable to any pathogen
- GenSig.

- **Part 4:**

- More computational tools for Immunologists, many applicable for any pathogen
- Vaccine design and evaluation tools, applicable to any pathogen

Entry page at <https://hiv.lanl.gov/>



HIV DATABASES

The HIV databases contain comprehensive data on HIV genetic sequences and immunological epitopes. The website also gives access to a large number of tools that can be used to analyze and visualize these data. This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Interagency Agreement No. AAI12007-001-00000. Our content is reviewed by an [Editorial Board](#).

[SEQUENCE DATABASE ▶](#) [IMMUNOLOGY DATABASE ▶](#)

[OTHER VIRUSES ▶](#)

News: [Archived News ▶](#)

[HIV Molecular Immunology 2020](#)
HIV Molecular Immunology 2020 is now available online. The PDF version is hypertext enabled and features clickable table-of-contents, indexes, references and links to external web sites. 27 January 2022


[2020 Alignments](#)
The 2020 Web, Filtered Web, Super Filtered Web, and Consensus Alignments are now available [online](#). The curated web alignments contain a full range of sequences available through the end of 2020. New consensus sequences are available, described by [Linchangco et al. 2022](#). 24 January 2022

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


Operated by Triad National Security, LLC for the U.S. Department of Energy's National Nuclear Security Administration
© Copyright Triad National Security, LLC. All Rights Reserved | [Disclaimer/Privacy](#)



DEPT OF HEALTH & HUMAN SERVICES



Los Alamos
NATIONAL LABORATORY



NATIONAL INSTITUTES OF HEALTH



Los Alamos
NATIONAL LABORATORY

- Sequence DB
- Immunology DB
- HXB2 Feature DB
- Env Feature DB
- Neutralization DB
- HCV DB
- HFV DB
- COVID-19 Genome Analysis Pipeline

HIV Sequence Database

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

[How to Cite this Database](#)

[Editorial Board](#)

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

[Genome Browser](#) uses jBrowse to display diverse data about the HIV-1 genome and proteome

[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

News:

[Archived News >](#)

[HIV Molecular Immunology 2020](#)

HIV Molecular Immunology 2020 is now available online. The PDF version is hypertext enabled and features clickable table-of-contents, indexes, references and links to external web sites. 27 January 2022

[2020 Alignments](#)

The 2020 *Web*, *Filtered Web*, *Super Filtered Web*, and *Consensus* Alignments are now available [online](#). The curated web alignments contain a full range of sequences available through the end of 2020. New consensus sequences are available, described by [Linchangco et al. 2022](#). 24 January 2022

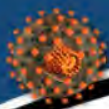
last modified: Tue Sep 7 15:54 2021

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

Operated by Triad National Security, LLC for the U.S. Department of Energy's National Nuclear Security Administration
© Copyright Triad National Security, LLC. All Rights Reserved | [Disclaimer/Privacy](#)



Each header on our pages is a pull-down menu of choices.



HIV sequence database

DATABASES **SEARCH** **ALIGNMENTS** **TOOLS** **PUBLICATIONS** **INFO**

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

[Genome Browser](#) uses jBrowse to display diverse data about the HIV-1 genome and proteome

[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

[How to Cite this Database](#)

[Editorial Board](#)

Multiple
paths to
most tools

News:

[Archived News](#) ▶

[HIV Molecular Immunology 2020](#)

HIV Molecular Immunology 2020 is now available online. The PDF version is hypertext enabled and features clickable table-of-contents, indexes, references and links to external web sites. 27 January 2022

[2020 Alignments](#)

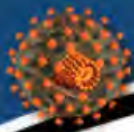
The 2020 *Web*, *Filtered Web*, *Super Filtered Web*, and *Consensus Alignments* are now available [online](#). The curated web alignments contain a full range of sequences available through the end of 2020. New consensus sequences are available, described by [Linchangco et al. 2022](#). 24 January 2022

last modified: Tue Sep 7 15:54 2021

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

Operated by Triad National Security, LLC for the U.S. Department of Energy's National Nuclear Security Administration
© Copyright Triad National Security, LLC. All Rights Reserved | [Disclaimer/Privacy](#)





HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS INFO

All kinds of basic information about HIV and about our database

Tutorials and Basic Information

Tutorials

[HIV Database presentations](#) from conference workshops

[FAQs](#) about the HIV Sequence Database

[FAQs](#) about the HIV Immunology Database

[Sequence quality control](#) explains several common problems with sets of viral sequences

[How to make a phylogenetic tree](#) explains how to build a phylogenetic tree

[HIV numbering](#) relative to reference strain HXB2

[SIV numbering](#) relative to reference strain SIVmm239

Articles

[Stalking the AIDS Virus \[PDF\]](#) article from 2003 LANL Research Quarterly about HIV Database research on the HIV-immune system interaction as a step toward an AIDS vaccine

[Novel approach leads to potential HIV-1 vaccine](#) 2018 LANL news release about the HIV mosaic vaccine

Reference Information

[Circulating recombinant HIV-1 sequences](#) details about all documented CRFs

[HIV-1 gene map](#) illustrates the organization of HIV-1, including HXB2 breakpoints

[HXB2 annotated spreadsheet \(.xls\)](#) provides a fully-annotated sequence of HXB2 with base-by-base detail

[HIV and SIV subtype nomenclature](#) gives an overview of HIV and SIV subtype nomenclature, particularly HIV-1 groups and subtypes

[Primate immunodeficiency virus nomenclature](#) lists SIV species and nomenclature

[How the HIV database classifies sequences](#) explains how recombinants are named and annotated

[Common sequence formats for alignments](#) shows examples of common sequence formats for alignments

[How to cite this Database](#) explains how to cite this website and the printed HIV compendia

[Codes and symbols in sequences](#) decodes the symbols and IUPAC codes that appear in sequences and alignments

[Codon table](#) gives the translation of nucleotides into amino acids

[Links](#) HIV/AIDS resources and bioinformatics tools on other websites

Previous workshop presentations

last modified: Wed Feb 23 11:34 2022

Questions or comments? Contact us at

seq-info@lanl.gov

Yes! We do respond to this e-mail address!

Search Interface

■ Results (what you want)

Can download aligned or unaligned sequences

Alignments based on multiple pairwise alignments – alignments are good, but need re-alignment (GeneCutter for example) for an optimal alignment

Select all or a subset of sequences for download (one per patient for example)

Sequences can be re-ordered by clicking on fields at the top of the page, and names customized

■ Searches (how you get it)

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 (what you can do with it)

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

■ Help (if all else fails, read the instructions!)

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

Today's Sequence Search example workflows

- Assemble a country-wide whole-genome data set:
 - Get all available complete genome sequences from a given country (Brazil)
 - Add in subtype reference sequences and make a phylogenetic tree for quick evaluation
 - Download the sequences as a phylogenetically sorted alignment; *look at the alignment!*
 - Clean up alignment and extract spliced coding sequences (GeneCutter); *look at the alignment(s) again!*
- Other approaches (search and evaluation)
 - Geography Search interface
 - Advanced Search (ask us afterwards!)

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: 2022-02-20

[Advanced Search](#)

Sequence Information

[Upload accession file](#) No file selected.

[Accession number](#)

[Sequence name](#)

[Sequence length](#)

[exact](#) ☒ [Sampling year](#)

[Sampling country](#)

[Virus](#)

[Subtype](#)
No subtype
A
A1
A2
A3

☐ Include [recombinants](#)

☐ More sequence information

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

[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

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

☐ Publication Information

☐ Patient Information

☐ Geographical Information

☐ Amino Acid Motif Search

☐ Output

☐ Include [problematic](#) sequences

[% of non-ACGT](#)

List records per page

Show results selected ☐

[Advanced Search](#)

We will search
for country =
Brazil (BR)

We will search
for complete
genomes.

Results for HIV-1 complete genomes from Brazil

Displaying 1 - 100 of 435 sequences found:

Note: 17 **problematic** sequences were removed from this result.

record
 to
 100 records per page

Click on field name to sort in ascending or descending order

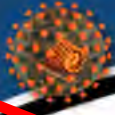
#	Select	Patient Code (id)	Accession Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism	
1	<input checked="" type="checkbox"/>	Blast BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	<input type="checkbox"/>	Blast BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3	<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_BF1	BRAZIL	1999		8767	HIV-1
12	<input checked="" type="checkbox"/>	Blast BR97(10885)	AY455779	94BR_RJ_97	BF1	BRAZIL	1994		8962	HIV-1
13	<input checked="" type="checkbox"/>	Blast RJ2(10886)	AY455780	99UFRJ_2	BF1	BRAZIL	1999		9045	HIV-1
14	<input checked="" type="checkbox"/>	Blast BR41(15452)	AY455781	94BR_RJ_41	BF1	BRAZIL	1994		8864	HIV-1
15	<input checked="" type="checkbox"/>	Blast RJ16(10887)	AY455782	99UFRJ_16	BF1	BRAZIL	1999		9002	HIV-1
16	<input checked="" type="checkbox"/>	Blast RJ9(10888)	AY455783	99UFRJ_9	BF1	BRAZIL	1999		9040	HIV-1
17	<input checked="" type="checkbox"/>	Blast BR59(10884)	AY455784	94BR_RJ_59	BF1	BRAZIL	1994		8898	HIV-1
18	<input checked="" type="checkbox"/>	Blast BR58(10883)	AY455785	94UFRJ_58	BF1	BRAZIL	1994		8898	HIV-1
19	<input checked="" type="checkbox"/>	Blast	AY727522	04BR012	C	BRAZIL	2004		9050	HIV-1

"One sequence
 per patient"
 removes very
 similar
 sequences
 (available if a
 region is
 selected)

For real use, download background data

Select a few sequences and make a tree.

We can add a reference set to our data and align them all together.



HIV sequence database

DATABASES

SEARCH

ALIGNMENTS

TOOLS

PUBLICATIONS

INFO

search site

Search

Make Tree

Download Sequences

Save Background Info

Make Histogram

Geography

Clear

Displaying 1 - 100 of 435 sequences found:

Note: 17 problematic sequences were removed from this result.

Select all

Unselect all

Invert selection

Show all

One sequence/patient


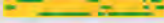

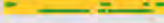





















Select

 record to

List

 100 records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code	Accession Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism	
		(id)								
1	<input type="checkbox"/>	Blast BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	<input checked="" 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 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 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_BF1	BRAZIL	1999		8767	HIV-1
12	<input type="checkbox"/>	Blast BR97(10885)	AY455779	94BR_RJ_97	BF1	BRAZIL	1994		8962	HIV-1
13	<input checked="" type="checkbox"/>	Blast RJ2(10886)	AY455780	99UFRJ_2	BF1	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	BF1	BRAZIL	1999		9002	HIV-1
16	<input type="checkbox"/>	Blast RJ9(10888)	AY455783	99UFRJ_9	BF1	BRAZIL	1999		9040	HIV-1
17	<input checked="" type="checkbox"/>	Blast BR59(10884)	AY455784	94BR_RJ_59	BF1	BRAZIL	1994		8898	HIV-1
18	<input type="checkbox"/>	Blast BR58(10883)	AY455785	94UFRJ_58	BF1	BRAZIL	1994		8898	HIV-1
19	<input type="checkbox"/>	Blast	AY727522	04BR013	C	BRAZIL	2004		9050	HIV-1
20	<input checked="" type="checkbox"/>	Blast	AY727523	04BR021	C	BRAZIL	2004		8958	HIV-1
21	<input type="checkbox"/>	Blast	AY727524	04BR038	C	BRAZIL	2004		9042	HIV-1
22	<input checked="" type="checkbox"/>	Blast	AY727525	04BR073	C	BRAZIL	2004		8997	HIV-1
23	<input type="checkbox"/>	Blast	AY727526	04BR137	31_BC	BRAZIL	2004		8795	HIV-1
24	<input type="checkbox"/>	Blast	AY727527	04BR142	31_BC	BRAZIL	2004		9057	HIV-1
25	<input type="checkbox"/>	Blast 107(10943)	AY771528	RRFPM107	RF1	BRAZIL	1999		8798	HIV-1

TreeMaker tool

Choice of outgroup influences the presentation of the tree.

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

Alternatively, leave blank for midpoint rooting

Optional mailback, and tree title

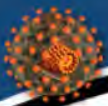
The screenshot shows the TreeMaker tool interface for the HIV sequence database. The top navigation bar includes links for DATABASES, SEARCH, ALIGNMENTS, TOOLS, PUBLICATIONS, and INFO, along with a search site input and a Search button. The main form is divided into several sections:

- Model parameters:** Includes a Distance model dropdown set to "General-time-reversible (GTR)", Gap handling radio buttons for "strip gaps before analysis" and "treat as missing" (the latter is selected), and Site rates radio buttons for "Equal" and "Gamma Shape" (the latter is selected with a value of 0.5).
- Reference sequences (GENOME):** Includes radio buttons for "All", "A-L" (selected), "A-L + CRFs", "N, O, P, CPZ", and "Menu select only". A list of reference sequences is displayed below.
- Outgroup:** Includes a list of reference sequences and a list of database sequences. A red arrow points to the "Reference sequences" list, and another red arrow points to the "Database sequences" list.
- Results link:** Includes a text input for "Email a link to the results to this address" (btf@lanl.gov) and a text input for "with job title" (FewBrazilGenomes).

Buttons for "Submit" and "Reset" are located at the bottom left of the form.

These settings may change relative branch lengths somewhat, but rarely alter the tree topology.

Our Brazilian sequences



HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS INFO

Download Your Tree Results

This tree contains 62 sequences and is 11382 characters long, including insertions.

View tree:

- Phenogram: [pdf](#) [png](#) [ATV](#) (a Java-based phylogenetic tree viewer)
- Radial (unrooted): [pdf](#) [png](#)
- Newick

Analyze tree:

- [Find a root by TreeRate](#)
- [See branch length](#)
- [Create a color-coded tree using Rainbow Tree](#)
- [Examine topology using TreeTopo](#)

Download alignment used for tree building

- [Fasta alignment \(before gapstripping\)](#)
- [Fasta alignment in tree order \(before gapstripping\)](#)

Job ID:

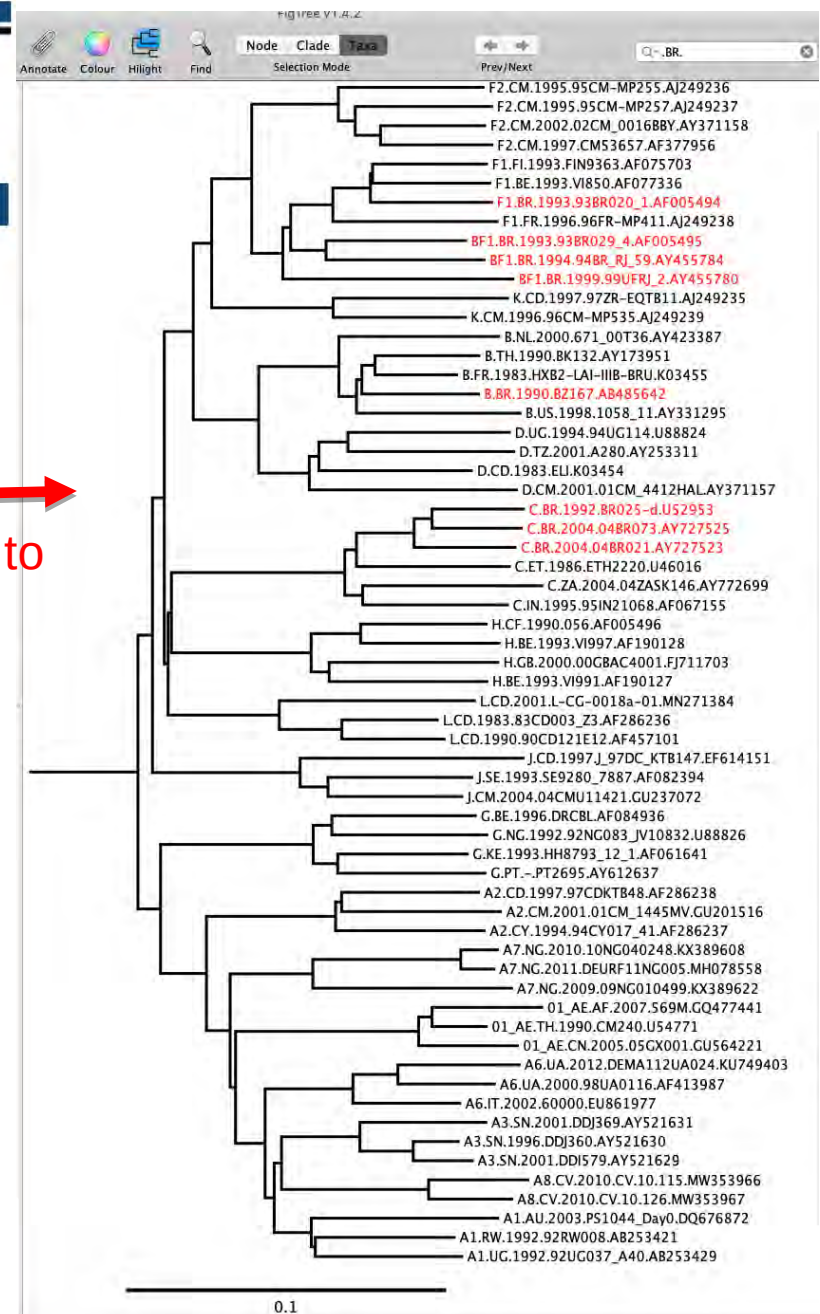
MAP/3108267c

Notice: An email with a link to the results has been sent to btbf@lanl.gov

last modified: Tue Oct 26 08:42 2021

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

Use FigTree or other programs to view Newick tree files.



Save alignment, use BioEdit or Aliview to view/adjust (but see Align Multi-tool).

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



HIV sequence database

Save alignment, use BioEdit, Aliview, or SeAl to view.

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

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)

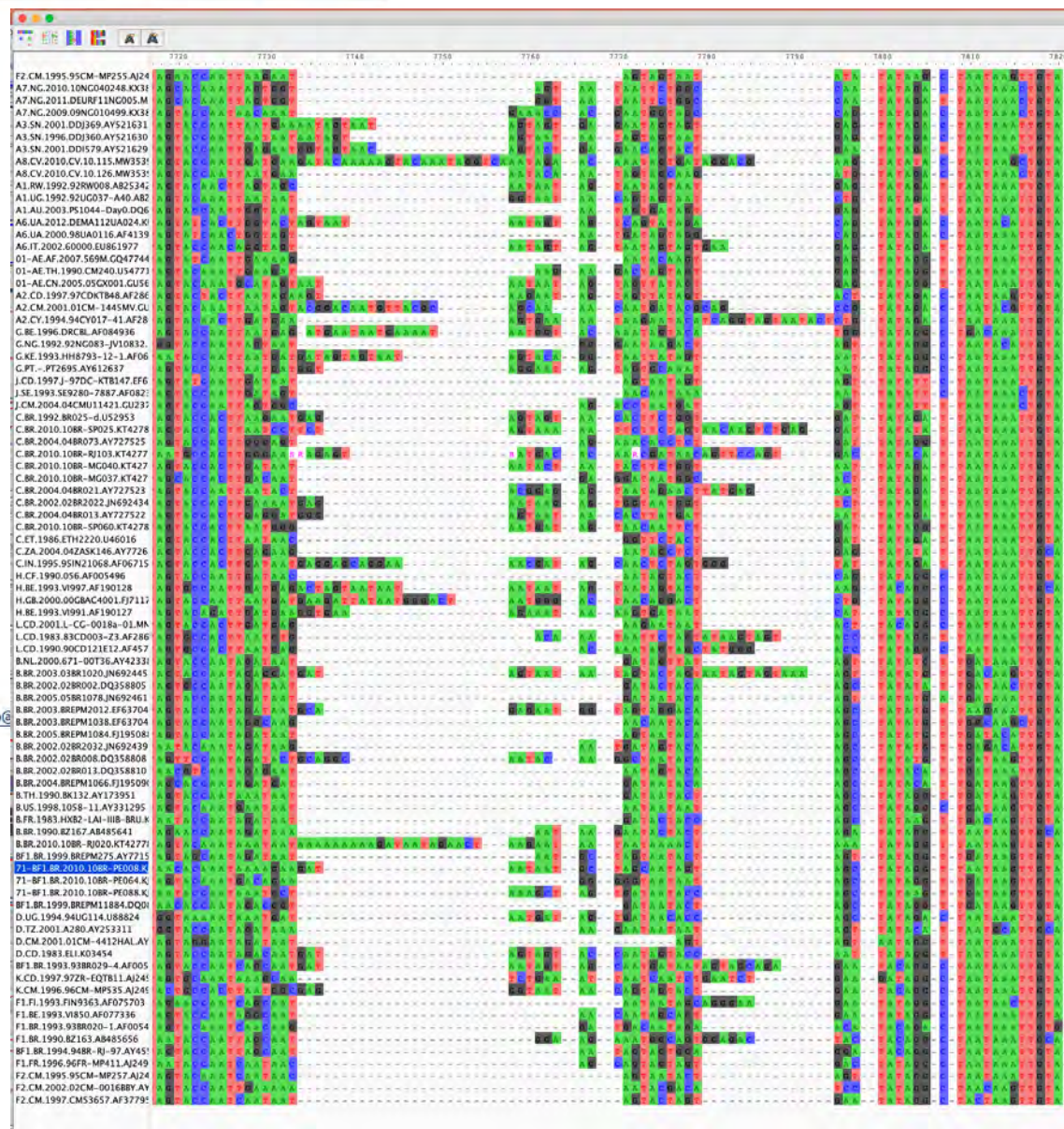
Alignment used for tree building

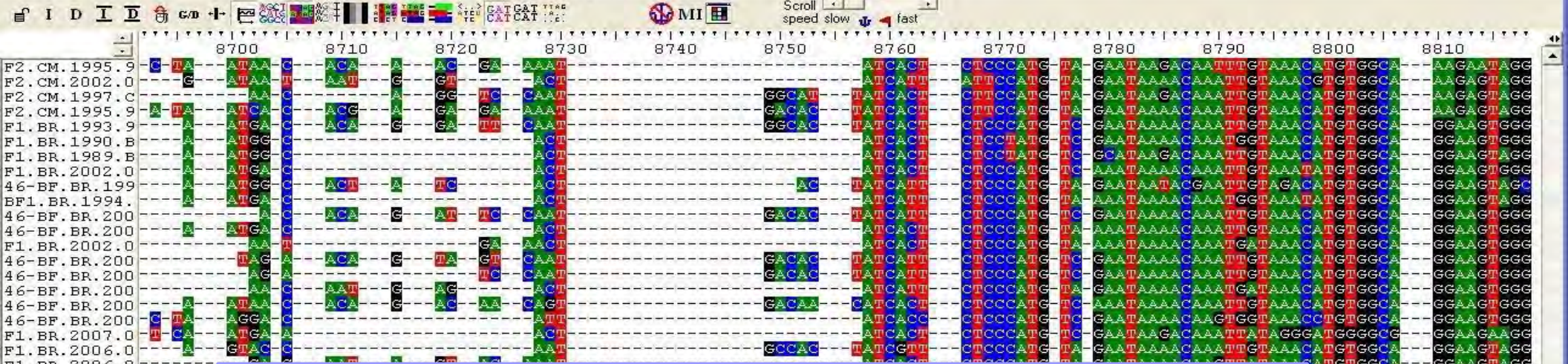
- Download fasta alignment (before gapstripping)
- Download fasta alignment in tree order (before gapstripping)
- Download fasta alignment (after gapstripping)
- Download Newick Tree File

last modified: Thu May 7 07:35:00 2009

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

Brazil Genomes Plus Subtype Reference Set

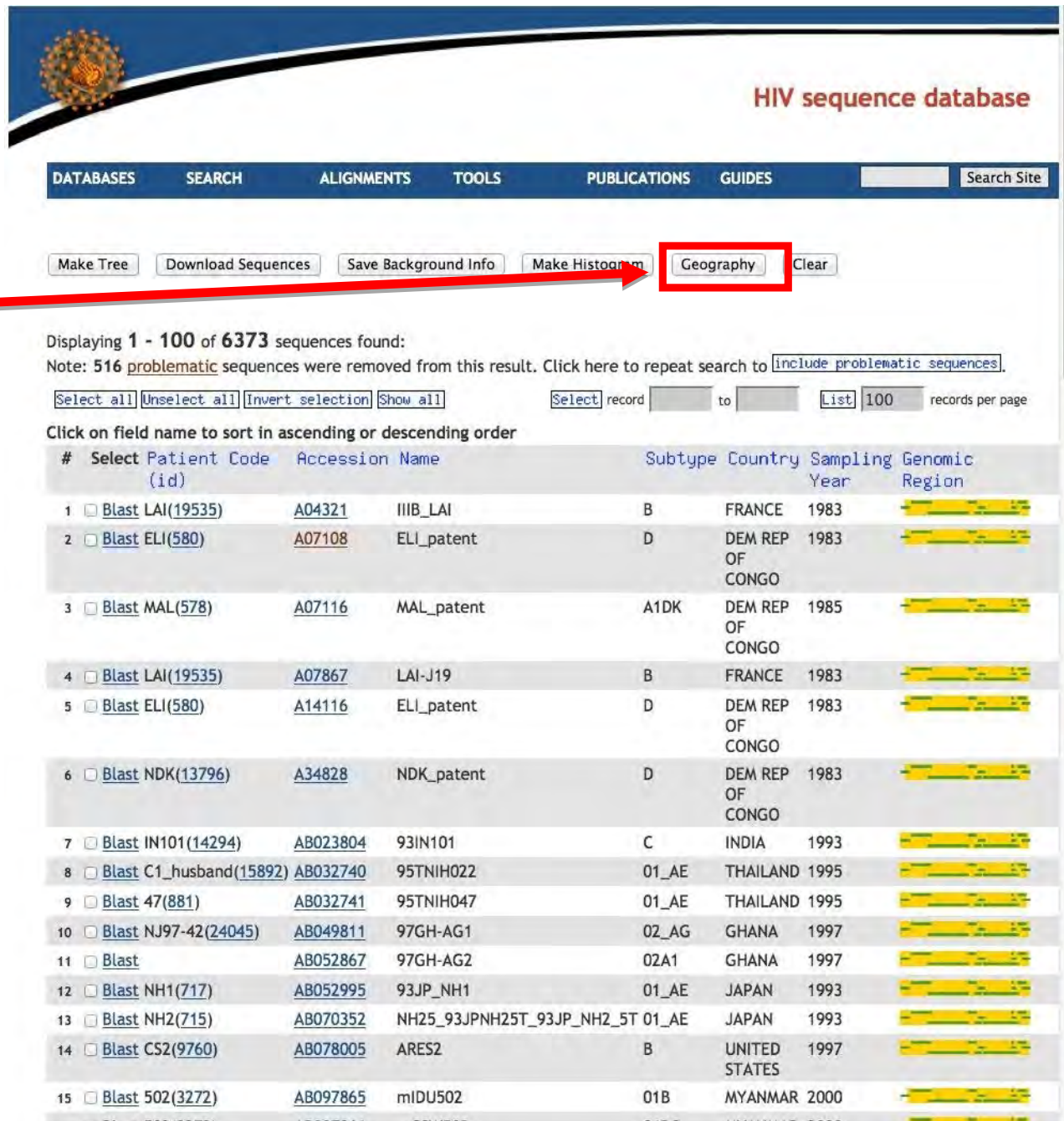




Quick Alignment from Search Interface has many "broken codons"

GeneCutter tool aligns by intact codon
(so sequences are "codon-aligned")

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 6373 sequences found:
Note: 516 problematic sequences were removed from this result. Click here to repeat search to [include problematic sequences](#).

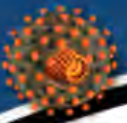
Select all Unselect all Invert selection Show all Select record to List 100 records per page

Click on field name to sort in ascending or descending order

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

New search: all sequences from Brazil.

Then look at the distribution of the sequences over the genome


















 HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS INFO

Displaying 1 - 100 of 27937 sequences found:
Note: 9720 problematic sequences were removed from this result.

record to 100 records per page

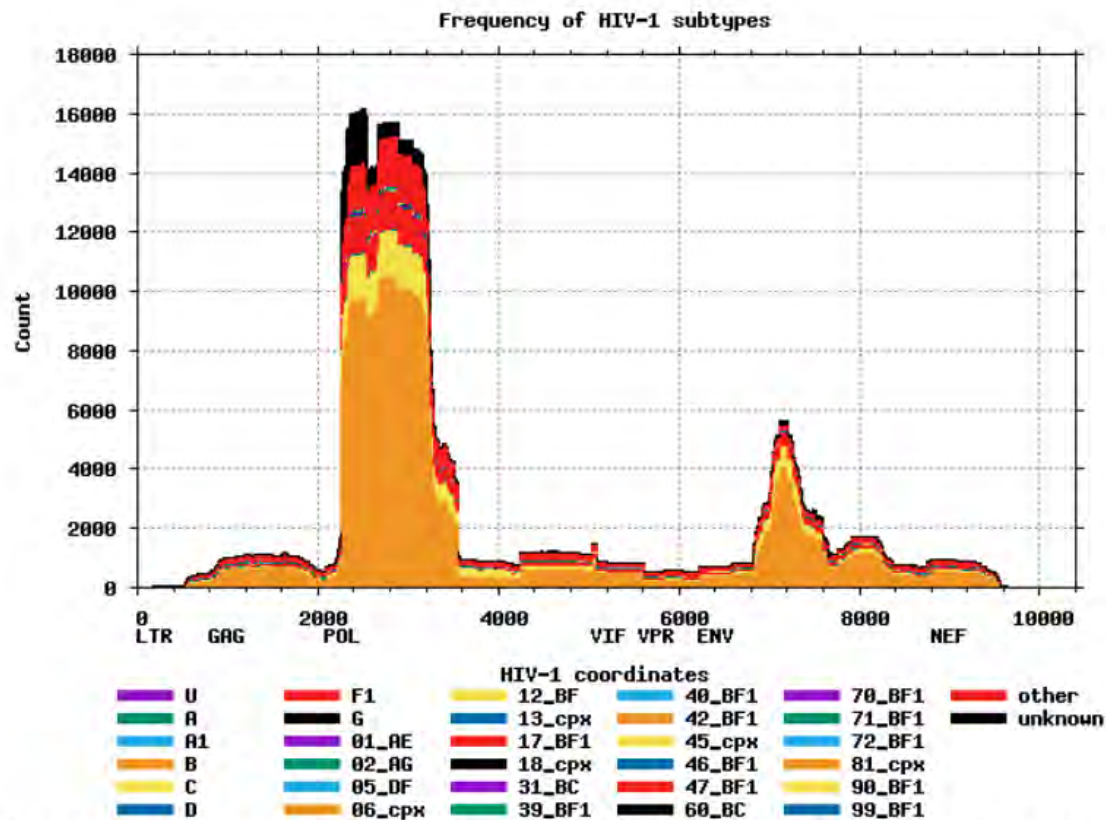
Click on field name to sort in ascending or descending order

#	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
15	<input type="checkbox"/> Blast	BR023(13877)	AF009376	92BR023	BC	BRAZIL	1992		1176	HIV-1
16	<input type="checkbox"/> Blast	BR024(659)	AF009377	92BR024	B	BRAZIL	1992		1176	HIV-1
17	<input type="checkbox"/> Blast	BR025(586)	AF009378	92BR025	B	BRAZIL	1992		1176	HIV-1

Histogram output

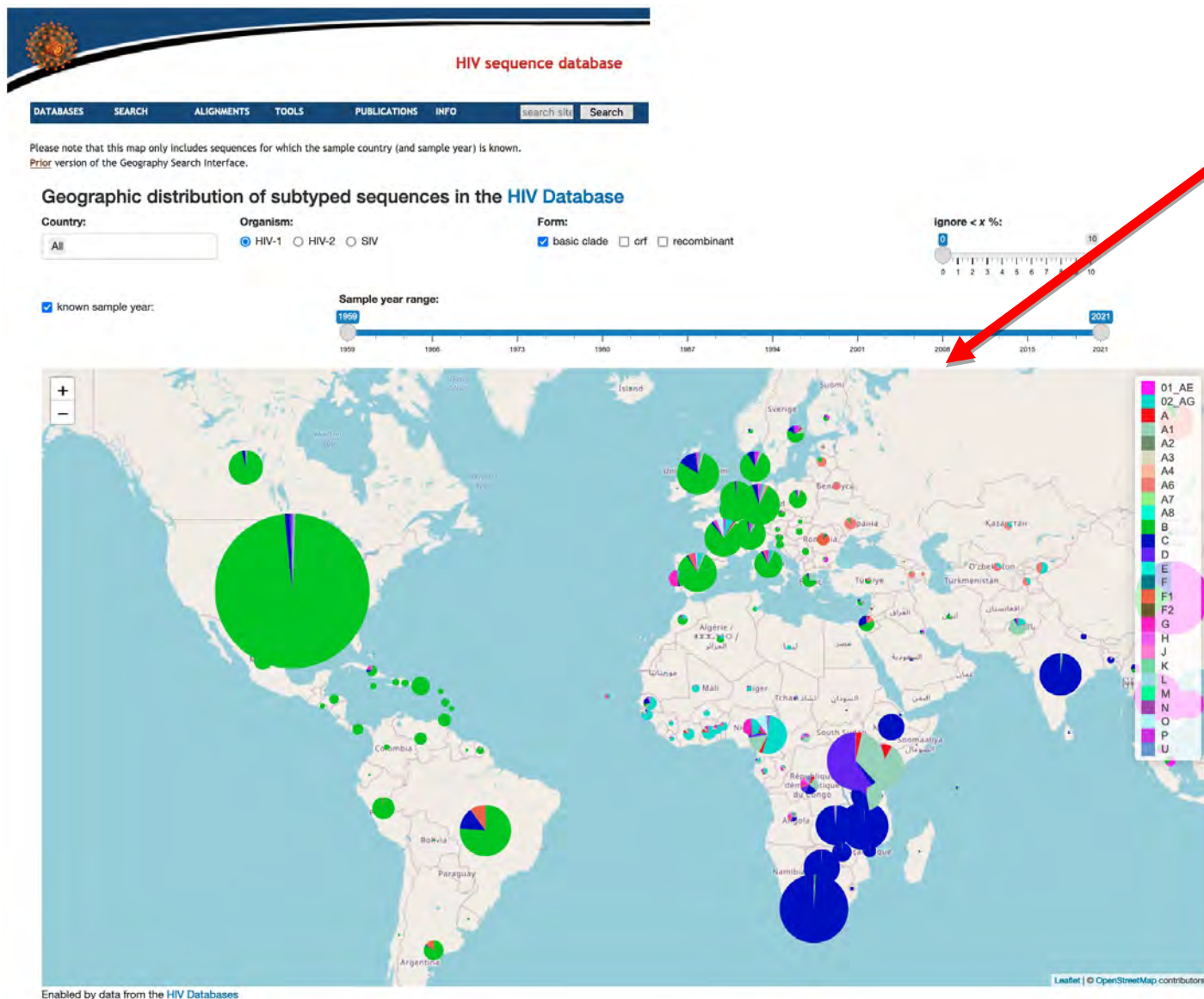


Histogram result for your query:



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.

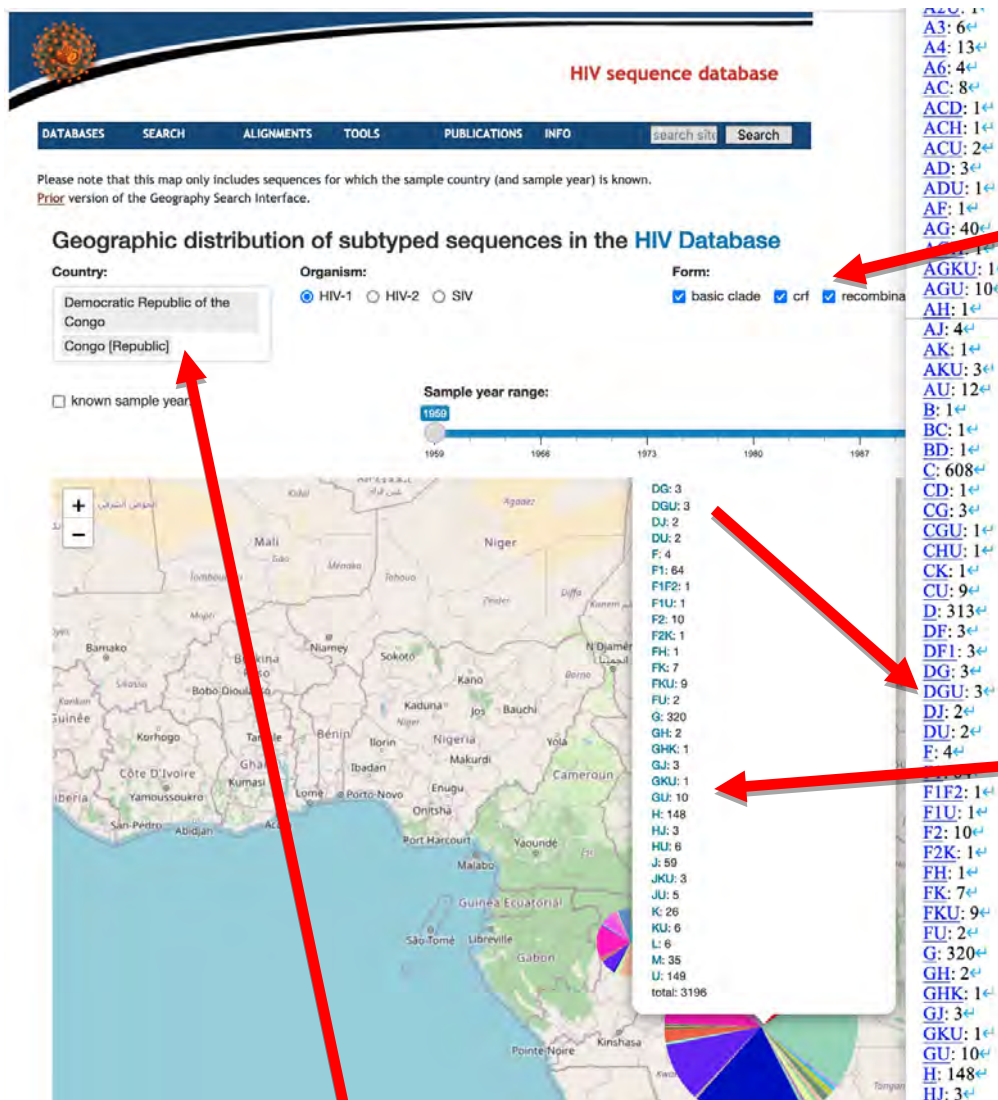
Geography output



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

More sequences in the HIV database are from USA than from South Africa. But South Africa has more infected people. Beware of this type of sampling bias.

Geography output



You must select these check-boxes if you are interested in the recombinants.

Data from the page can be copy-pasted to a text editor, and the links to the sequence data are still included.

You must delete the “all” from this list, when selecting one or a few countries.

Pre-Built Sequence alignments

- Based on both manual and HMM alignments
- Manually curated
- Alignments are in reading frame (codon aligned)
- Contain non-redundant data (one sequence per patient)
- Compendium alignments show a small “readable” subset
- Reference alignments contain up to four representatives of each subtype (CRFs optional).
 - Useful to provide context for newly generated sequences!
- Protein alignments with frameshifts compensated
- Subtype consensus and “maximum likelihood ancestors” are available for reagent production
- Special interest alignments
 - Sequence sets (“authors’ alignments”) of particular research interest
 - Suggestions and additions welcome!

HIV Sequence Alignments

- **Web Alignments** are nucleotide and protein alignments that represent the fullest spectrum of sequences in the database.
- **Filtered Web Alignments** are a filtered subset of sequences from the web alignments. These alignments are cleaner, but contain slightly less information.
- **Subtype Reference Alignments** contain approximately 4 representatives of each subtype.
- **Compendium Alignments** are the subset of sequences printed in the [HIV Sequence Compendium](#).
- **Consensus/Ancestral Sequences** include a consensus for each subtype, an M-group consensus-of-consensuses, and some ancestral sequences.
- **RIP Alignment** contains a consensus for each subtype and reference sequences for all groups, subtypes, and CRFs.

Before use, please read the additional information below.

Options

Alignment type:

Organism:

Region:

Subtype:

DNA/Protein:

Year:

Format:

Get Alignment Reset

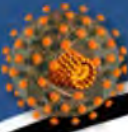
All (complete) = one per patient, all sequences for which we have a complete genome, or a complete gene.

Subtype Reference = 4 representatives of each subtype, plus one of each Circulating-inter-subtype-Recombinant-Form (CRF) of the M group, plus 4 O group, N group, P group, and SIV-CPZ

Consensus/Ancestral computed from master alignment periodically (at this point in the pandemic there is little year-to-year change).

Explanations of the content of the different alignments are shown lower on the webpage.

The HIV database sequence analysis tool set



HIV sequence database

DATABASES	SEARCH	ALIGNMENTS	TOOLS	PUBLICATIONS	GUIDES	Search Site
Programs and Tools <p>Search Interface retrieves HIV and SIV sequences, which can be aligned and used to build trees</p> <p>Geography Search Interface retrieves HIV sequences based on geographical distribution</p> <p>Genome Browser uses jBrowse to display diverse data about HIV-1 genome and proteome</p> <p>Tools for working with sequences lists all our online tools, by function</p>	Alignments <p>HIV Premade Alignments includes Consensus and Ancestral Sequences, Subtype Reference Alignments, and Complete Alignments</p>	HIV S	Index of all tools	Heatmap	Protein Feature Accent	
			Alignment Slicer	Hepitope	Quality Control	
			AnalyzeAlign	Highlighter	QuickAlign	
			AnnotateTree	HIV BLAST	Rainbow Tree	
			Branchlength	HIVAlign	Recombinant HIV-1 Drawing Tool	
			CATNAP	Hypermute	RIP	
			Codon Alignment	IQ-TREE	SeqPublish	
			CombiNAbert	jpHMM at GOBICS	Sequence Locator	
			Consensus Maker	Mosaic Vaccine Tool Suite	SNAP	
			ELF	Motif Scan	SUDI Subtyping	
ElimDupes	N-Glycosite	SynchAlign				
Entropy	PCOORD	Translate				
Epigraph	PepMap	TreeMaker				
FindModel	PeptGen	TreeRate				
Format Converter	PhyloPlace	Variable Region Characteristics				
Gap Strip/Squeeze	PhyML	VESPA				
GenBank Entry Generation	Pixel	External Tools				
Gene Cutter	Poisson-Fitter					
Genome Browser	PrimerDesign-M					

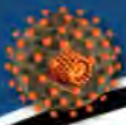
Click top level to link to full page of tools

News:

[IQ-TREE interface](#)

IQ-tree is a fast and effective stochastic algorithm for finding ML trees. We have developed a convenient web server for building trees with this method. A nice feature of this method is the ability to output a table of site-specific rates of evolution for each position in the alignment. 18 September 2017

[Archived News](#)



HIV sequence database

[DATABASES](#)[SEARCH](#)[ALIGNMENTS](#)[TOOLS](#)[PUBLICATIONS](#)[INFO](#)

HIV Database Tools

- Tools specific for HIV/SIV
- General use tools with some HIV/SIV-specific features
- General use tools

Analysis and Quality Control

- [Entropy](#) quantifies positional variation in an alignment using Shannon Entropy
- [GenSig](#) identifies genetic signatures. It can work on any phenotype file given in conjunction with a codon-aligned nucleotide alignment of a protein coding region
- [Glycan Shield Mapping](#) shows mapping absent hole-causing potential N-linked glycosylation sites (PNGS) on predicted glycan shields for an ENV sequence
- [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 for quick troubleshooting of HIV-1 sequences; optional step prepares sequence submission for GenBank
- [RAPR](#) (Recombinant Analysis PRogram) uses the Wald-Wolfowitz Runs Test to check for recombination in every triplet in the alignment.
- [RIP](#) (Recombinant Identification Program) detects HIV-1 subtypes and recombination
- [SNAP](#) calculates synonymous/non-synonymous substitution rates

Phylogenetics

- [AnnotateTree](#) creates a colored and weighted phylogenetic tree
- [Branchlength](#) calculates branch lengths between internal and end nodes; now included in the [TreeRate](#) tool
- [FindModel](#) finds which evolutionary model best fits your sequences
- [IQ-TREE](#) is a fast and effective stochastic algorithm for finding Maximum Likelihood trees, including site-specific rates of evolution at each alignment position
- [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
- [Rainbow Tree](#) Color code phylogenetic tree branches according to labels in the sequence names
- [TreeMaker](#) generates a Neighbor Joining phylogenetic tree
- [TreeRate](#) finds the phylogenetic root of a tree and calculates branch lengths and evolutionary rate

Immunology

- [CATNAP](#) (Compile, Analyze, Tally NAb Panels) provides meta-analysis of published neutralization panel data
- [CombiNAb](#) predicts and analyzes combination antibody neutralization scores using IC₅₀ and/or IC₉₅ for individual antibodies

Color code squares indicate whether the tool is only for HIV/SIV or not.

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.

More below!

[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 list a selection of external tools of significance in HIV informatics.

Many of these tools are essential, such as either BioEdit or Aliview for alignment viewing and correction.

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

Tools (a selection)

■ Analysis and Quality Control

Entropy identifies regions of proteins that are more conserved, or less conserved.

Hypermut identifies genomic regions affected by APOBEC-induced hypermutation.

Quality Control performs HyperMut, RIP subtyping, Treemaker, GeneCutter, etc...

N-Glycosite finds potential N-linked glycosylation sites.

RIP (Recombinant Identification Program) detects HIV-1 subtypes and recombination.

AnalyzeAlign in depth analysis of epitopes, continuous or discontinuous.

Variable Region Characterization unique tool for unaligned/unalignable V-regions

■ Alignment and sequence manipulation

Gene Cutter and **HIValign** align your sequences and extract protein-coding reading frames.

Align Multi-Tool simplifies a broad range of alignment processing

■ Phylogenetics

TreeMaker generates a neighbor-joining phylogenetic tree.

PhyML generates a maximum likelihood phylogenetic tree.

IQ-TREE generates a fast approximate maximum likelihood phylogenetic tree.

TreeRate finds the phylogenetic root of a tree and calculates evolutionary rate.

Rainbow Tree Adds colors and symbols to trees.

AnnotateTree maps quantitative information to branch weights and colors.

■ Format and display

Highlighter highlights differences within an alignment of nucleotide sequences.

Pixel makes compact images of large alignments.

Recombinant HIV drawing tool makes graphical representations of recombinant genomes

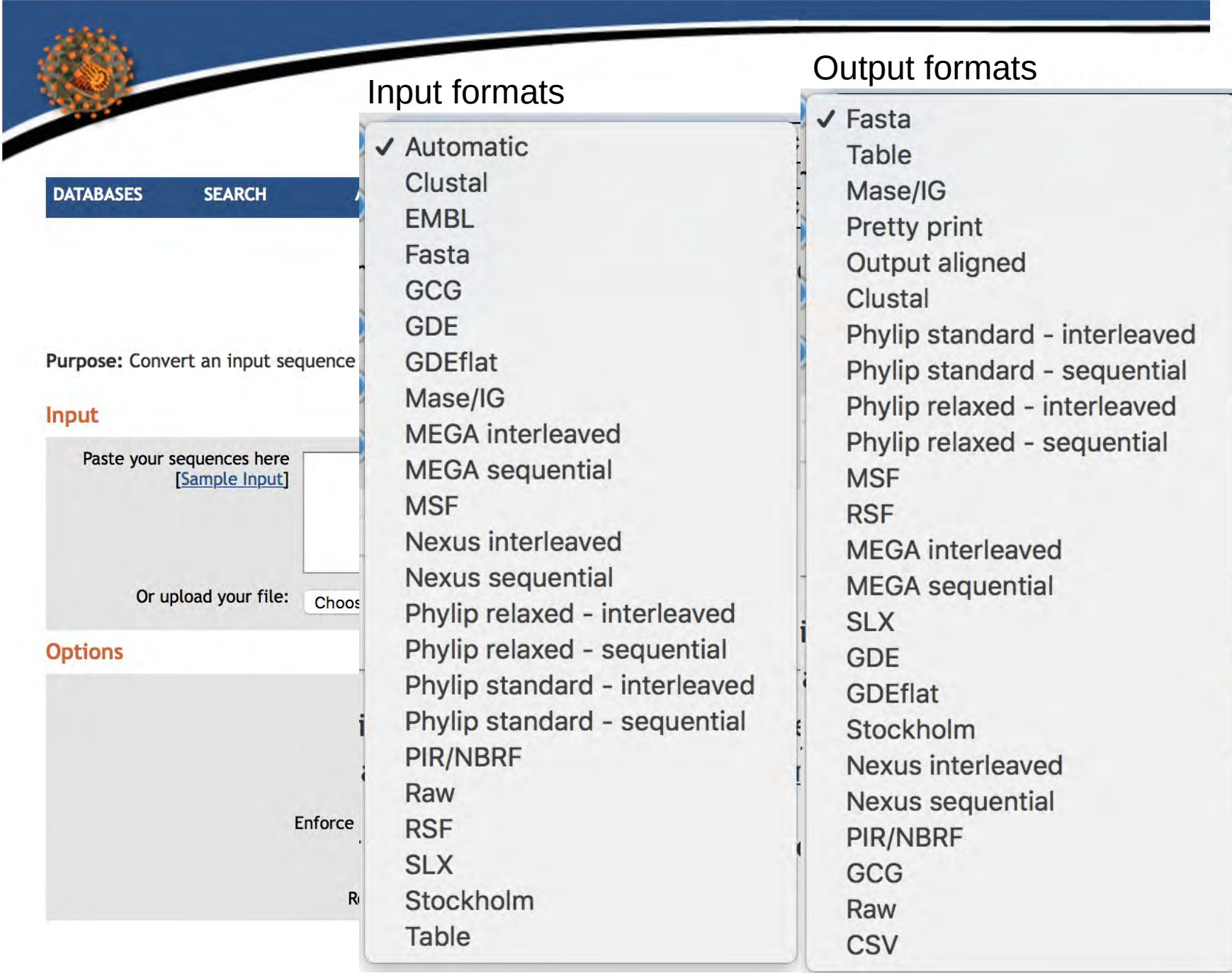
Genome Browser shows structural and immunological features of HIV

Format Converter transforms sequence data file formats

Format Converter

File formats:
"the secret
shame of
bioinformatics"

This tool helps
make life
better, easing
handling of
legacy data
and
transferring
data between
applications.



The screenshot displays the 'Format Converter' web interface. At the top left is a logo featuring a stylized orange and yellow sphere. Below it, a navigation bar contains 'DATABASES' and 'SEARCH' links. The main heading is 'Purpose: Convert an input sequence'. Under the 'Input' section, there is a text area for 'Paste your sequences here' with a '[Sample Input]' link, and a file upload option 'Or upload your file:' with a 'Choose' button. The 'Options' section is partially visible. Two dropdown menus are open: 'Input formats' and 'Output formats'. The 'Input formats' list includes: Automatic (checked), Clustal, EMBL, Fasta, GCG, GDE, GDEflat, Mase/IG, MEGA interleaved, MEGA sequential, MSF, Nexus interleaved, Nexus sequential, Phylip relaxed - interleaved, Phylip relaxed - sequential, Phylip standard - interleaved, Phylip standard - sequential, PIR/NBRF, Raw, RSF, SLX, Stockholm, and Table. The 'Output formats' list includes: Fasta (checked), Table, Mase/IG, Pretty print, Output aligned, Clustal, Phylip standard - interleaved, Phylip standard - sequential, Phylip relaxed - interleaved, Phylip relaxed - sequential, MSF, RSF, MEGA interleaved, MEGA sequential, SLX, GDE, GDEflat, Stockholm, Nexus interleaved, Nexus sequential, PIR/NBRF, GCG, Raw, and CSV.

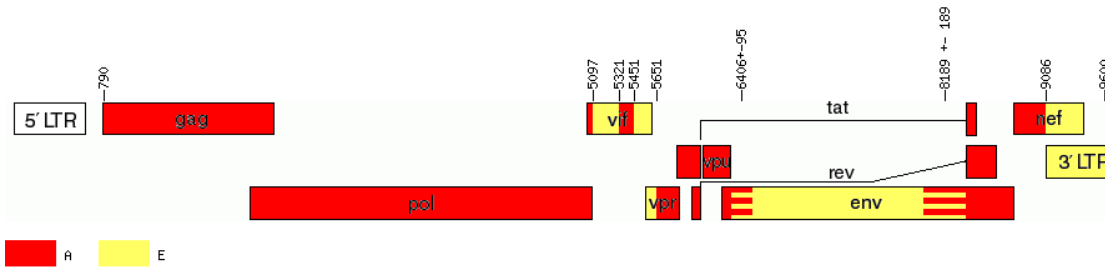
Input formats

- ✓ Automatic
- Clustal
- EMBL
- Fasta
- GCG
- GDE
- GDEflat
- Mase/IG
- MEGA interleaved
- MEGA sequential
- MSF
- Nexus interleaved
- Nexus sequential
- Phylip relaxed - interleaved
- Phylip relaxed - sequential
- Phylip standard - interleaved
- Phylip standard - sequential
- PIR/NBRF
- Raw
- RSF
- SLX
- Stockholm
- Table

Output formats

- ✓ Fasta
- Table
- Mase/IG
- Pretty print
- Output aligned
- Clustal
- Phylip standard - interleaved
- Phylip standard - sequential
- Phylip relaxed - interleaved
- Phylip relaxed - sequential
- MSF
- RSF
- MEGA interleaved
- MEGA sequential
- SLX
- GDE
- GDEflat
- Stockholm
- Nexus interleaved
- Nexus sequential
- PIR/NBRF
- GCG
- Raw
- CSV

Common problems with HIV sequences



- Genome structure
 - multiple overlapping reading frames, splicing, frame-shifting, error-prone replication
- Viral biology
 - hypermutation
 - recombination
- Laboratory/Data processing
 - contamination
 - location identification
 - fragile sequence identifiers

(partial) solutions

- Alignment / reading frame tools
 - Gene Cutter and HIValign align your sequences and chop out proteins.
 - Pixel makes compact images of large alignments.
 - Entropy identifies regions of proteins that are more conserved, or less conserved.
- Hypermution/recombination/contamination tools
 - Hypermot identifies sequences that have been hyper-mutated by APOBEC-3G or other restriction factors.
 - Highlighter reveals discordance within groups of putatively related sequences.
 - RIP (Recombinant Identification Program) detects recombination between HIV-1 subtypes.
 - TreeMaker generates a neighbor-joining phylogenetic tree.
 - Quality Control performs HyperMut, RIP subtyping, Treemaker, GeneCutter, etc...
- Location identification
 - Sequence locator unambiguously locates DNA and amino-acid sequences relative to a standard reference.
 - GenBank Entry Generation does what it says
- Sequence identifiers
 - Search interface simplifies generation of uniform sequence names with useful information.
 - Align Multi-tool allows re-annotation of sequence names

Gene Cutter

Unconventional Alignment/Homology program specialized for HIV

- copes with indels (“dead” viruses), IUPAC ambiguities, overlapping (multi-frame) coding sequences, “unalignable” variable regions
- produces DNA alignments by codon, as well as amino-acid alignments for multiple genes
- Aligns to reference sequence (HXB2 or SIV-Mac239) via HMMer
- Splits sequences into genes, and translates each gene to protein

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 (frame-shifted)

Including HXB2 as a reference may improve results

Does *NOT* address hypermutation or recombination

(see “Hypermut” or “RIP”)

Gene Cutter

Gene Cutter: Sequence Alignment and Protein Extraction

Purpose: Gene Cutter is a sequence alignment and protein extraction tool. It can be used for any set of nucleotide sequences for HIV-1, HIV-2 or SIV.

Gene Cutter can:

- align your nucleotide sequences (if they aren't already aligned)
- clip pre-defined coding regions from a nucleotide alignment
- codon-align the coding regions
- generate nucleotide and protein alignments of the cut regions

Details: The reference sequence used by this tool is [HXB2\(Accession #K03455\)](#) for HIV-1 or [SMM239\(Accession #M33262\)](#) for HIV-2 or SIV. Gene coordinates are based on these reference sequences. This version of Gene Cutter doesn't require a reference sequence to be included in your input nucleotide alignment. Gene Cutter will also accept **unaligned sequence sets**. Gene Cutter uses Hmmer with a training set of the full-length genome alignment and will give a better multiple alignment than many computationally-based alignment programs. Misalignments at the ends of a coding region may result in a few amino acids/bases not appearing in the output for that coding region.

In some sequences, an insertion will be compensated within a short distance by a deletion, or vice versa. As these frameshifts may not inactivate the protein, if a compensating mutation is within 5 amino acids of an initial frameshift, the shifted reading frame is left intact. Otherwise, the frame shift is marked with the hash symbol (#), and the translation is continued in the correct reading frame beyond the offending codon. Stop codons are marked by a dollar sign (\$).

The **best results** will be obtained if you submit an alignment that has been hand-aligned and contains the correct reference sequence. For more information, see [Gene Cutter Explanation](#).

Input

Select the organism: HIV1 (HXB2)

Paste your sequences
[\[Sample Input\]](#)

Or upload your file: /Users/btf/Desktop/Outputs/OurData-PlusRefset.FASTA [Browse...](#)

Check box if appropriate ☐ Sequences are unaligned

Options

Region(s) to align and extract: Env CDS

☐ Insert [HXB2\(Accession #K03455\)](#) for HIV-1 or [SMM239\(Accession #M33262\)](#) for HIV-2 or SIV into the result set

☐ Remove [HXB2\(Accession #K03455\)](#) for HIV-1 or [SMM239\(Accession #M33262\)](#) for HIV-2 or SIV from the result set

☒ Codon align the region

Translation options

☐ Codons containing an [IUPAC character](#) are shown as "X".

☐ Codons containing an [IUPAC character](#) in a silent position are translated; others are shown as "X".

☐ Codons containing an [IUPAC character](#) are translated.

☒ Do not translate to amino acids

Note: codons containing "-" are always translated to either "-" (gap) or "#" (partial codon)

[Submit](#) [Reset](#)

Please be patient. Your input file must download to our server, where the actual work is performed. This can take several

Input is our data plus the “reference Set” and any other sequences we chose to add from the search interface.

Input: GeneCutterInput.FASTA

For this exercise, we want the Env gene, codon aligned, but not translated to proteins.

Output:
GeneCutterOutput.FASTA

Gene Cutter Results

Gene Cutter Mailback Form

Please enter the email address to send the results set:

Submit email address

- Results are stored on our server

An HTML link is e-mailed to the user when the run is complete

For this workshop, we will provide example files.

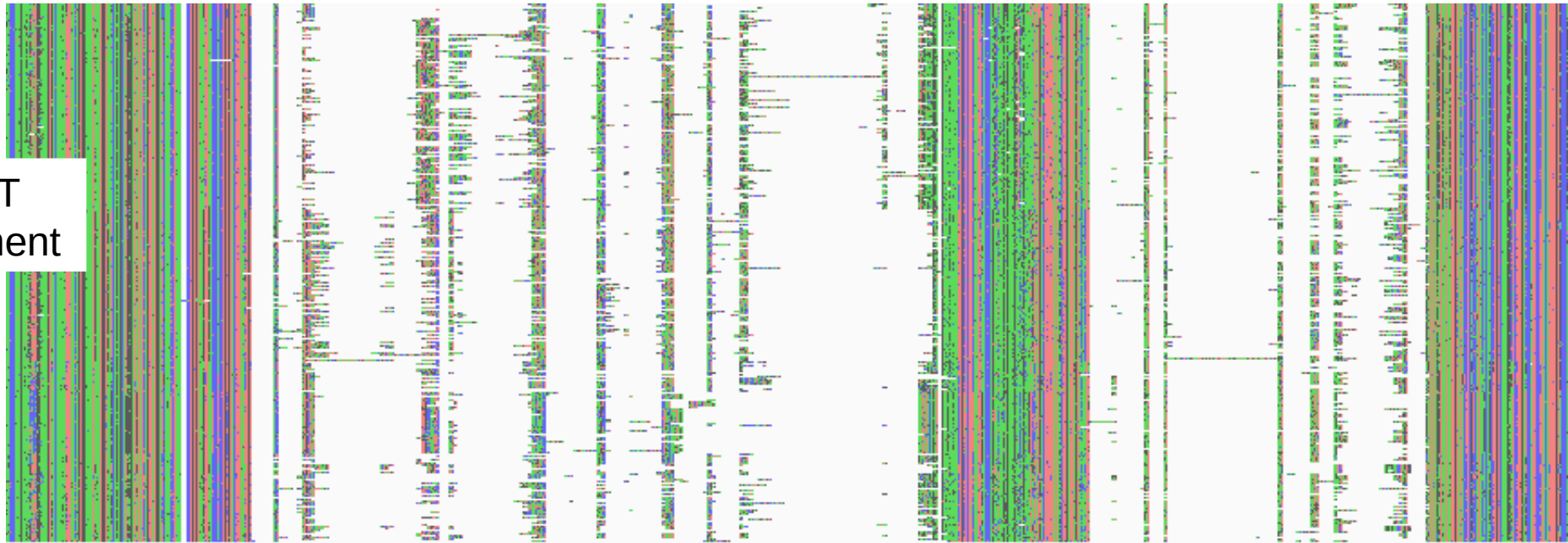
Gene Cutter Alignment

HIV V1/V2 sequences:
GeneCutterInput.FASTA

Result saved in Outputs folder
Alignments viewed with Pixel

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

MAFFT
alignment



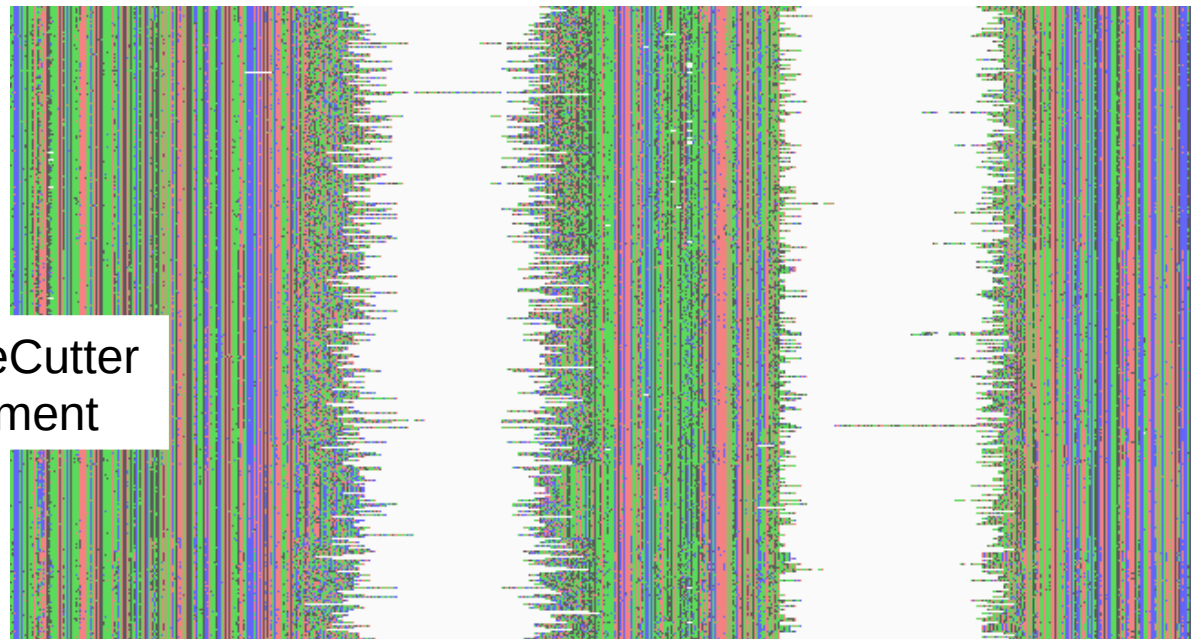
GeneCutter alignment:
“unalignable” regions compressed,
bases codon-aligned throughout.

File: GeneCutterOutput.FASTA

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

**Options: translations of all
genes in proper reading frames**

GeneCutter
alignment



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
- PhyML and IQTree interfaces are also available

<https://www.hiv.lanl.gov/content/sequence/PHYML/interface.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. Any
organism.

OR upload an
alignment file
here.

Tree parameters

Include reference sequences (HIV-1/CPZ only) ☐

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

Browse...

Tree parameters

Include reference sequences (HIV-1/CPZ only) ☐

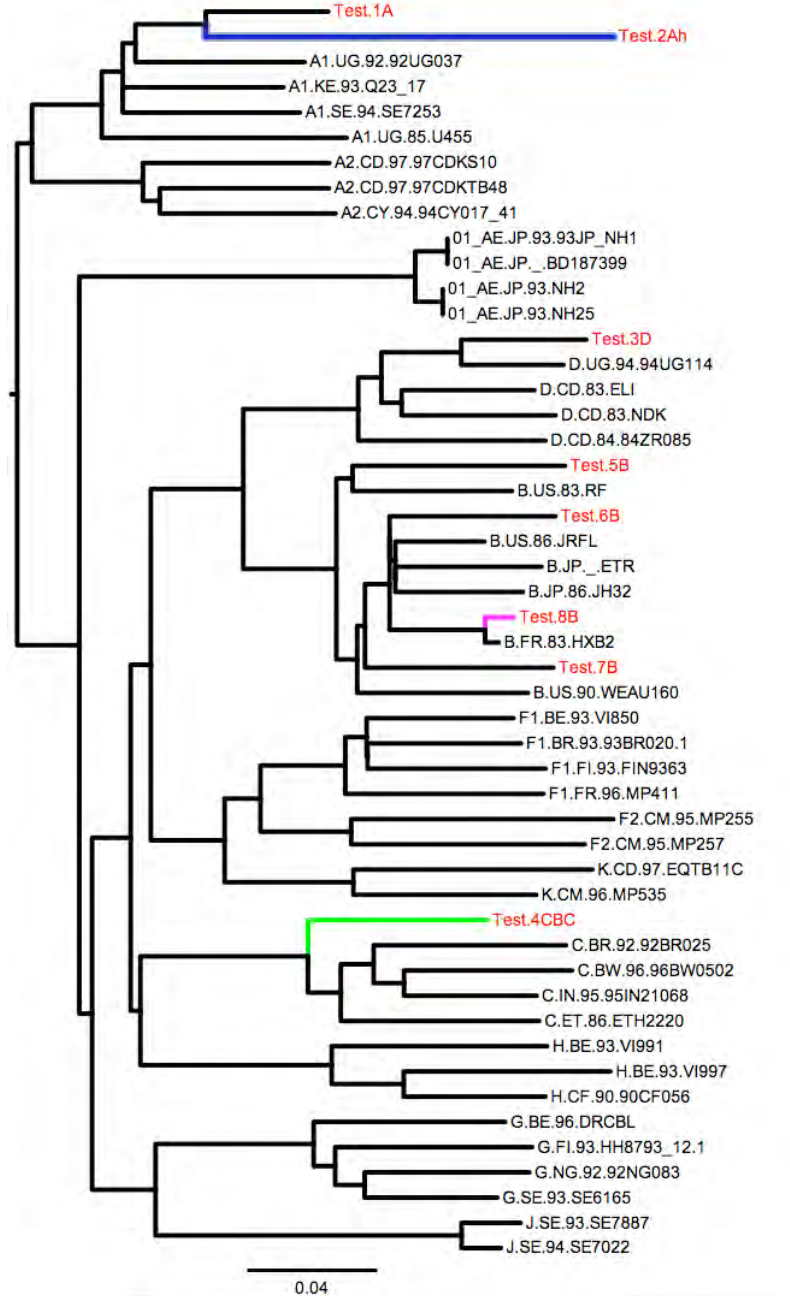
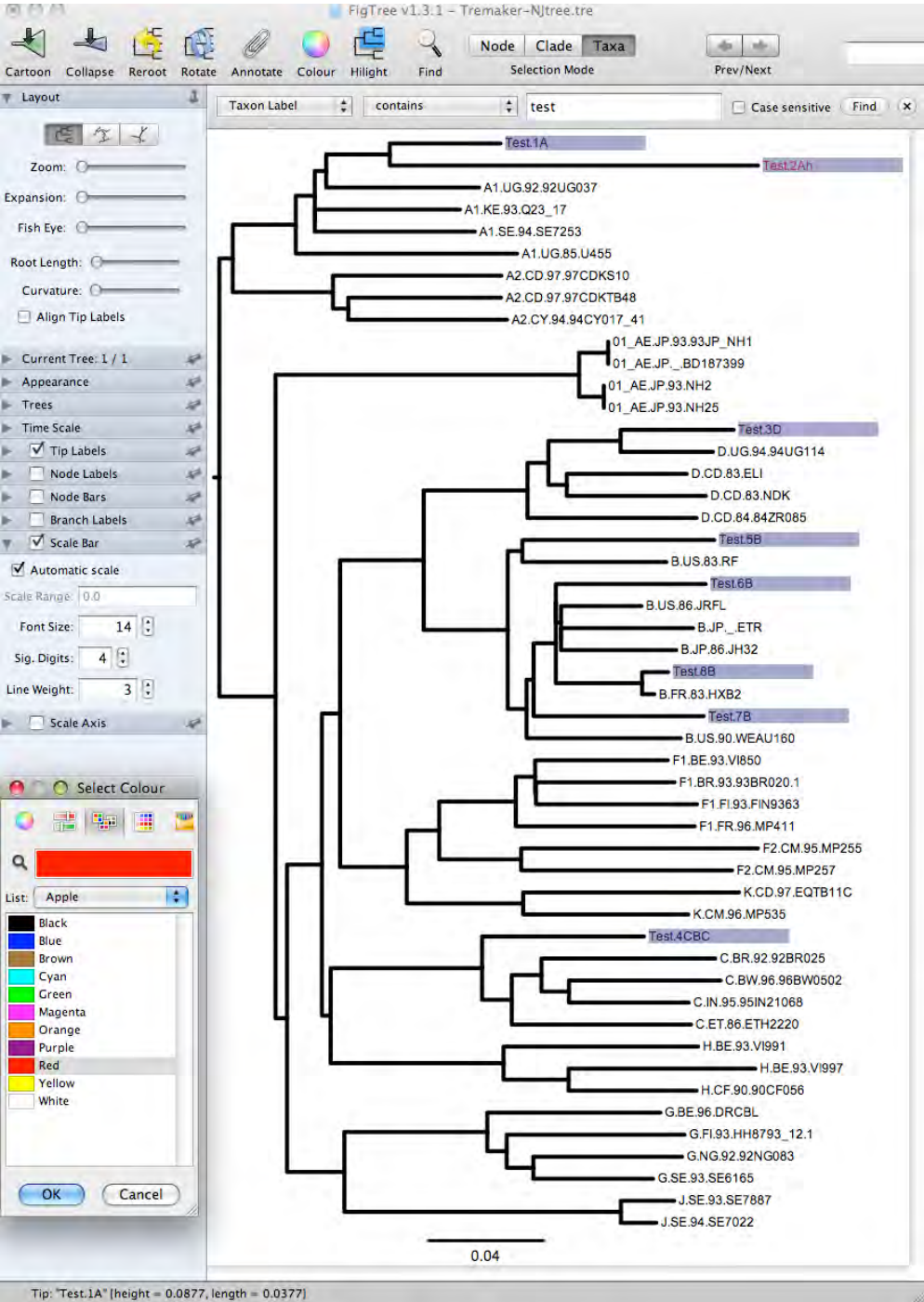
Submit

Reset

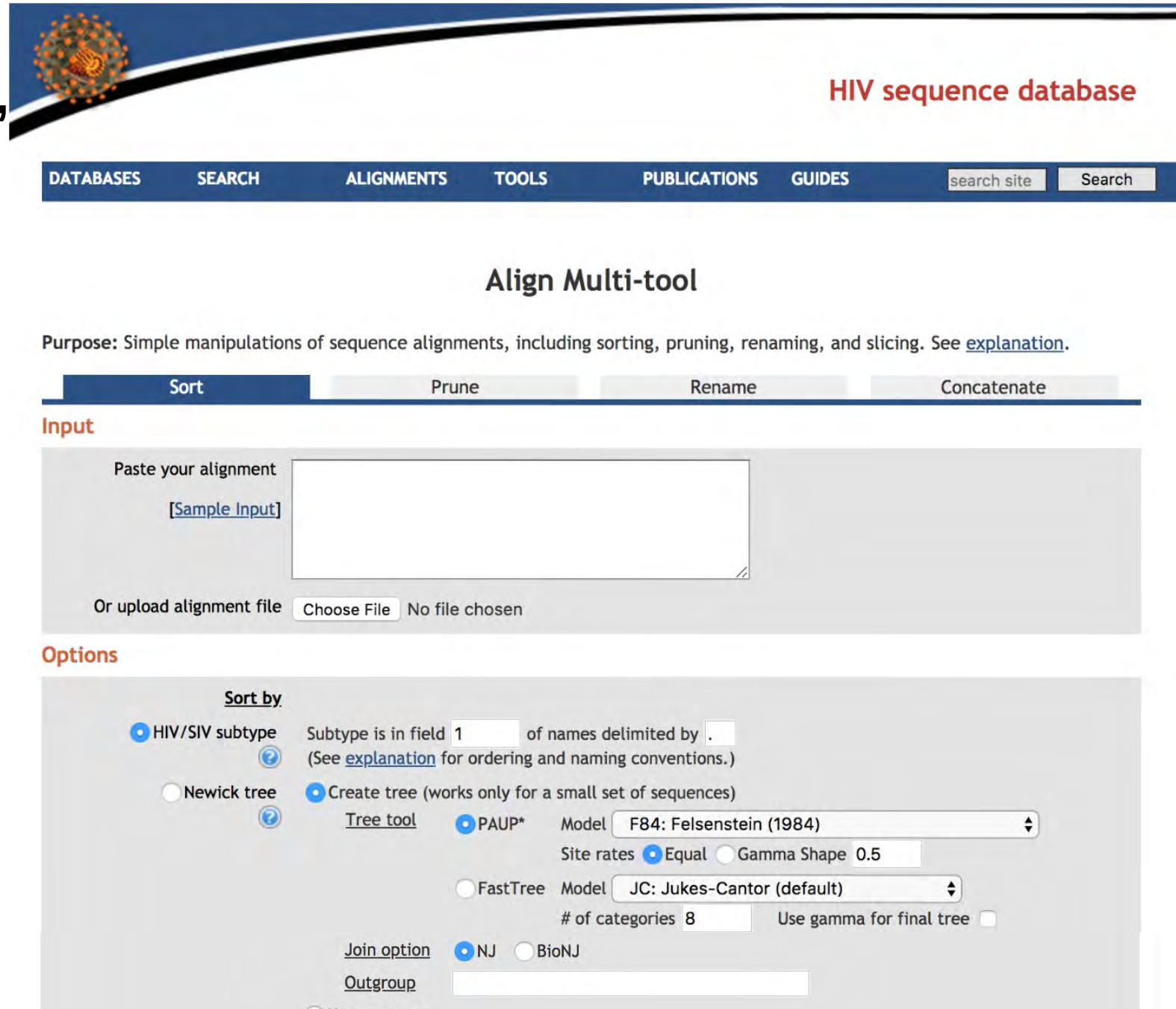
For this exercise
use the sample
input.

Include the
reference
sequences.

<http://tree.bio.ed.ac.uk/software/figtree/>



Alignment Multi-Tool



The screenshot shows the HIV sequence database website. At the top, there is a navigation bar with links: DATABASES, SEARCH, ALIGNMENTS, TOOLS, PUBLICATIONS, GUIDES. A search bar is on the right. Below the navigation bar, the title "Align Multi-tool" is centered. The purpose of the tool is described as "Simple manipulations of sequence alignments, including sorting, pruning, renaming, and slicing." There are four tabs: Sort, Prune, Rename, and Concatenate. The "Sort" tab is selected. Under the "Input" section, there is a text area for pasting an alignment, a link for sample input, and a file upload section. The "Options" section includes "Sort by" (HIV/SIV subtype, Newick tree, Create tree), "Tree tool" (PAUP*, FastTree), "Model" (F84: Felsenstein (1984), JC: Jukes-Cantor (default)), "Site rates" (Equal, Gamma Shape), "# of categories" (8), "Join option" (NJ, BioNJ), and "Outgroup".

HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES search site Search

Align Multi-tool

Purpose: Simple manipulations of sequence alignments, including sorting, pruning, renaming, and slicing. See [explanation](#).

Sort Prune Rename Concatenate

Input

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

Or upload alignment file Choose File No file chosen

Options

Sort by

☒ HIV/SIV subtype Subtype is in field 1 of names delimited by .
(See [explanation](#) for ordering and naming conventions.)

☐ Newick tree

☒ Create tree (works only for a small set of sequences)
[Tree tool](#)

☒ PAUP* Model F84: Felsenstein (1984)
Site rates ☒ Equal ☐ Gamma Shape 0.5

☐ FastTree Model JC: Jukes-Cantor (default)
of categories 8 Use gamma for final tree ☐

Join option ☒ NJ ☐ BioNJ

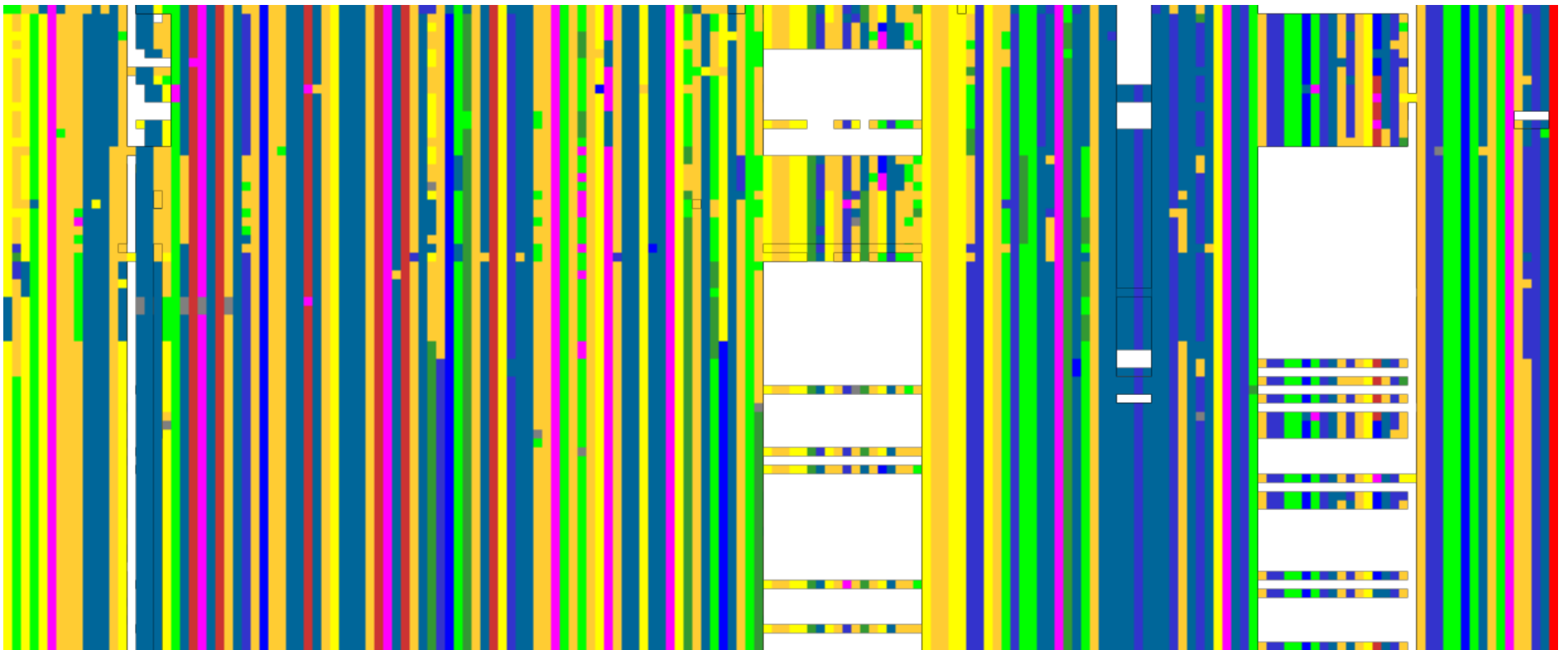
Outgroup

- Provides a suite of useful functions operating on alignments: **Sorting, pruning, renaming, concatenation**

- Sort *by tree*, by length, or by field(s) in sequence name
- Simplifies adding annotation adding to sequence names (by accession-number-based database search)
- Down-selects (or combines) both row-wise (sequence) and column-wise (region) slices of alignments, preserving sequence names

Alignment Multi-Tool

- Sort ***by tree***, by length, or by field(s) in sequence name
 - Clusters things that should be aligned similarly



Align Multi-tool results

Input & Options

Sort by Newick tree, create a tree

tool=FastTree, model=Whelan-And-Goldman 2001, # of categories=8, tree join=BioNJ

Top sequence picked: None

Number of sequences: 73

Run ID: ALIGN_MULTITOOL/nLAULg5zJu

Results

[\[Download\]](#) [\[Go back with this output\]](#)

```
>MAC.US.x.93062.AY607704
VVPFIPFAAAQQRGP--RKTIKCWNCGKEGHSARQCRAPRRQGCWKCGKMD
HVMAKCPDRQAGFLGLGPWGKKPRNFPMAQVHQGLM-----
----PTAPPEDPAVDLLKNYMLGKQOREKQRESRXKPYKEV-----
-----TEDLLHLNSLFGGDQ$
>MAC.US.x.96093.AY611489
VVPFIPFAAAQQRGP--RKXIKCWNCGKEGHSARQCRAPRRQGCWKCGKMD
HVMAKCPDRQAGFLGLGPWGKKPRNFPMAQVHQGLM-----
----PTAPPEDPAVDLLKNYMLGKQOREKQKESREKPYKEV-----
-----TEDLLHLNSLFGGDQ$
>MAC.US.x.1937.AY611495
VVPFIPFAAAQQRGP--RKPIKCWNCGKEGHSARQCRAPRRQGCWKCGKMD
HVMAKCPDRQAGFLGXGPWGKKPRNFPMAQVHQGLM-----
----PTAPPEDPAVDLLKNYMLGKQOREKQRESREKPYKEV-----
-----TEDLLHLNSLFGGDQ$
>MAC.US.x.96072.AY611491
VVPFIPFAAAQQRGP--RKPIKCWNCGKEGHSARQCRAPRRQGCWKCGKMD
HVMAKCPDRQVGFLGLGPWGKKPRNFPMAQVHQGLM-----
----PTAPPEDPAVDLLKNYMLGKQOREKQRESREKPYKEV-----
-----TEDLLHLNSLFGGDQ$
>MAC.US.x.81035.AY599200
VVPFIPFAAAQQRGP--RKPIKCWNCGKEGHSARQCRAPRRQGCWKCGKMD
HVMAKCPDRQXGFLGLGPWGKKPRNFPMAQVHQGLM-----
----PTAPPEDPAVDLLKNYMLGKQOREKQRESKEKPYKEV-----
-----TEDLLHLNSLFGGDQ$
>MAC.US.x.17EC1.AY033233
VVPFIPFAAAQQRGP--RKPIKCWNCGKEGHSARQCRAPRRQGCWKCGKMD
HVMAKCPDRQAGFLGLGPWGKKPRNFPMAQVHQGLM-----
```

Align Multi-tool

Purpose: Simple manipulations of sequence alignments, including sorting, pruning, renaming, and slicing. See [explanation](#).

Sort

Prune

Rename

Concatenate

Input

Paste your alignment

[\[Sample Input\]](#)

Or upload alignment file

Choose File

no file selected

Options

Prune sequences by

☒ Name match



e.g., DQ676872, F[1-2]

(for pattern, simple regex allowed)

Match type

☒ Exact word

☐ Pattern ☐ case insensitive

Match against

☒ full name

☐ field of names delimited by

☐ Custom list



Choose File

no file selected

Invert



☐ Remove other than the above specified sequences

Squeeze gaps

☐ Delete columns that are entirely gaps

Align Multi-tool

Purpose: Simple manipulations of sequence alignments, including sorting, pruning, renaming, and slicing. See [explanation](#).

Sort

Prune

Rename

Concatenate

Input

Paste your alignment

[\[Sample Input\]](#)

Or upload alignment file

Choose File

no file selected

Options

Export names

☐ Export Save sequence names to text file

Check uniqueness

☐ Check Find duplicate names

Replace names with

☒ HIV database
field values



Accession must be in input seq name to search database

☐ Seq name is accession

☒ Accession is in field of names delimited by

Numbering fields composing a new name

Subtype Country Year Name

Accession Patient code SE ID Patient ID

Field separator

Missing character for unavailable values

☐ Sequential numbers

Align Multi-tool

Purpose: Simple manipulations of sequence alignments, including sorting, pruning, renaming, and slicing. See [explanation](#).




Sort

Prune

Rename

Concatenate


Input

Upload alignment file(s)    Choose File no file selected Range 1,9-end,(empty)=full range

[\[Sample Input\]](#)

AddAlignment

Click & drag arrows to change order of alignments

Range numbers refer to  ☒ alignment columns, including gaps ☐ residues of 1st sequence

Submit

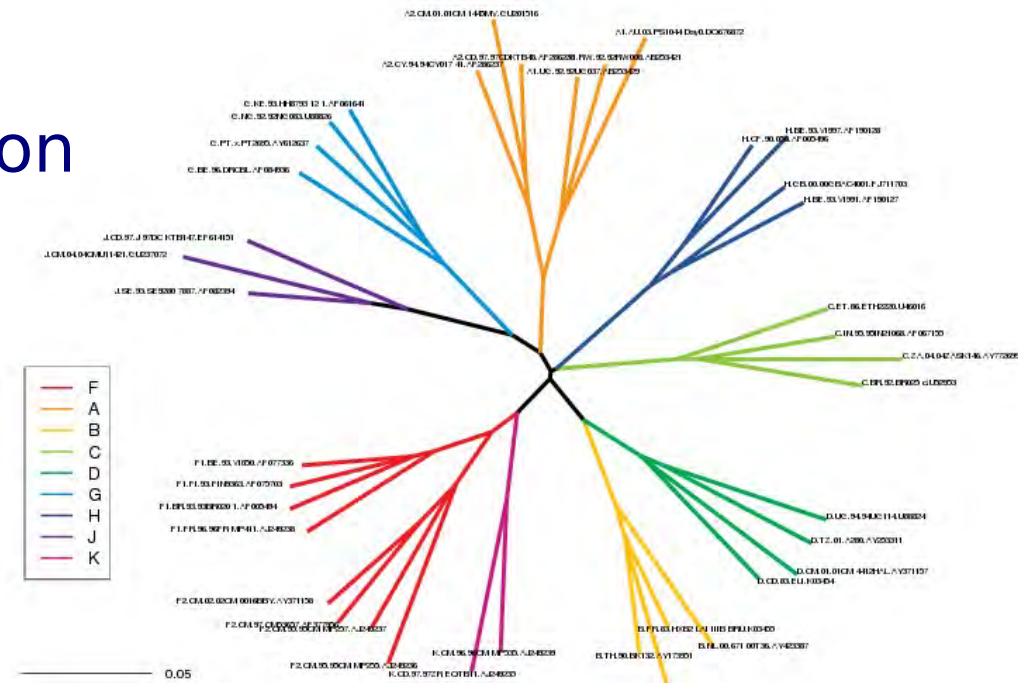
Reset

- The different functions of the multi-tool are a convenience.
- Annoying manipulations are easier and less error-prone.
- The ability to "stack" operations ("go back with this output") makes this a power tool.
- Easy manipulation of sequence names (by database query or by pairwise list) makes tree annotation much easier.

Making ~~decorative~~ informative trees for publication

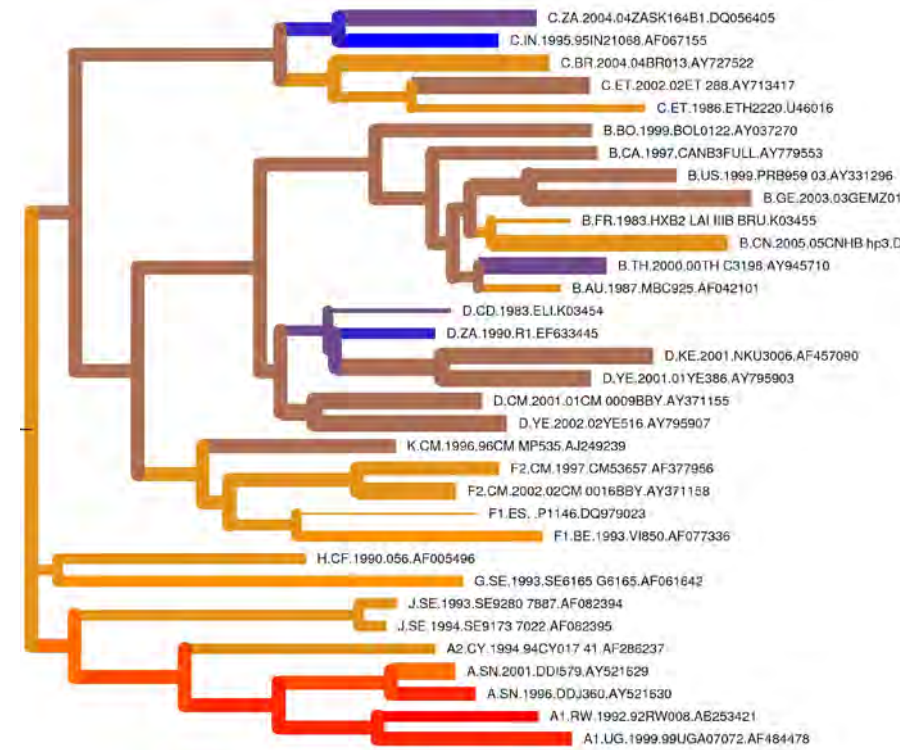
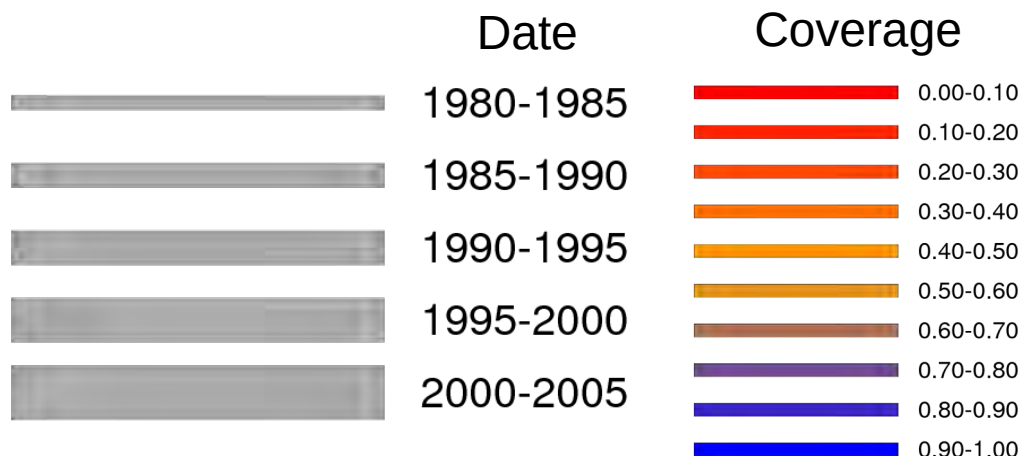
Rainbow Tree

- Colors branches based on strings in sequence names



AnnotateTree

- maps quantitative information to branch weights and colors.



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)

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:

(Results for sequence Test.8B)

Location in genome (red bar).

Numeric coordinates (useful for entry on search form) for DNA and amino-acids in all reading frames, with translations

Alignment of the query sequence to HXB2 (Similarity 97.8%):

```
Query  ATGAGAGTGA  AGGAGAAATA  TCAGCACTTG  TGGAGATGGG  GGTGGAAATG  50
      ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::
HXB2  ATGAGAGTGA  AGGAGAAATA  TCAGCACTTG  TGGAGATGGG  GGTGGAGATG  6274

Query  GGGCACCATG  CTCCTTGGGA  TATTGATGAT  CTGTAGTGCT  ACAGAAAAAT  100
      ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::
HXB2  GGGCACCATG  CTCCTTGGGA  TGTTGATGAT  CTGTAGTGCT  ACAGAAAAAT  6324

Query  TGTGGGTCAC  AGTCTATTAT  GGGGTACCTG  TGTGGAAGGA  AGCAACCACC  150
      ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::
HXB2  TGTGGGTCAC  AGTCTATTAT  GGGGTACCTG  TGTGGAAGGA  AGCAACCACC  6374

Query  ACGCTATTTT  GTGCATCAGA  TGCTAAAGCA  TATGATACAG  AGGTACATAA  200
      ::  ::::::::::  ::::::::::  ::::::::::  ::::::::::  ::
HXB2  ACTCTATTTT  GTGCATCAGA  TGCTAAAGCA  TATGATACAG  AGGTACATAA  6424
```

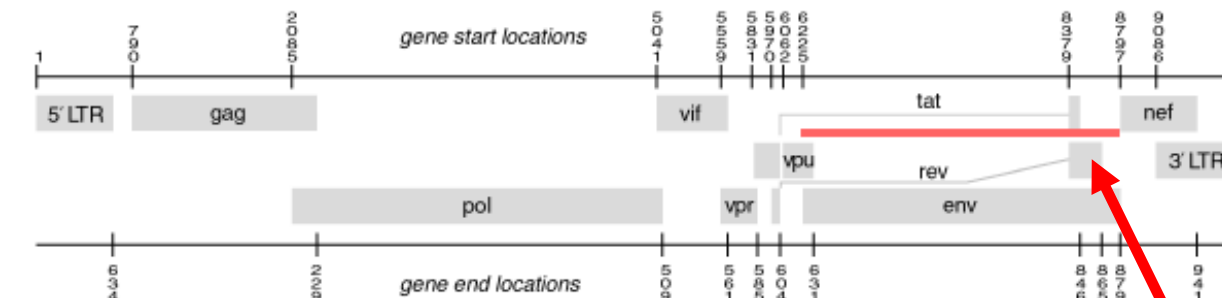


Table of genomic regions touched by query sequence. (Protein translation of query shown in blue.)				
CDS	Nucleotide position relative to CDS start in HXB2	Nucleotide position relative to query sequence start	Nucleotide position relative to HXB2 genome start	Amino Acid position relative to protein start in HXB2
Vpu	164 → 249	1 → 86	6225 → 6310	55 → 83
XESEGISALVEMGVEMGHAPWDIDL*				
gp160	1 → 2571	1 → 2586	6225 → 8795	1 → 857
Notice: length of gp160 portion of query (2586) is greater than its length in HXB2 (2571).				
MRVKEKYQHLWRWGWKGMTLLGILMICSATEKLWVTVYGVVPVWKEATTTLFCASDAKAYDTEVHNWVATHACVPTDPNPQEVVLNVTFNFMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVSLKCTDLGNATNTNSSNTSSSGEMMKEGEIKNCSFNISTSIRGKVQKEYAFYKLDIIPIDNDTTSYTLTSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNKNTFNGTGPCTNVSTVQCTHGIRPVVSTKLLNGSLAEVEVIRSANFTDNAKTIIVQLNQSVEINCTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAHCNISRAKWNATLKQJASKLREQFGNKTIIIFKQSSGGDPEIVTHSFNCGGEFFYCNCSTQLFNSTWFNSTWSTEGSNNTTEGSDTITLPCRIFQFINMWQEVGKAMYAPPISGQIRCSSNITGLLLRDGGNNNNGSEIFRPGGGDMRDNWRSESYKYKVVKIEPLGVAPTAKARRRVQREKRAVGIGALFLGFLGAAGSTMGARSLTLTVQARQLLSGIVQQNNLLRAIEAQHLLQLTVWGIKQLQARILAVERYLKDQQLLGIWGCSEKLICTAVPWNASWSNKSLEQIWNMMTWMEWDREINNNTSLIHSLEESQNNQKEKNEQELVELDK*ASLWNWFNITN*LWYIKIFIMIVGGGLYGLRIVFAVLSIVNRVRQGYSPLSFQTHLPTPRGPDPEGIEEGGERDRDRSIRLVNGLSLIWDLRSLCLFSYHRLDLLLIVTRIVELLGRRGWALKYWWNLQYWSQELKNSAVSLLNATAIAVAEGTDRVIEVVQGACRAIRHIPRRIRQGLERILL*				
gp120	1 → 1533	1 → 1548	6225 → 7757	1 → 511
Notice: length of gp120 portion of query (1548) is greater than its length in HXB2 (1533).				
MRVKEKYQHLWRWGWKGMTLLGILMICSATEKLWVTVYGVVPVWKEATTTLFCASDAKAYDTEVHNWVATHACVPTDPNPQEVVLNVTFNFMWKNDMVEQMHEDIISLWDQSLKPCVKLTPLCVSLKCTDLGNATNTNSSNTSSSGEMMKEGEIKNCSFNISTSIRGKVQKEYAFYKLDIIPIDNDTTSYTLTSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNKNTFNGTGPCTNVSTVQCTHGIRPVVSTKLLNGSLAEVEVIRSANFTDNAKTIIVQLNQSVEINCTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAHCNISRAKWNATLKQJASKLREQFGNKTIIIFKQSSGGDPEIVTHSFNCGGEFFYCNCSTQLFNSTWFNSTWSTEGSNNTTEGSDTITLPCRIFQFINMWQEVGKAMYAPPISGQIRCSSNITGLLLRDGGNNNNGSEIFRPGGGDMRDNWRSESYKYKVVKIEPLGVAPTAKARRRVQREKR				
gp41	1 → 1038	1549 → 2586	7758 → 8795	1 → 346
AVGIGALFLGFLGAAGSTMGARSMTLTVQARQLLSGIVQQNNLLRAIEAQHLLQLTVWGIKQLQARILAVERYLKDQQLLGIWGCSEKLICTAVPWNASWSNKSLEQIWNMMTWMEWDREINNNTSLIHSLEESQNNQKEKNEQELVELDK*ASLWNWFNITN*LWYIKIFIMIVGGGLYGLRIVFAVLSIVNRVRQGYSPLSFQTHLPTPRGPDPEGIEEGGERDRDRSIRLVNGLSLIWDLRSLCLFSYHRLDLLLIVTRIVELLGRRGWALKYWWNLQYWSQELKNSAVSLLNATAIAVAEGTDRVIEVVQGACRAIRHIPRRIRQGLERILL*				
Tat2	1 → 91 216 → 306 (Tat)	2170 → 2260	8379 → 8469	1 → 31 73 → 102 (Tat)
XPTSQPRGDPTGPKE*KKKVERETETDPFD*				
Rev2	1 → 275 77 → 351 (Rev)	2170 → 2444	8379 → 8653	1 → 92 26 → 117 (Rev)
XPPNPEGTQRARRNRRRRWRERQRQIHSISERILSTYLGRSAEPVPLQLPLLERLTLDNEDCGTSGTQGVGSPQILVESPTVLESGTK*				

Variable Region Characteristics

Purpose: Variable Region Characteristics analyzes protein sequences for V1, V2, V3, V4, V5 and reports length, glycosylation sites, and net charge.

Details: The tool accepts a set of aligned protein sequences in Fasta, IG, table, and other formats, along with an optional reference sequence.

Select Regions

If you input an HIV alignment that includes HXB2

Make sure you understand the [explanation](#) before

V1: ☐ Full

V2: ☐ Full

V1+V2: ☐ Full

V3: ☐ Full

V4: ☐ Full

V5: ☐ Full

Alignment

Title of Analysis

Paste your alignment here

[Use Sample Input](#)

[Clear Input Data](#)

Or upload a data file no file selected

Prefix Summary

If your sequence names have information such as clade embedded as an alphanumeric prefix (A1_ or A1. or A1- or A1*) in the name, and you would like a summary by those values, click the

☐ Include a prefix summary

Select Positions

☐ Use Alignment positions to

☐ Use Reference positions to

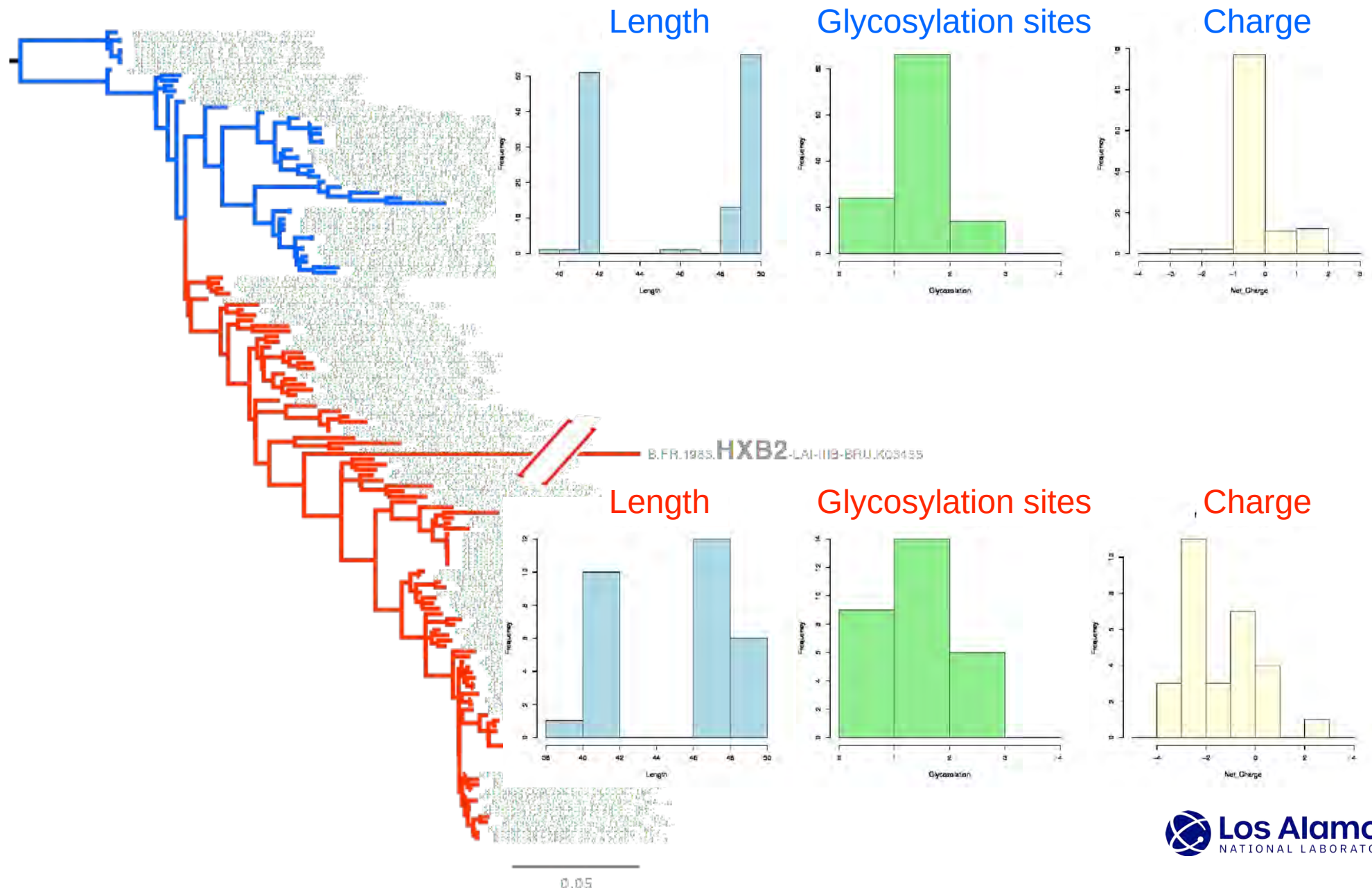
Net Charge Options

You may choose how net charge is computed:

☒ KRH = +, DE = - (default)

☐ KR = +, DE = -

Variable Region Characteristics

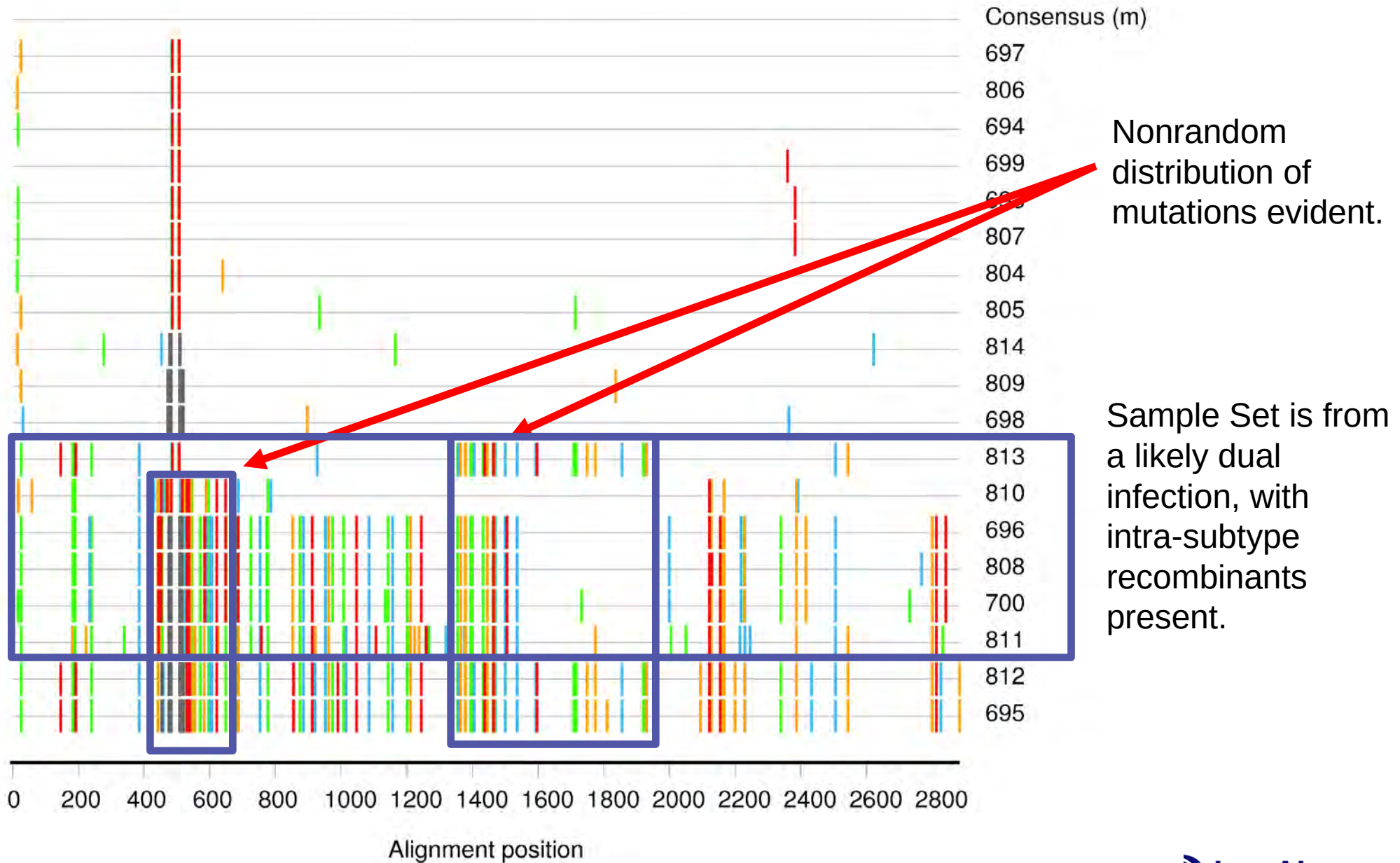


Highlighter

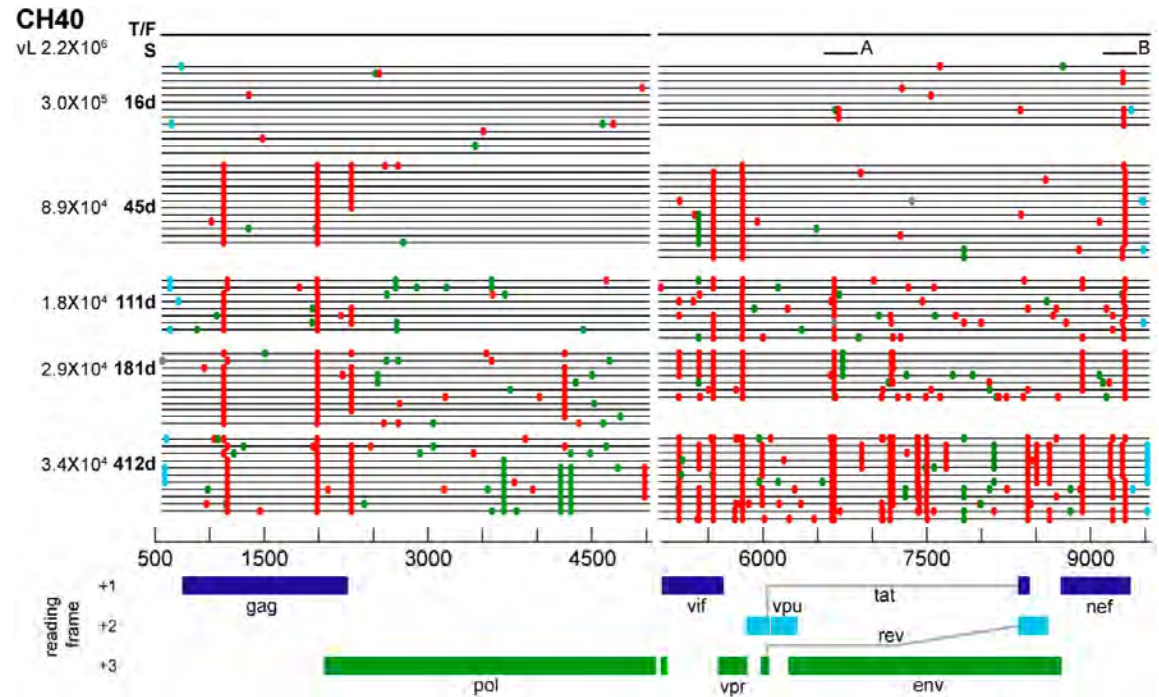
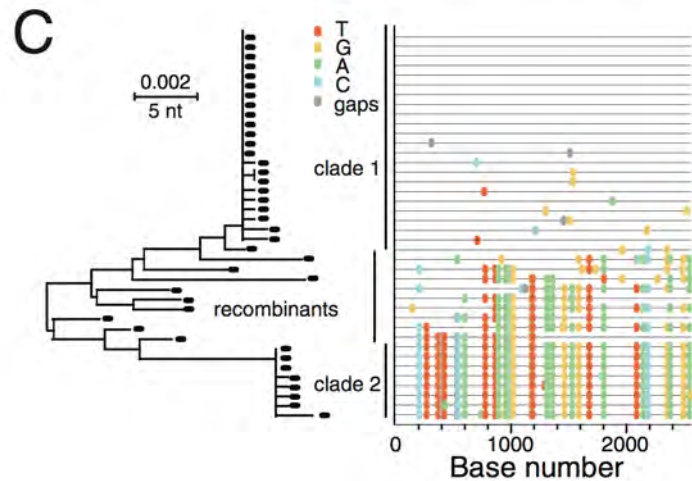
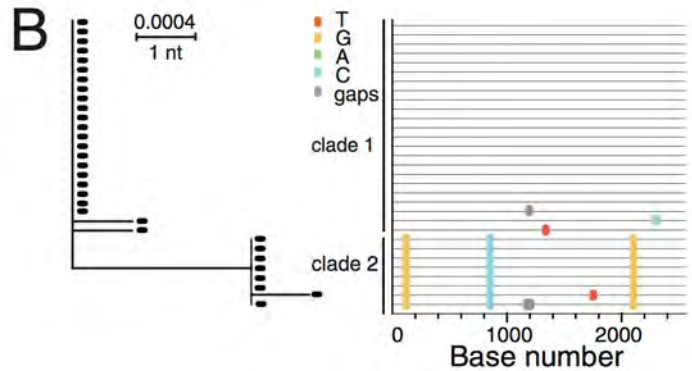
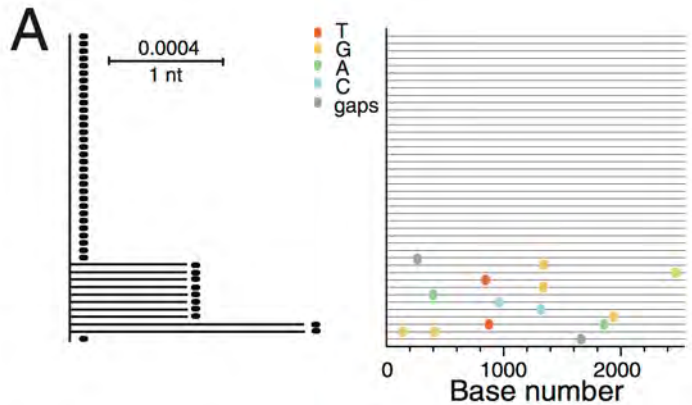
- Highlights mutations relative to a reference strain, particularly useful for intra-patient analyses.
- Highlights:
 - syn/non-syn
 - transition/transversion
 - APOBEC motifs
- Sorts on similarity
- Visualize recombination of closely related sequences

Highlighter sample data

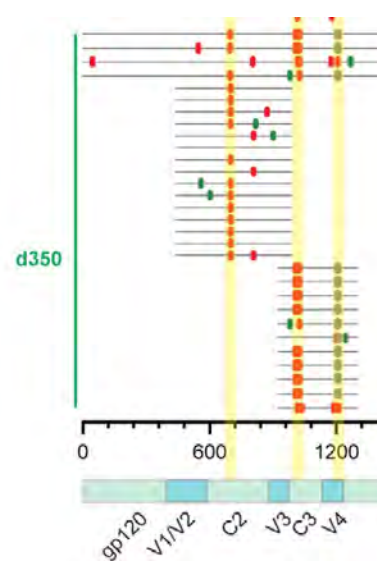
Mismatches compared to master



Highlighter examples



J.F. Salazar-Gonzalez, M.G. Salazar, B.F. Keele et al. (2009) J. Exp. Med. 206:1273-1289

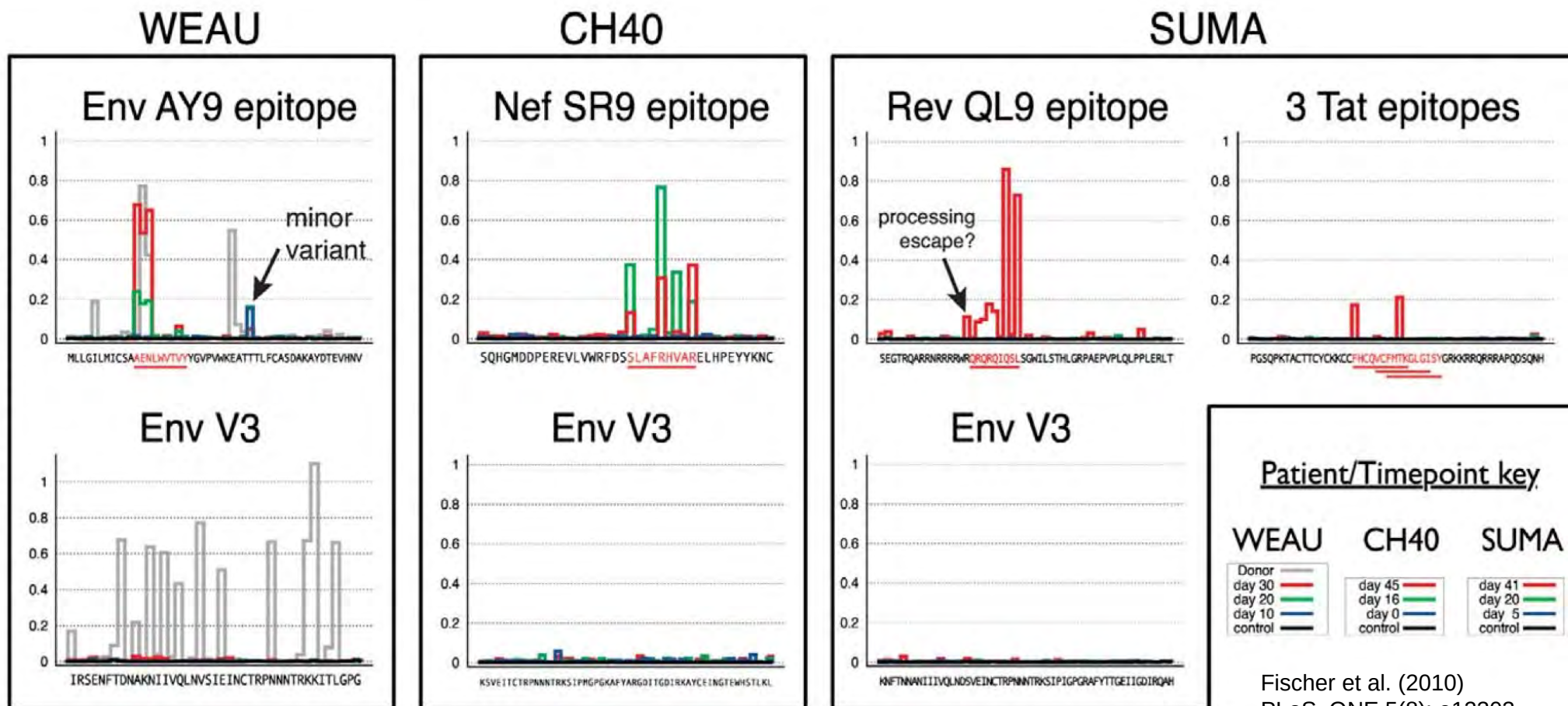


- Infection multiplicity
- CTL escape
- Antibody escape

Entropy

- Quantifies per-site variability within a sequence.
- Highlights regions of rapid evolution:
 - CTL or antibody epitopes
 - Reveals dynamics of site changes (e.g. immune escape)

Shannon Entropy

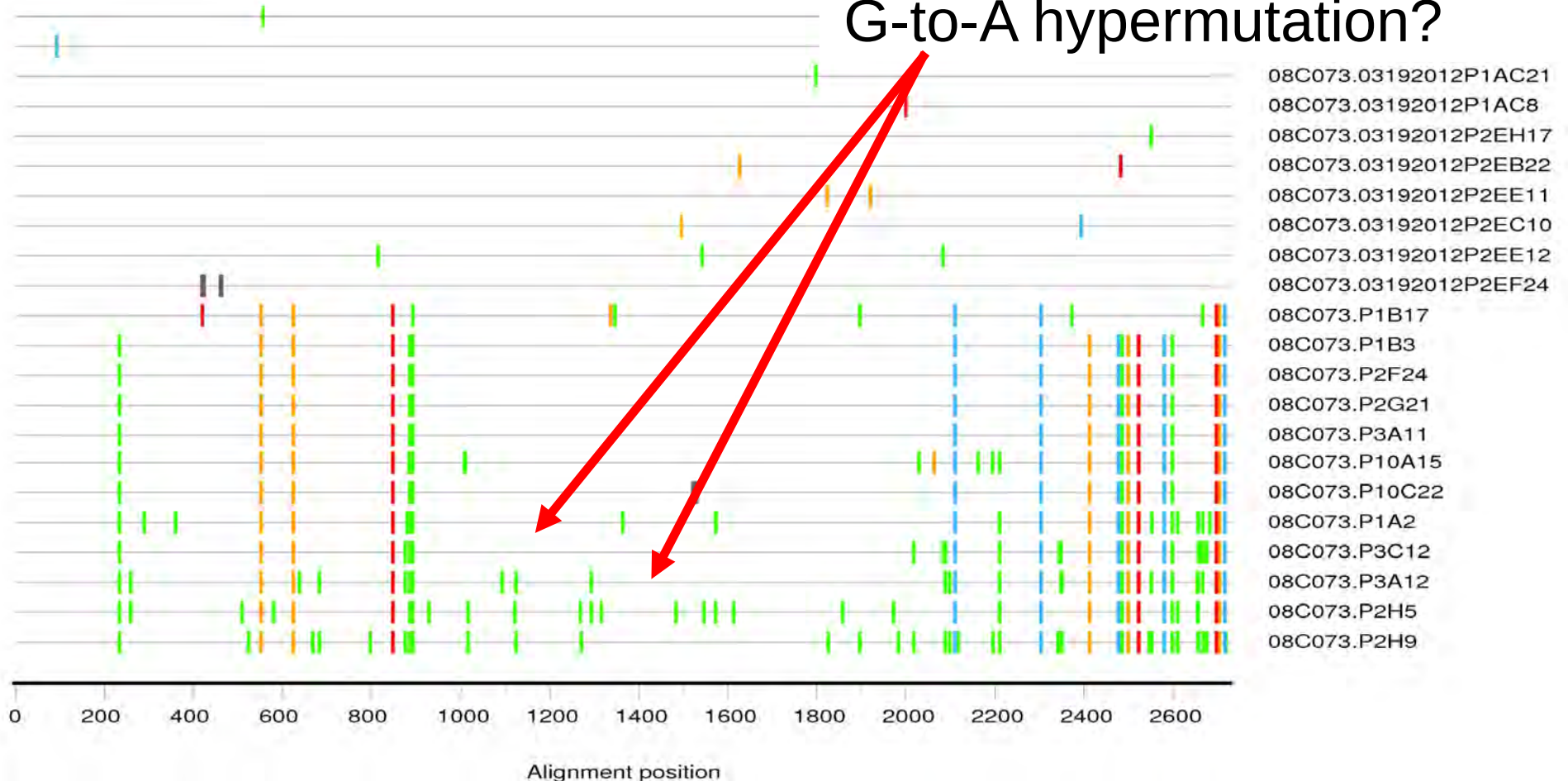


Hypermutation

Mismatches compared to master(s)

■ Question:

Are all those “A” residues the effect of *APOBEC*-mediated G-to-A hypermutation?



Hypermur Tool

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, Abfalterer, 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:

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

- Assesses statistical signal of hypermutation
- Detects APOBEC-3G mediated G-to-A hypermutation as default
- Can be adapted to detect any fuzzy motif in relation to a control pattern
- An “easy version” is included in the QC tool
- Some datasets are enriched for hypermutation, even when counts for individual sequences aren’t significant.

Hypermut results

Hypermut 2.0

Your pattern definitions are as follows. Where there is no pattern (i.e., just '...') all sequences will match.

Pattern Upstream From → To Downstream

'Mut' ... G → A RD ...
'Control' ... G → A YN|RC ...

Results

'Potential Mut' or 'Potential Control' means a match to the corresponding Upstream, From, and Downstream patterns above, while an actual 'Mut' matches those and the To pattern as well. We consider a P-value less than 0.05 to indicate a hypermutant when using the default patterns.

Sequence:	Muts:	Out of:	Controls:	Out of:	Rate Ratio:	Fisher Exact P-value:
(Select for graphing)	(Match Sites)	(Potential Mut Sites)	(Control Muts)	(Potential Controls)	(A/B)/(C/D)	(~P(Muts,Poten.Muts-Muts, Cntrlis,Poten.Cntrlis-Cntrlis))
<input checked="" type="checkbox"/> Seq2	0	71	0	54	undef	1
<input checked="" type="checkbox"/> Seq5	4	69	1	52	3.01	0.282669
<input checked="" type="checkbox"/> Seq7	26	71	1	54	19.77	5.35061e-07
<input checked="" type="checkbox"/> Seq14	48	71	9	54	4.06	8.26961e-09

A significant excess of G-to-A mutations at APOBEC match sites (compared to non-match sites) indicates hypermutation.

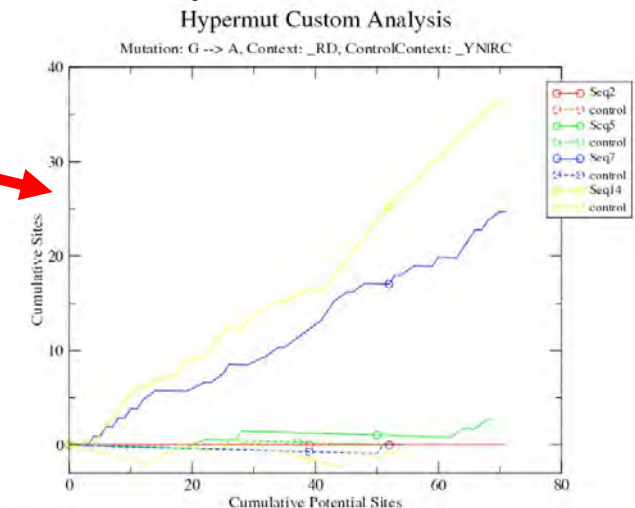
View Sites Along Sequence

Type of graph:
☒ Locations of Matches
☐ Cumulative Matches (try me!)

[Graph Matches](#) (opens in a new window)

Optional Controls:
 Show region: From to
 Graph Title: Hypermut Custom Analysis
 Access xmgrace compatible [datafile](#).

Cumulative mutation Graph is useful



Original Hypermut Output

The input file has 5 sequence(s)

[Download](#) the following info as text file

Sequence Length: 645

Compared to SEQ1, 264 As, 126 Gs, 92 Cs, 163 Ts

GRAPH TABLE

[Dinuc Context:](#)

Sequence names	Ratio	#diffs	perc	Gs	#A→G	#G→A	GG	GAG	GC	GT	OBSERVED CHANGES
SEQ2	0/0	1	0.00	0	0	0	0	0	0	0	TC
SEQ5	5/2	33	3.97	2	5	2	3	0	0	0	CA, MA, CA, CT, CA, CA, AT, CT, AC, A, M, A, C, M

output:
[pdf version here.](#)
[postscript version here.](#)

Quality Control Tool

- Incorporates existing HIV database tools
 - GeneCutter
 - RIP, BLAST
 - HyperMut (simple version)
 - Neighbor-joining Trees
- Output is an email with link to a summary report
- Why use it?
 - Prepare sequences for GenBank submission
 - Prevent database pollution
 - Avoid embarrassment!
- <http://www.hiv.lanl.gov/content/sequence/QC/index.html>

Quality Control Tool

<https://www.hiv.lanl.gov/content/sequence/QC/index>



Quality Control

HIV-1 Sequence Quality Analysis

Purpose: (1) Examines sets of HIV-1 nucleotide sequences for common problems. (2) Prepares HIV-1 sequence sets, together with related data, for submission to GenBank.

Input: The tool accepts HIV-1 nucleotide sequences in [Fasta](#) format. Before using, please read the [QC/GenBank Tool Explanation](#). If you have already performed QC analyses and you only want to generate a Sequin file, you can also use the [GenBank Entry Generation](#) tool.

Input

Details

QC analysis: This tool will perform a set of tests to help you find problems with your sequences. The [QC/GenBank Tool Explanation](#) gives details about how to assess the results of these analyses. QC results will include:

- subtype (from [RIP](#)),
- most similar database sequence (from [HIV BLAST](#)),
- phylogenetic tree of each single sequence with subtype references (from [Neighbor TreeMaker](#)),
- phylogenetic tree of all sequences together with subtype references (from [Neighbor TreeMaker](#)),
- number of stop codons and frameshifts (from [GeneCutter](#)),
- hypermutation (from [HyperMut](#)).

Preparing GenBank submissions: This tool can also be used to prepare HIV-1 sequences for GenBank submission. This step is *not* required if you only want to do the QC analysis.

Related Links:

[QC/GenBank Tool Explanation](#)
[Sequence Quality Control Tutorial](#)

[GenBank Entry Generation](#)

After the QC analyses, you can continue directly to the GenBank entry creation tool.

GenBank preparation procedure requires a comma separated (CSV) spreadsheet of annotations, as described on the help pages.

http://www.hiv.lanl.gov/content/sequence/QC/field_help.html

Easy to enter in spreadsheet (export as CSV format), or in text editor

Quality Control Tool

- Summary of results from analysis programs
- Useful for helping to determine subtype, hypermutation, mislabeling of samples, spotting (some) lab strain contaminants

Summary 9876

Job # **9876**

Title QC_Submission

[NJ Tree \(all sequences\)](#)

Select ☐ All ☐ None

Name	Blast	RIP Subtype	Tree	Stop Codons	Frameshifts	Hypermutation	GeneCutter Result
<input type="checkbox"/> sequence1	EU577525 US B 99	B	NJ Tree	0	3	Not Detected	GeneCutter Result
<input type="checkbox"/> sequence2	EU577511 US B 100	B	NJ Tree	0	3	Not Detected	GeneCutter Result
<input type="checkbox"/> sequence3	EU846964 BE B 100	B	NJ Tree	0	1	Possible	GeneCutter Result
<input type="checkbox"/> sequence42	KU499345 GB D 100	D,G	NJ Tree	0	2	Not Detected	GeneCutter Result

Select ☐ All ☐ None

Please select *all* sequences that you want to submit. They should be submitted to GenBank as a single Sequin file.

[Create GenBank submission file from selected sequences](#)

[Download Summary](#)

Click on each result link to see details of analyses.

Optional: prepare results to submit to GenBank.

Rules for (HIV) Sequence Data

LOOK AT YOUR SEQUENCES in a dedicated alignment editor

- toggle between DNA and amino-acid views
- look at large and small scales throughout the alignment
- don't implicitly trust machine alignment

Build a tree with samples plus references *as part of initial analysis*

- Include sequences from all your samples
- Include reference sequences (sensible outgroups!)
- Include sequences previously generated in your lab! (Paranoia pays!)

Use robust sequence names

- for public sequences include accession numbers
- for new sequences include patient/subject IDs and relevant metadata (e.g. sample timepoints, tissue type)
- **Make highlighter plots for closely-related sequences**
- **Use sequence locator to check genome and protein coordinates (epitope locations!)**
- **Use our “Quality Control” pipeline for HIV sequences**
- **Submit to GenBank**
- **Call your mother**

Thank you for attending!

We are happy to help with research questions on the use of our tools and database.

We are thrilled to get ideas for further tool development!

Contact us: seq-info@lanl.gov or immuno@lanl.gov