HIV Database Workshop

www.hiv.lanl.gov

seq-info@lanl.gov

Presenters: Will Fischer, Brian Foley, Bette Korber

Database PIs: Bette Korber, Thomas Leitner, Karina Yusim, Brian Foley

Additional database staff: Werner Abfalterer, 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

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)***
- *Sequence Locator tool: HIV, SIV, HCV, HFV*
- *QuickAlign: HIV, SIV, HCV, HFV*
- ***Variable Region Characterization***
- ***Highlighter***
- ***Entropy***
- ***Hypermut***
- ***Quality Control (HIV)***

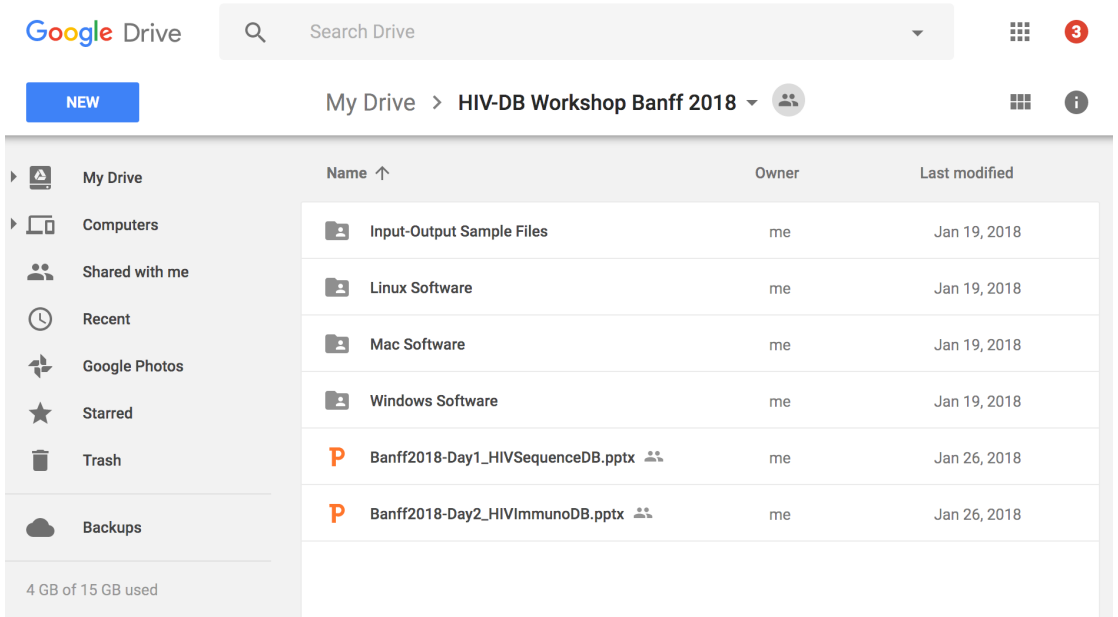
Los Alamos
NATIONAL LABORATORY

Datasets and programs

http://tinyurl.com/HIV-DB-2018

Aliview
Bioedit
Figtree

updated slides
datafiles



Los Alamos
NATIONAL LABORATORY

Entry page at http://www.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 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). Our content is reviewed by an Editorial Board.

SEQUENCE DATABASE ▸ VACCINE DATABASE ▸

IMMUNOLOGY DATABASE ▸ OTHER VIRUSES ▸

News: Archived News ▸

New tool: CombiNAber

The CombiNAber tool predicts neutralization by combinations of neutralizing antibodies using IC_{50} and/or IC_{80} data on individual antibodies as inputs. The output data are IC_{50}/IC_{80} titers, incomplete neutralization, instantaneous inhibitory potential (IIP), and coverage by multiple active antibodies for antibody combinations. Output analyses include ranking of antibodies/antibody combinations and figures showing the comparison of best antibodies/antibody combinations using the above metrics. *30 January 2017*

HIV-1 Antisense Protein alignments

We now provide alignments for the HIV-1 Antisense Protein (ASP). This information is included in the resources on the Special Interest Alignments page. *25 January 2017*

Los Alamos NATIONAL LABORATORY

HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Sit

- Sequence DB
- Immunology DB
- Vaccine DB
- HCV DB
- HFV DB

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

Site Statistics usage information for www.hiv.lanl.gov

How to Cite this Database

Editorial Board

News: Archived 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*

last modified: Mon Oct 31 13:51 2016

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

Los Alamos NATIONAL LABORATORY

Multiple paths to most tools

Los Alamos NATIONAL LABORATORY

HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Sit

- Tutorials
- CRFs
- HIV-1 Gene Map
- In-depth Annotation
- Neutralizing Antibody Resources & CATNAP
- 3D Structure
- Data Dictionary
- How to Cite this Database
- HIV Database News
- FAQs
- Links

All kinds of basic information about HIV and about our database

Tutorials and Basic Informatic

Previous workshop presentations

Tutorials

Keystone 2014 HIV sequence database workshop

Keystone 2014 HIV Immunology database workshop

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

How to use these databases summaries of workshops given at conferences

HIV numbering relative to reference strain HXB2

SIV numbering relative to reference strain SIVmm239

Articles

3D views of HIV macromolecular structures provides links to 3D views of HIV proteins

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

Reference Inform

Circulating recombinan tails about all documented CRFs

HIV-1 gene map illustra 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

FAQs answers basic questions about the HIV Sequence Database

Links HIV/AIDS resources and bioinformatics tools on other websites

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

last modified: Tue Aug 8 12:41 2017

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

Los Alamos NATIONAL LABORATORY

Search Interface

- **Results (what you want)**
  - Can download aligned or unaligned sequences
  - Alignments 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, 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

Los Alamos
NATIONAL LABORATORY

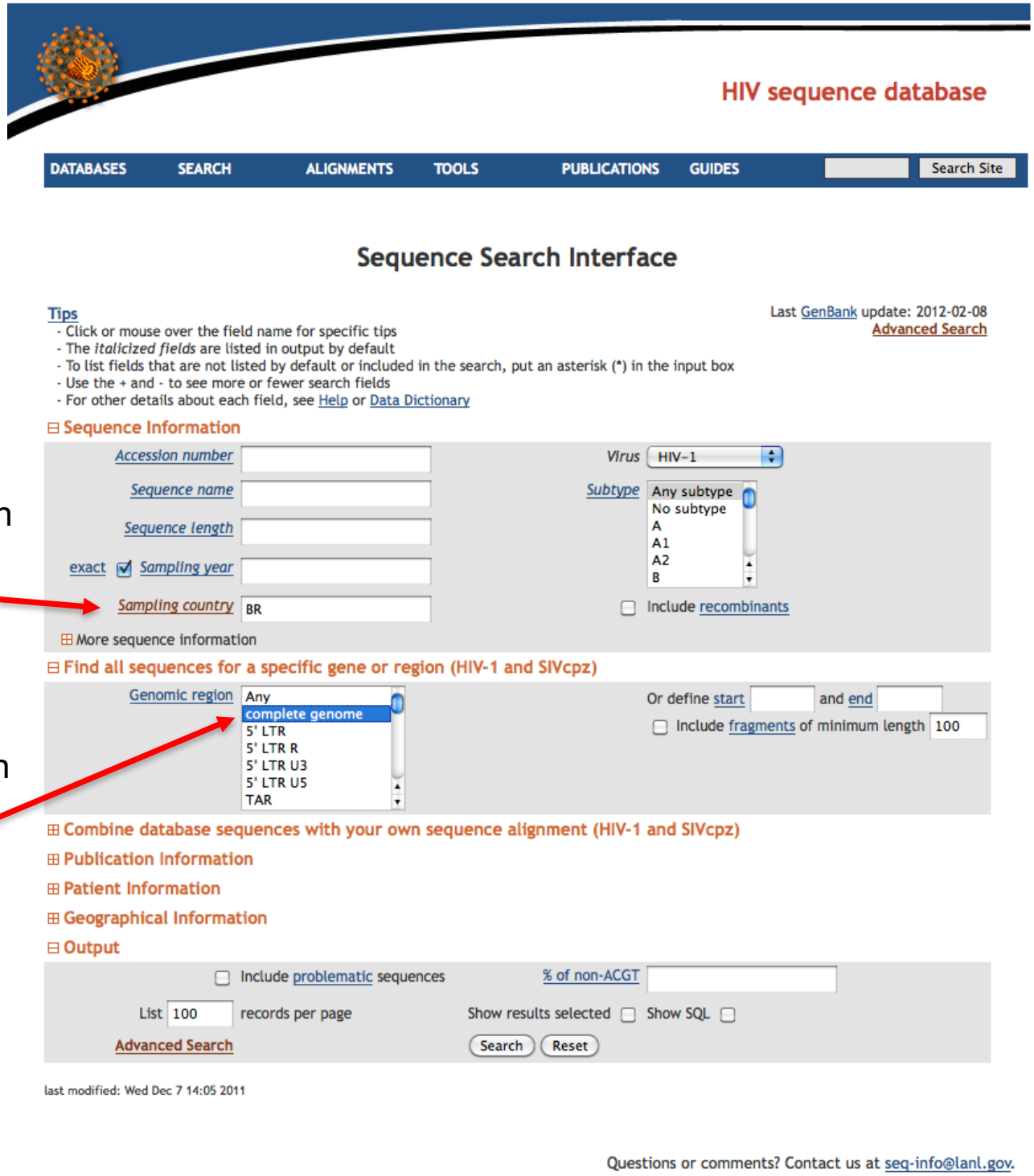
Today's Sequence Search example workflows

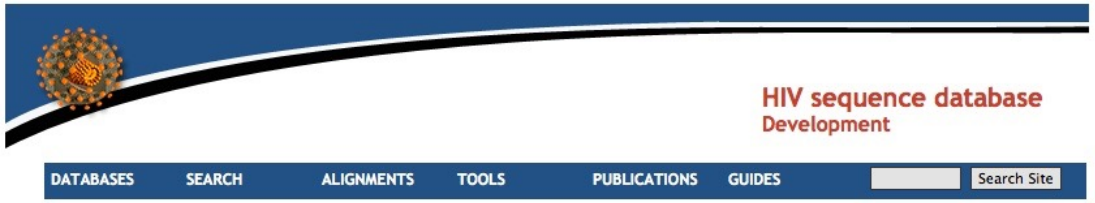
- Assemble a country-wide whole-genome data set:
  - Get all available complete genome sequences from a given country (Brazil)
  - Include patient health and sampling city information
  - Add in subtype reference sequences and make a phylogenetic tree for quick evaluation
  - Download the sequences as 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!)

Los Alamos
NATIONAL LABORATORY

We will search for country = Brazil (BR)

We will search for complete genomes.





Results for HIV-1 complete genomes from Brazil

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

Displaying **1 - 100** of **208** sequences found:

Note: **11** 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 100 records per page

Click on field name to sort in ascending or descending order

#	Select	Patient (id)	Code	Accession	Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1		Blast	BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2		Blast	BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3		Blast	BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990		9602	HIV-1
4		Blast	BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990		9602	HIV-1
		ast	BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993		8968	HIV-1
		ast	BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
		ast	BR004c(5320)	AF286228	98BR004	C	BRAZIL	1998		9016	HIV-1
		ast	BZ167(10007)	AY173956	BZ167	B	BRAZIL	1989		8940	HIV-1
		ast	BZ126(3090)	AY173957	BZ126	F1	BRAZIL	1989		9030	HIV-1
		ast	BZ163(4569)	AY173958	BZ163	F1	BRAZIL	1989		8991	HIV-1
		ast	RJ1(10882)	AY455778	99UFRJ_1	29_BF	BRAZIL	1999		8767	HIV-1
		ast	BR97(10885)	AY455779	94BR_RJ_97	BF	BRAZIL	1994		8962	HIV-1
		ast	RJ2(10886)	AY455780	99UFRJ_2	BF	BRAZIL	1999		9045	HIV-1
		ast	BR41(15452)	AY455781	94BR_RJ_41	BF1	BRAZIL	1994		8864	HIV-1
		ast	RJ16(10887)	AY455782	99UFRJ_16	46_BF	BRAZIL	1999		9002	HIV-1
		ast	RJ9(10888)	AY455783	99UFRJ_9	BF	BRAZIL	1999		9040	HIV-1
		ast	BR59(10884)	AY455784	94BR_RJ_59	BF	BRAZIL	1994		8898	HIV-1
18		Blast	BR58(10883)	AY455785	94UFRJ_58	BF	BRAZIL	1994		8898	HIV-1
19		Blast		AY727522	04BR013	C	BRAZIL	2004		9050	HIV-1
20		Blast		AY727523	04BR021	C	BRAZIL	2004		8958	HIV-1
21		Blast		AY727524	04BR038	C	BRAZIL	2004		9042	HIV-1
22		Blast		AY727525	04BR073	C	BRAZIL	2004		8997	HIV-1
23		Blast		AY727526	04BR137	31_BC	BRAZIL	2004		8795	HIV-1
24		Blast		AY727527	04BR142	31_BC	BRAZIL	2004		9057	HIV-1
25		Blast	107(10943)	AY771588	BREPM107	BF1	BRAZIL	1999		8798	HIV-1
26		Blast	108(10944)	AY771589	BREPM108	BF1	BRAZIL	1999		9058	HIV-1
27		Blast	269(10946)	AY771591	BREPM269	BF	BRAZIL	1999		8908	HIV-1
28		Blast	275(10947)	AY771592	BREPM275	BF	BRAZIL	1999		8671	HIV-1
29		Blast	278(10948)	AY771593	BREPM278	BF	BRAZIL	1999		8948	HIV-1

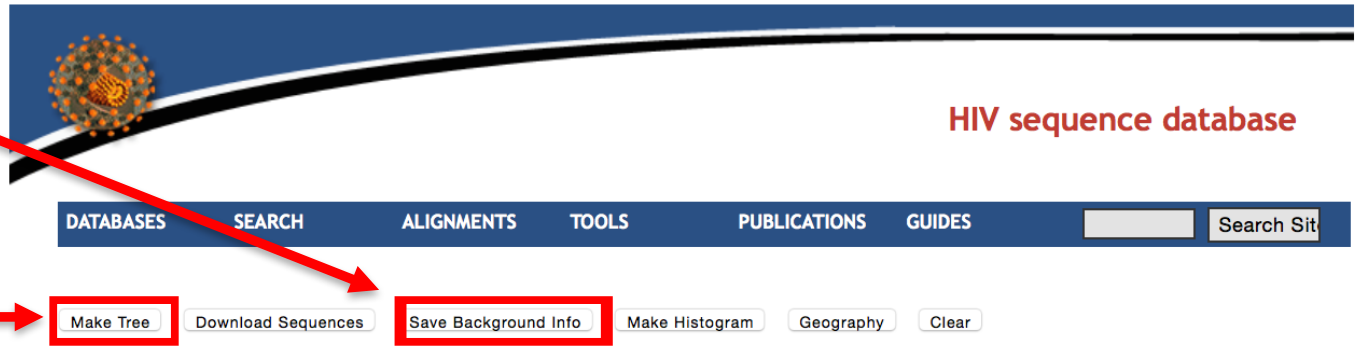
Choose "One sequence/patient" to remove very similar sequences (only available if a region is selected)

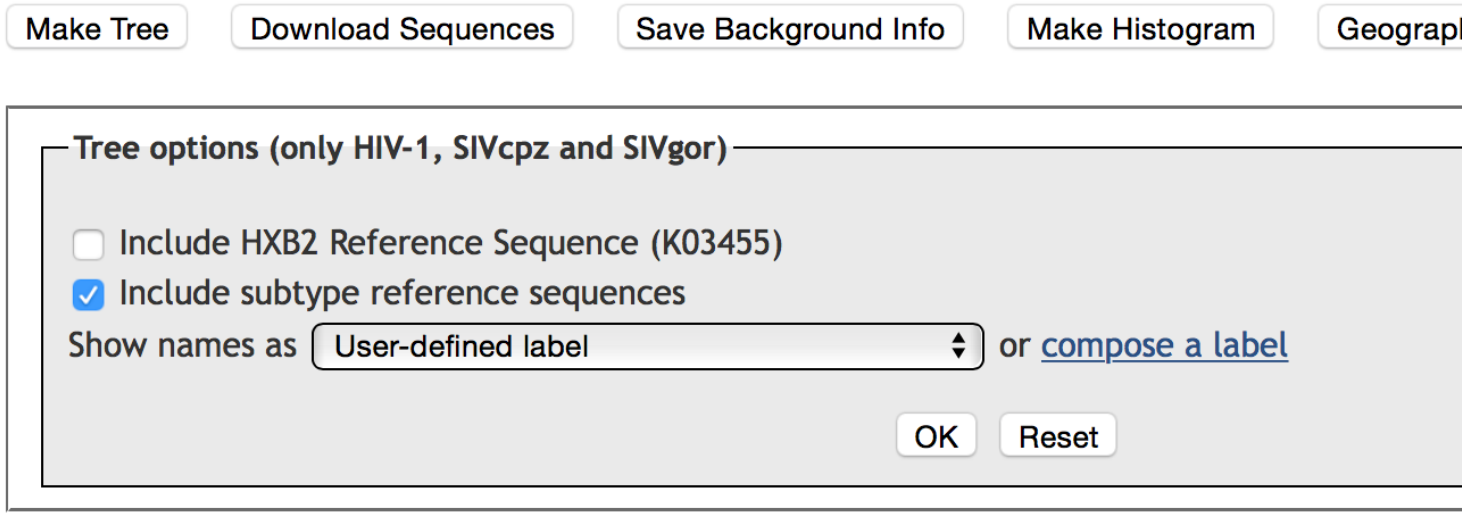
Los Alamos NATIONAL LABORATORY

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.





Displaying **1 - 100** of **391** sequences found:

Note: **14** problematic sequences were removed from this result. Click here to repeat search to

Select all | Unselect all | Invert selection | Show all | One sequence/patient | Select record

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Pa He
1	Blast	BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990	
2	Blast	BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990	

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.

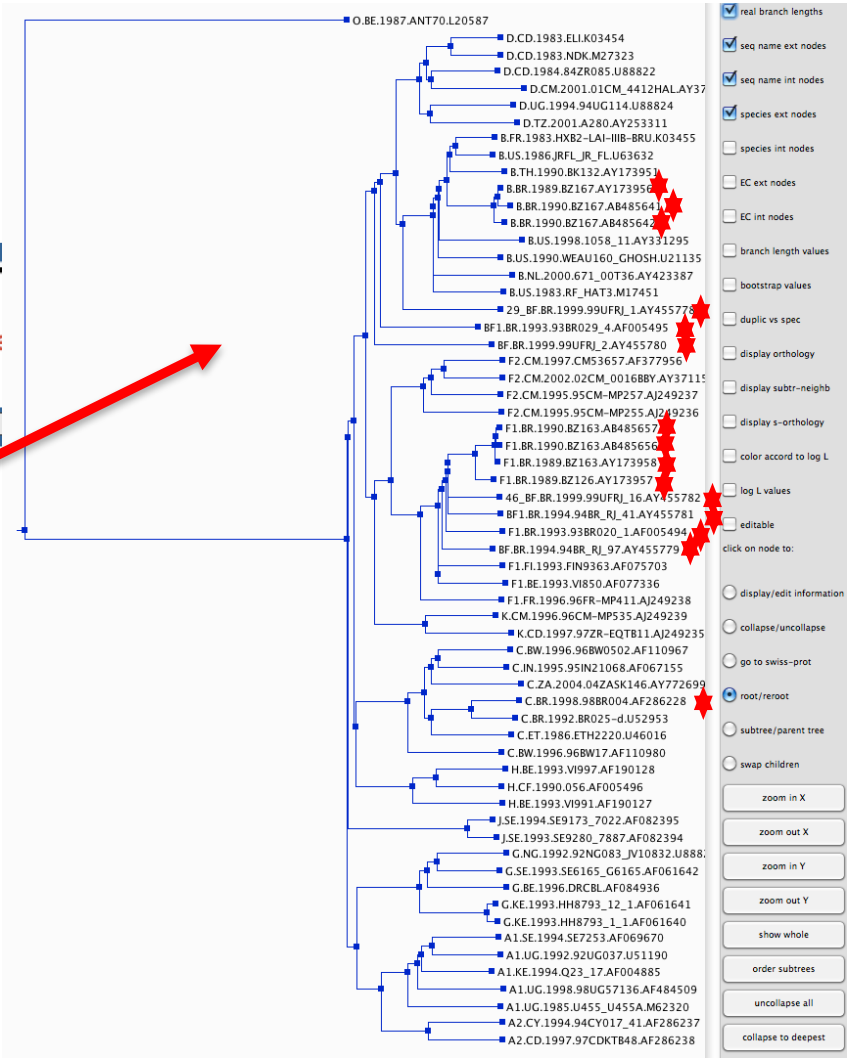
Alternatively, leave blank for midpoint rooting

Optional mailback, and tree title

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

Our Brazilian sequences

ATV is a java-based view for a quick look and manipulation; cannot save/print



Save alignment, use BioEdit or Aliview to view/adjust. Use FigTree or other programs to view Newick tree files.

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.

Los Alamos NATIONAL LABORATORY

HIV sequence database

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 Pheno
- View Phenogram

Radial:

- Download radial
- View radial (unr

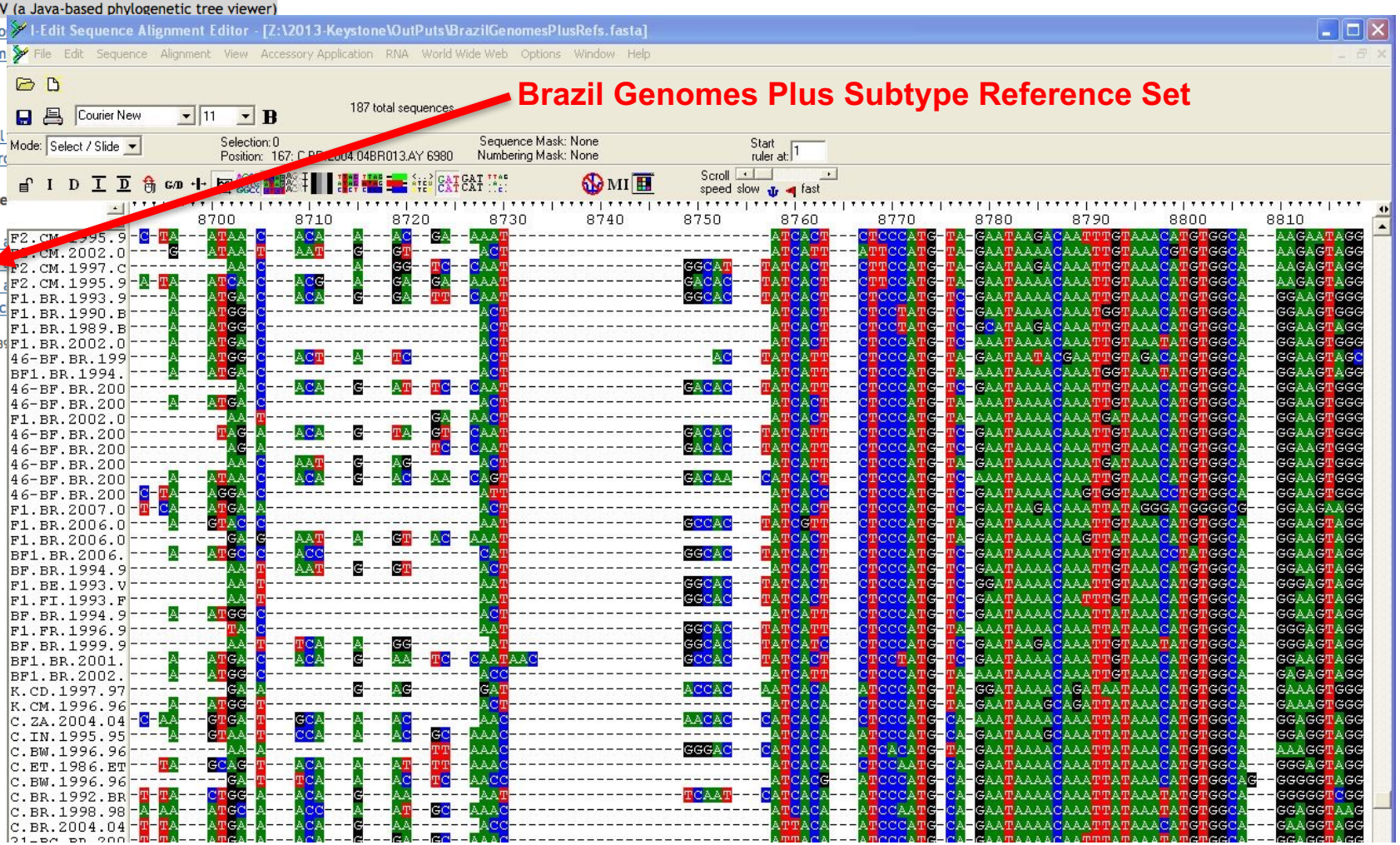
Alignment used for tre

- Download fasta
- Download fast
- Download fasta
- Download Newic

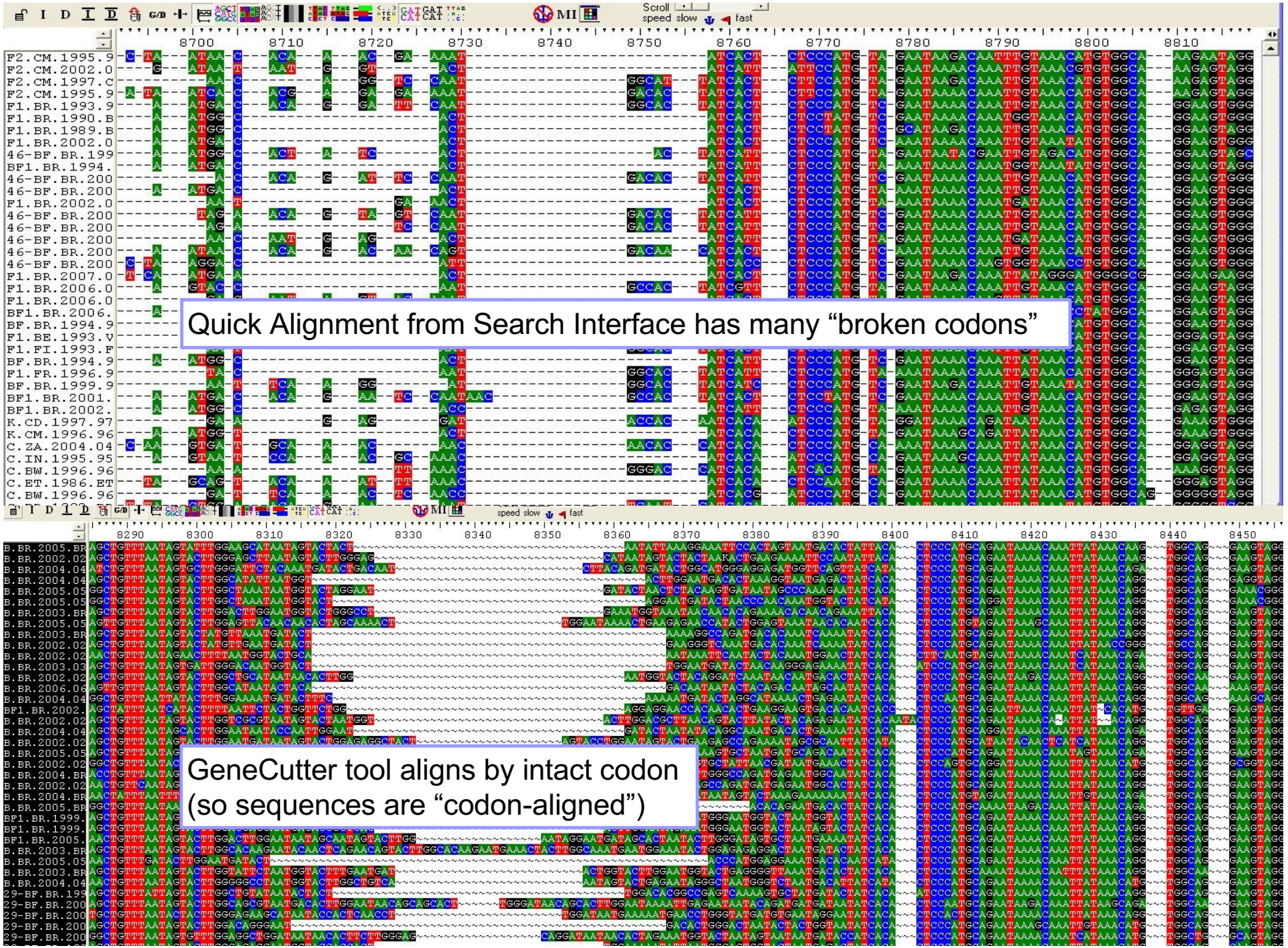
last modified: Thu May 7 07:3

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

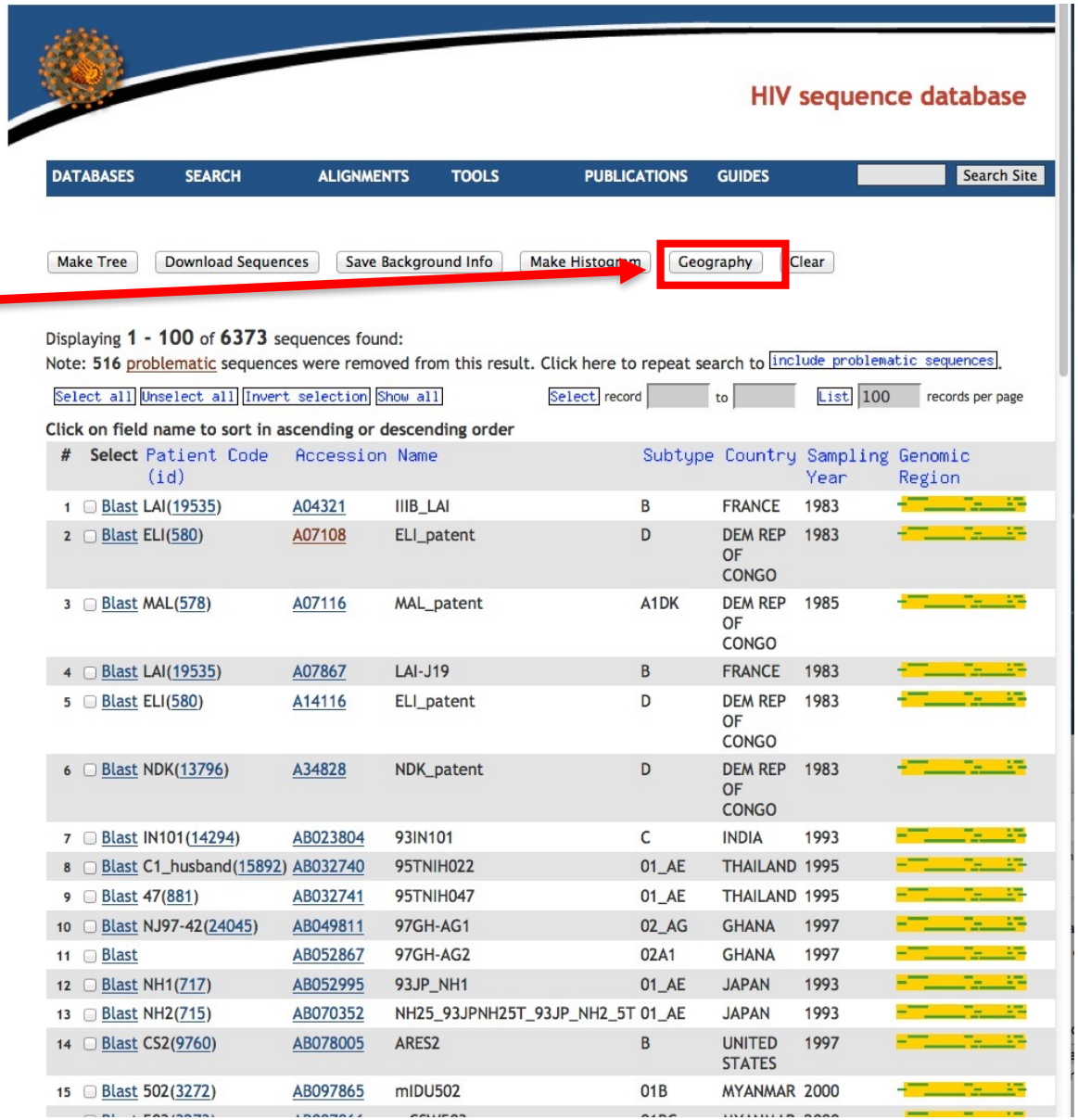
Sending the alignment through GeneCutter or HIV-Align will often improve tree quality.



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”)

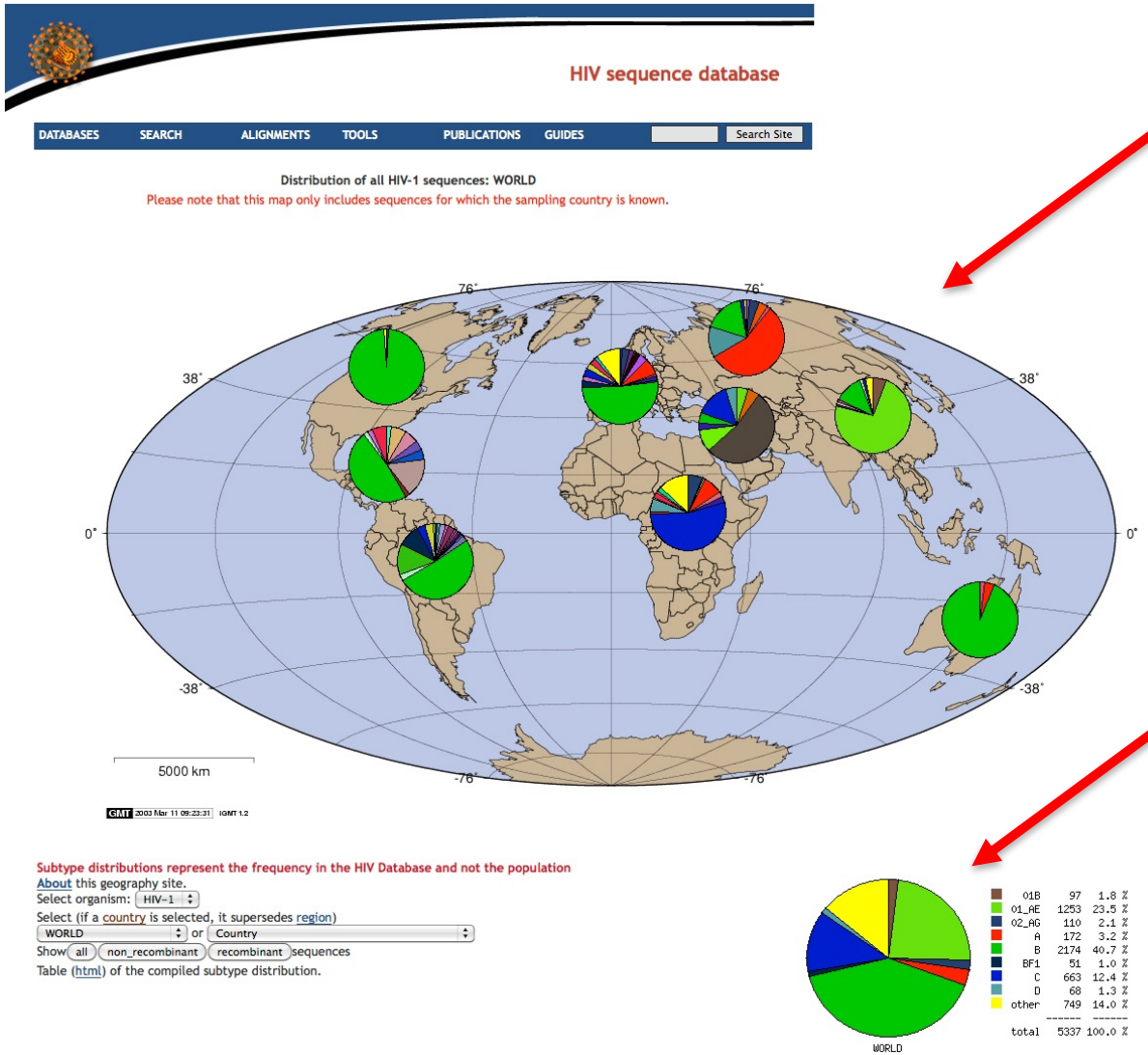
New search: all complete genomes; then look at geographic and subtype distribution of the sequences



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

Los Alamos NATIONAL LABORATORY

Geography output



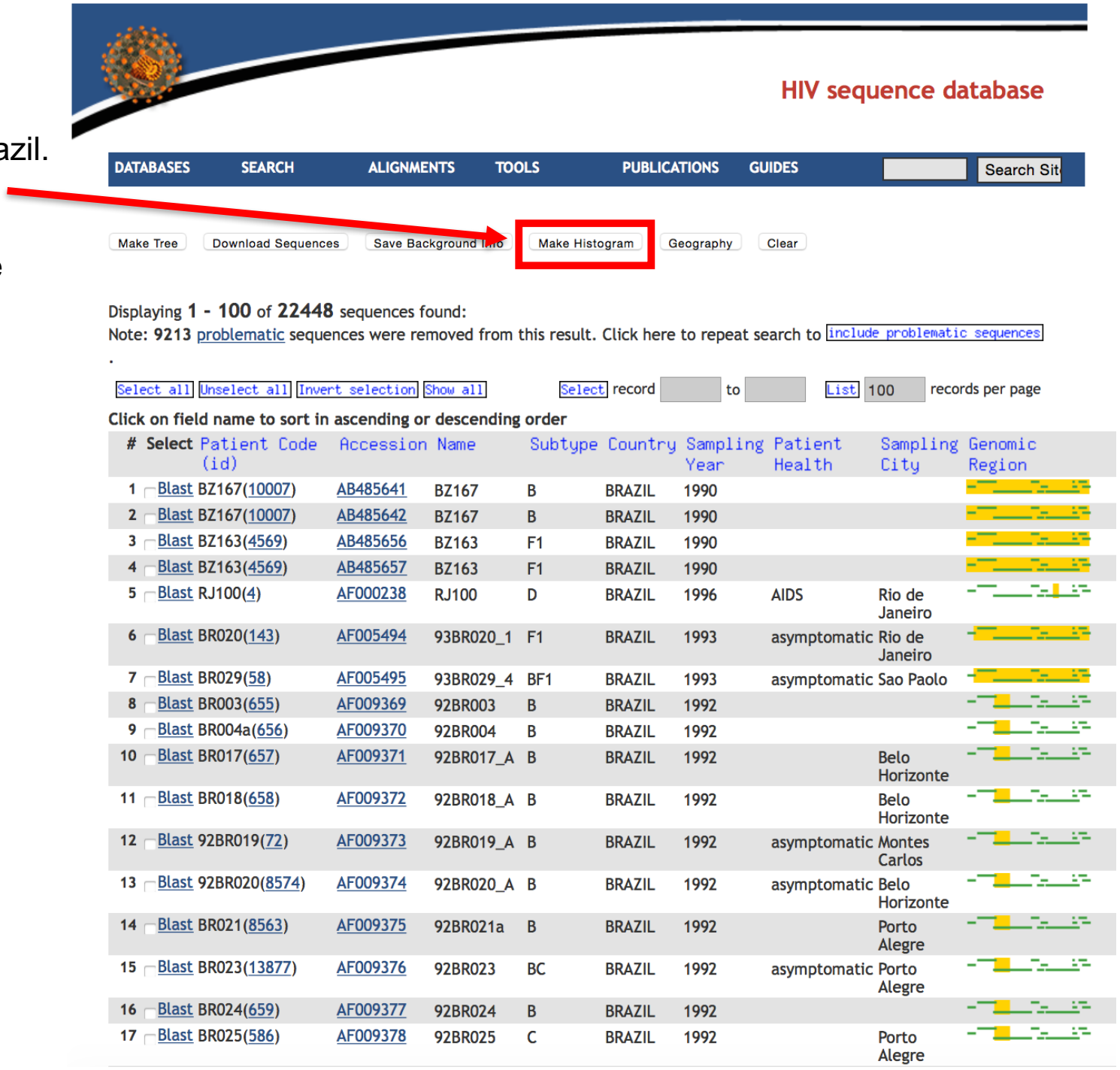
Each continent's pie chart is clickable to "zoom in" on that continent.

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.

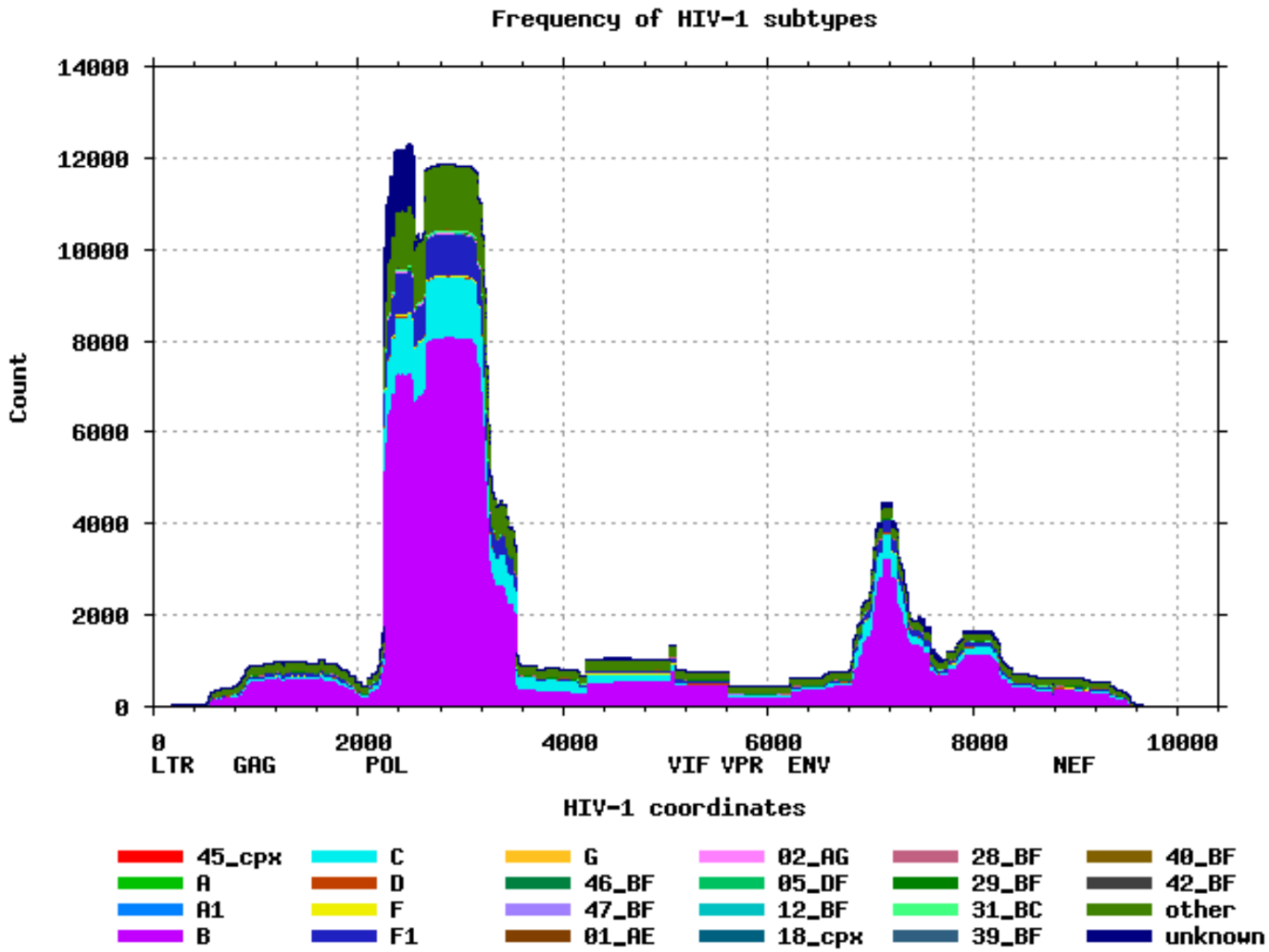
Los Alamos NATIONAL LABORATORY

New search: all sequences from Brazil. Then look at the distribution of the sequences over the genome



#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Patient Health	Sampling City	Genomic Region
1	Blast	BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990			
2	Blast	BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990			
3	Blast	BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990			
4	Blast	BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990			
5	Blast	RJ100(4)	AF000238	RJ100	D	BRAZIL	1996	AIDS	Rio de Janeiro	
6	Blast	BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993	asymptomatic	Rio de Janeiro	
7	Blast	BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993	asymptomatic	Sao Paolo	
8	Blast	BR003(655)	AF009369	92BR003	B	BRAZIL	1992			
9	Blast	BR004a(656)	AF009370	92BR004	B	BRAZIL	1992			
10	Blast	BR017(657)	AF009371	92BR017_A	B	BRAZIL	1992		Belo Horizonte	
11	Blast	BR018(658)	AF009372	92BR018_A	B	BRAZIL	1992		Belo Horizonte	
12	Blast	92BR019(72)	AF009373	92BR019_A	B	BRAZIL	1992	asymptomatic	Montes Carlos	
13	Blast	92BR020(8574)	AF009374	92BR020_A	B	BRAZIL	1992	asymptomatic	Belo Horizonte	
14	Blast	BR021(8563)	AF009375	92BR021a	B	BRAZIL	1992		Porto Alegre	
15	Blast	BR023(13877)	AF009376	92BR023	BC	BRAZIL	1992	asymptomatic	Porto Alegre	
16	Blast	BR024(659)	AF009377	92BR024	B	BRAZIL	1992			
17	Blast	BR025(586)	AF009378	92BR025	C	BRAZIL	1992		Porto Alegre	

Histogram output



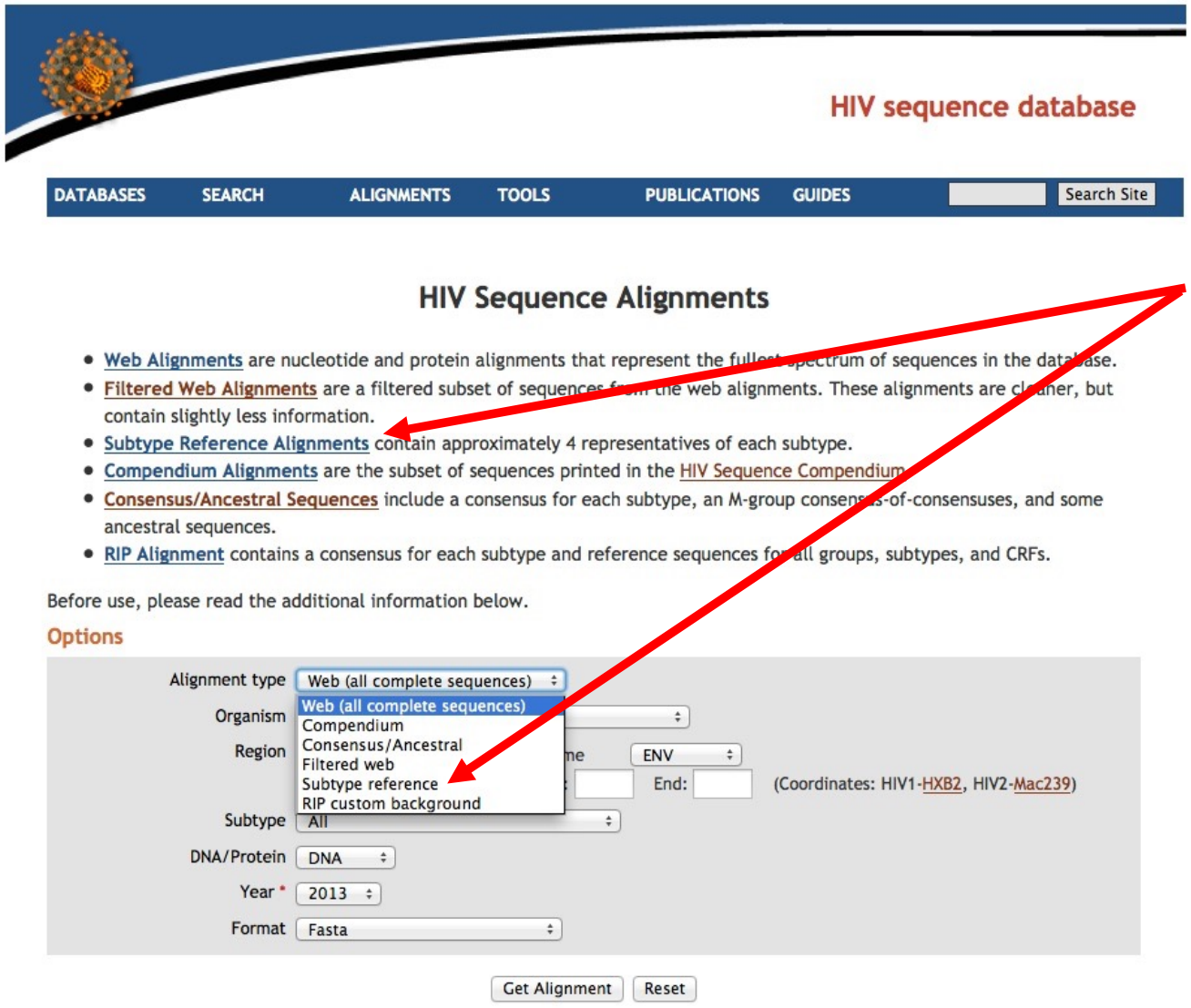
This histogram shows the distribution of sequences from your query across the entire HIV-1 genome. At each position across the genome, the number of sequences overlapping with that position is plotted. The colors represent different subtypes.

Los Alamos
NATIONAL LABORATORY

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 consensuses and “maximum likelihood ancestors” are available for reagent production
- Special interest alignments
  - Sequence sets (“authors’ alignments”) of particular research interest
  - Suggestions and additions welcome!

Los Alamos
NATIONAL LABORATORY



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

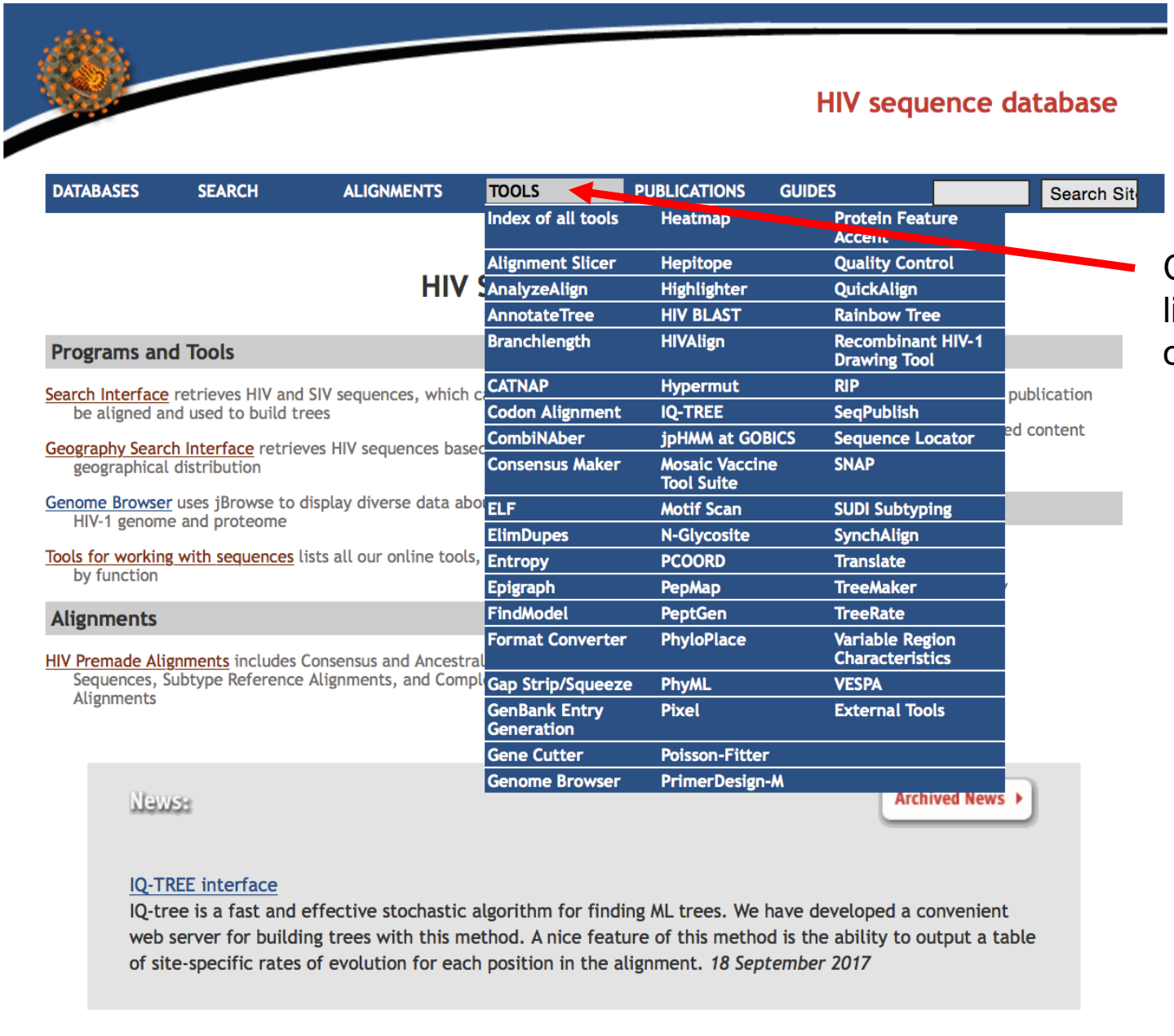
Subtype Reference = 4 representatives of each subtype, plus one of each Circulating-inter-subtype-Recombinant-Form (CRF) of the M group, plus 4 O group, N group, P group, and SIV-CPZ

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.

Los Alamos NATIONAL LABORATORY

The HIV database sequence analysis tool set



Click top level to link to full page of tools

Los Alamos NATIONAL LABORATORY

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

Epilign (QuickAlign) 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

More below!

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.

Los Alamos NATIONAL LABORATORY

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.

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

Los Alamos NATIONAL LABORATORY

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.
- **Phylogenetics**
  - **TreeMaker** generates a neighbor-joining phylogenetic tree.
  - **PhyML** generates a maximum likelihood phylogenetic tree.
  - **IQ-TREE** generates a fast 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

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.
- **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

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 specification**
  - **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.
- **Hypermutation/recombination/contamination tools**
  - **Hypermut** 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.

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)

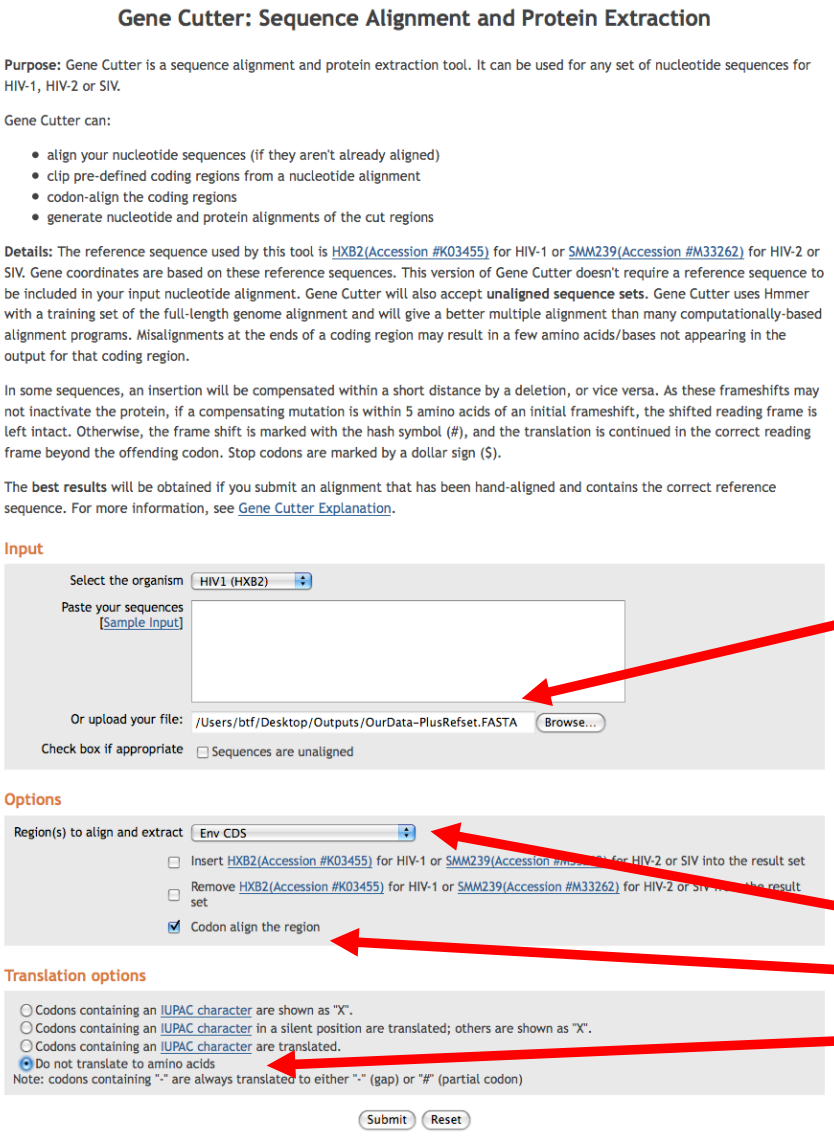
Including HXB2 as a reference may improve results

Does *not* address hypermutation or recombination

(see “Hypermut” or “RIP”)

Los Alamos
NATIONAL LABORATORY

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

Los Alamos NATIONAL LABORATORY

Gene Cutter Results

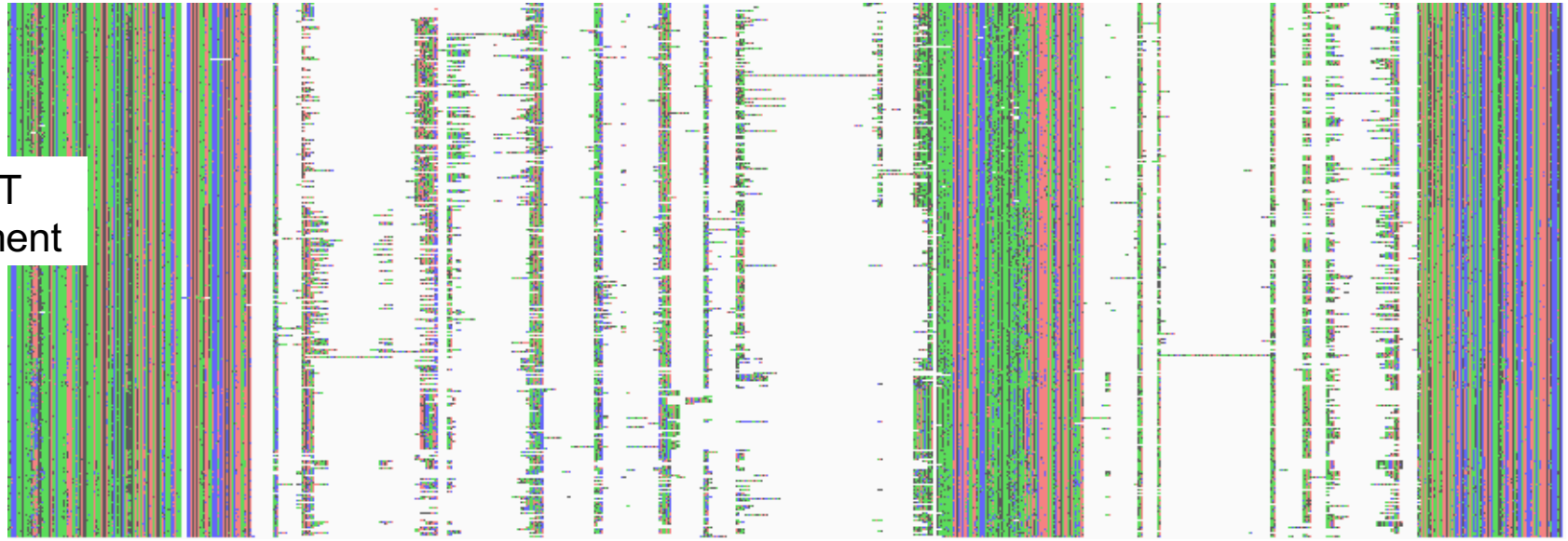
- Results are stored on our server
  - An HTML link is e-mailed to the user when the run is complete
  - For this workshop, we will provide example files.

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

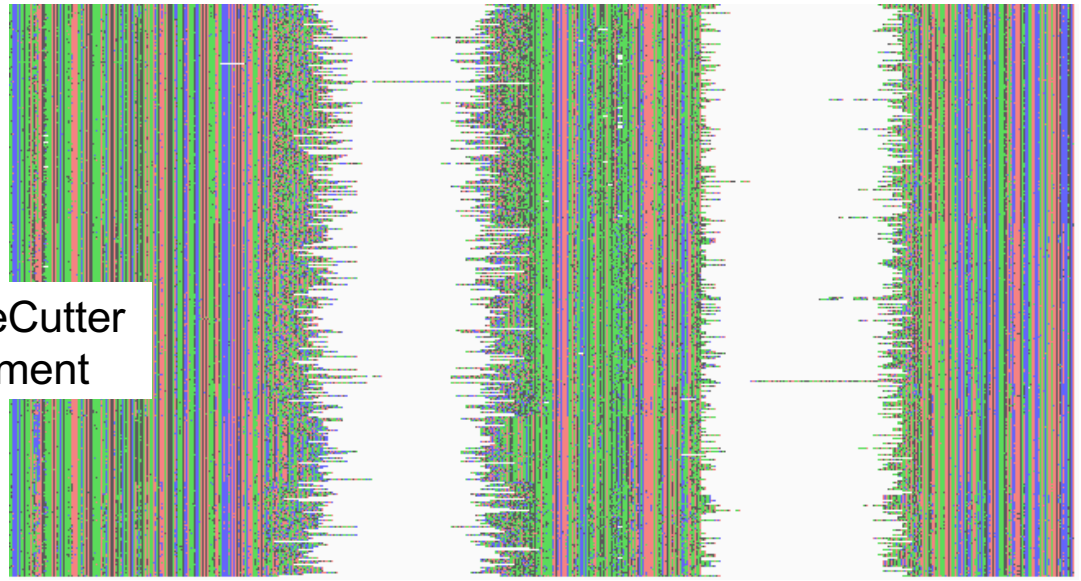
MAFFT alignment



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

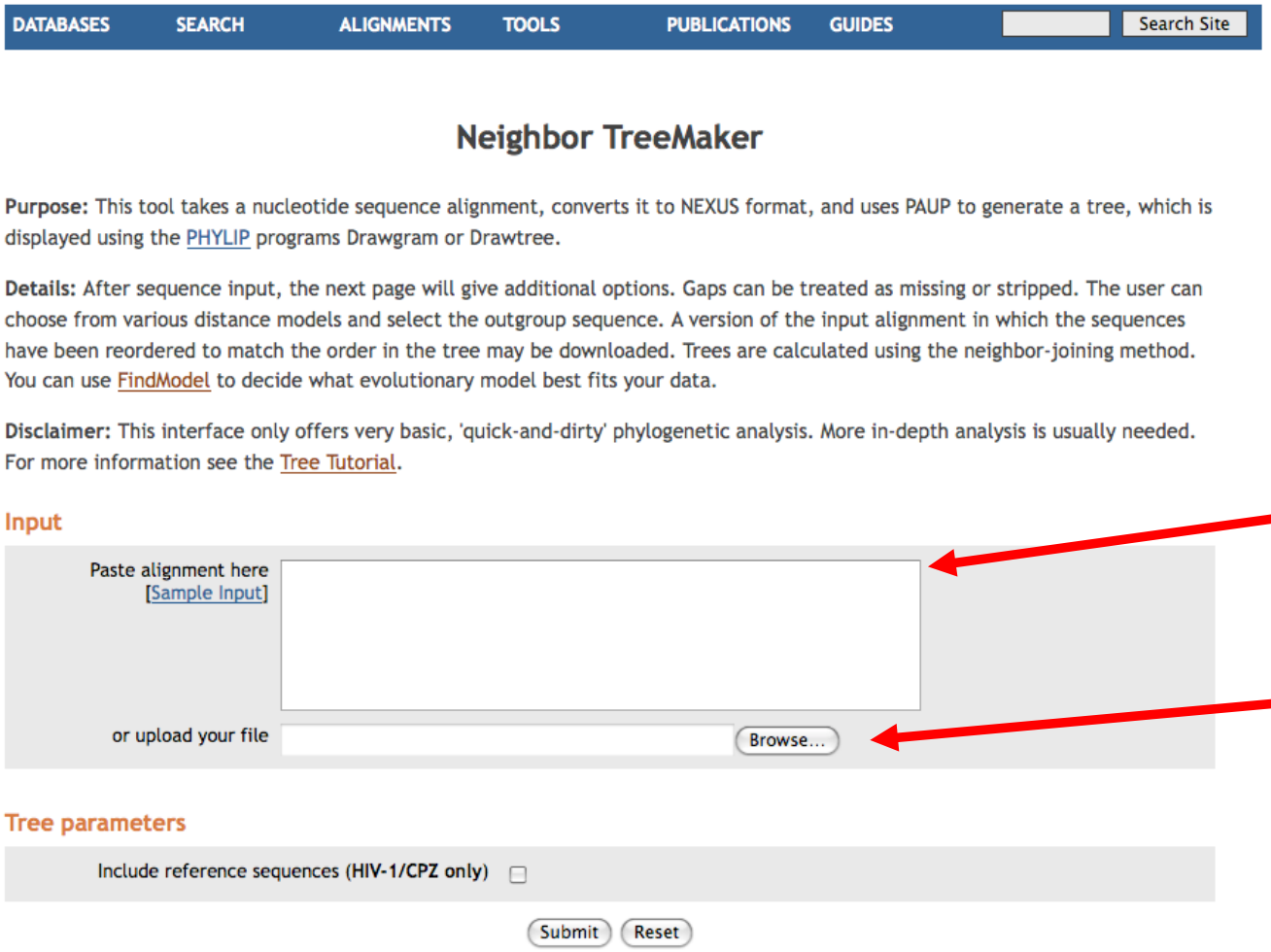
Check for phylogenetic relatives:

- 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 IQ-Tree Maximum Likelihood interfaces are also available

http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html
https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html

Los Alamos
NATIONAL LABORATORY

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

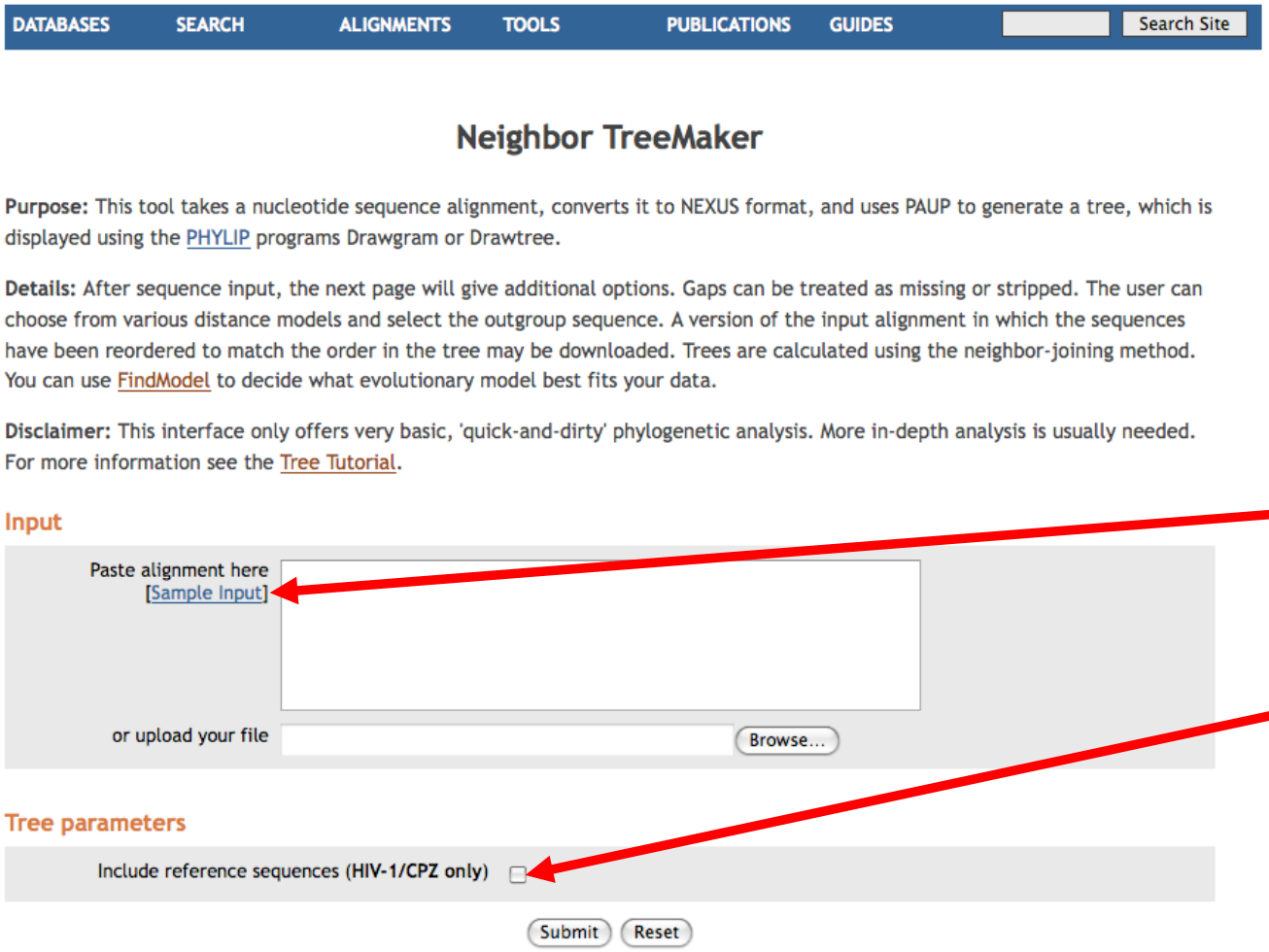


Paste or type a DNA **alignment** here. Any organism.

OR upload an alignment file here.

Los Alamos NATIONAL LABORATORY

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

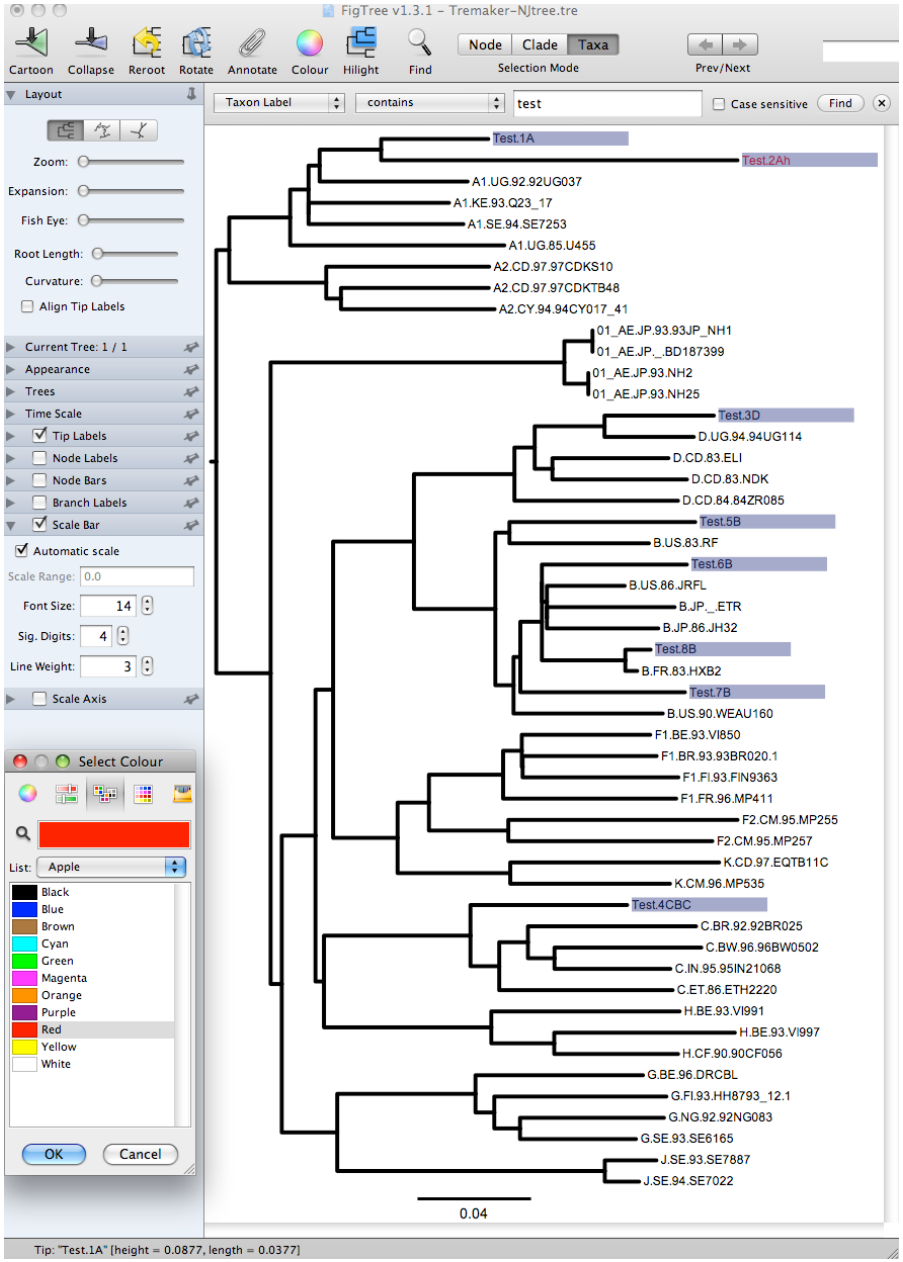


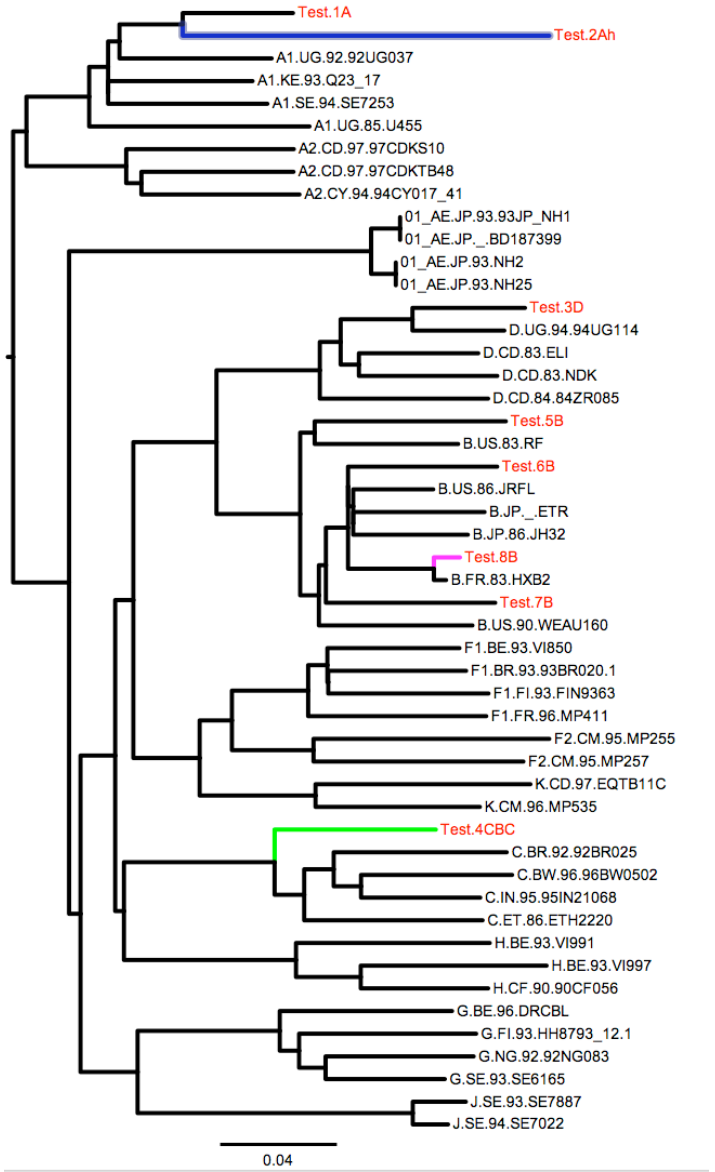
For this exercise use the sample input.

Include the reference sequences.

Los Alamos NATIONAL LABORATORY

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



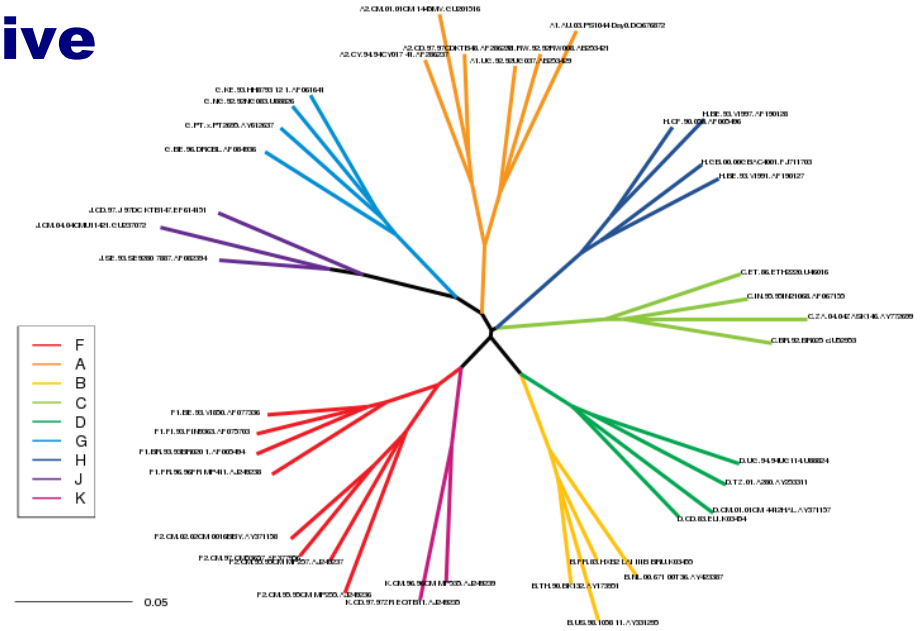


Los Alamos
NATIONAL LABORATORY

Making ~~decorative~~ informative trees for publication

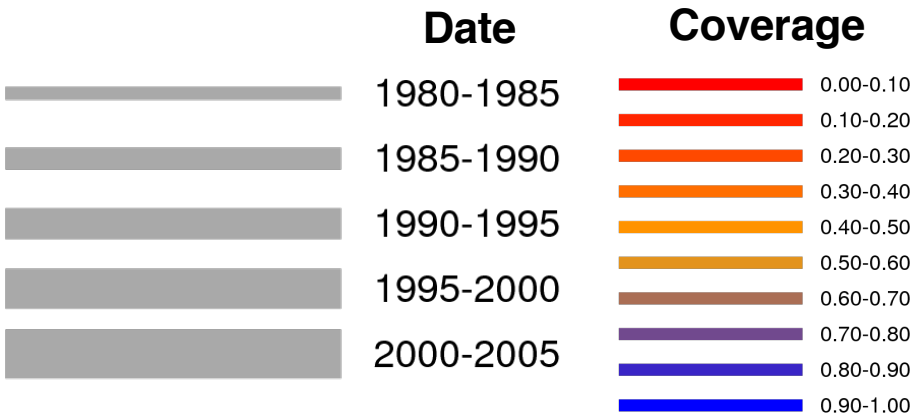
Rainbow Tree

- 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)

Los Alamos
NATIONAL LABORATORY

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 [Browse...]

-- OR --

Retrieve a region by its coordinates

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

Region [Complete]

Retrieve ◉ Nucleotide or ○ protein output

☐ include surrounding region

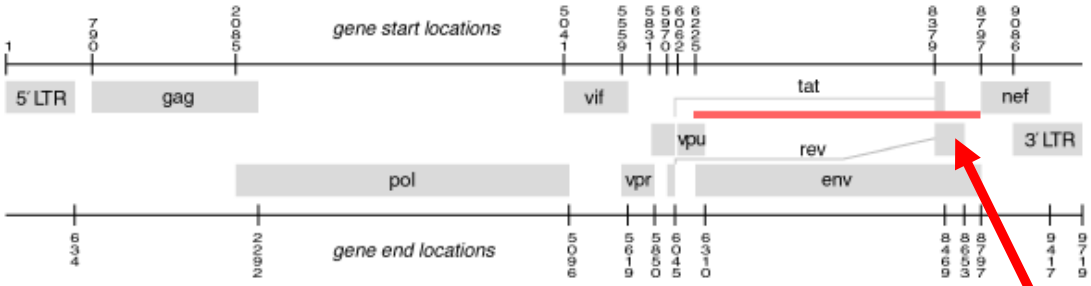
[Submit] [Reset]

Paste or type a DNA or protein sequence here.

OR enter numeric coordinates here.

Los Alamos NATIONAL LABORATORY

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

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
XESEGEISALVEMGVEMGHHAPWDIDDL*				
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).				
MRVKEKYQHLWRWGWKWGTMLLGILMICSATEKLWVTVYYGVPVWKEATTTLFCASDAKA YDTEVHNVWATHACVPTDPNPQEVVLVNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCV KLTPLCVSLKCTDLGNATNTNSSNTNSSSGEMMMEKGEIKNCSFNISTSIRGKVQKEYAF FYKLDIIPIDNDTTSYTLTSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNNKTFNG TGPCTNVSTVQCTHGIRPVVSTKLLLNGSLAEEEVVIRSANFTDNAKTIIVQLNQSVEIN CTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAHCNISRAKWNATLKQIASKLREQFGN NKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFNSTWFNSTWSTEGSNNTEGSDTIT LPCRIKQFINMWQEVGKAMYAPPISGQIRCSSNITGLLLTRDGGNNNNGSEIFRPGGGDM RDNWRSESYKYKVVKIEPLGVAPTKAKRRVVQREKRAVGIGALFLGFLGAAGSTMGARSM TLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARILAVERYLKDQQLLGI WGCSEKLICTTAVPWNASWSNKSLEQIWNNMTWMEWDREINNNTSLIHSLIEESQNQQEK NEQELVELDK*ASLWNWFNITN*LWYIKIFIMIVGGLVGLRIVFAVLSIVNRVRQGYSPL SFQTHLPTPRGPDRPEGIEEEGGERDRDRSIRLVNGSLALIWDDLRSLCLFSYHRLRDLL LIVTRIVELLGRRGWEALKYWWNLLQYWSQELKNSAVSLLNATAIAVAEGTDRVIEVVQG ACRAIRHIPRRIRQGLERILL*				
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).				
MRVKEKYQHLWRWGWKWGTMLLGILMICSATEKLWVTVYYGVPVWKEATTTLFCASDAKA YDTEVHNVWATHACVPTDPNPQEVVLVNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCV KLTPLCVSLKCTDLGNATNTNSSNTNSSSGEMMMEKGEIKNCSFNISTSIRGKVQKEYAF FYKLDIIPIDNDTTSYTLTSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNNKTFNG TGPCTNVSTVQCTHGIRPVVSTKLLLNGSLAEEEVVIRSANFTDNAKTIIVQLNQSVEIN CTRPNNNTRKSIRIQRGPGRAFVTIGKIGNMRQAHCNISRAKWNATLKQIASKLREQFGN NKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFNSTWFNSTWSTEGSNNTEGSDTIT LPCRIKQFINMWQEVGKAMYAPPISGQIRCSSNITGLLLTRDGGNNNNGSEIFRPGGGDM RDNWRSESYKYKVVKIEPLGVAPTKAKRRVVQREKR				
gp41	1 → 1038	1549 → 2586	7758 → 8795	1 → 346
AVGIGALFLGFLGAAGSTMGARSMTLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVW GIKQLQARILAVERYLKDQQLLGIWGCSEKLICTTAVPWNASWSNKSLEQIWNNMTWMEW DREINNNTSLIHSLIEESQNQQEKNEQELVELDK*ASLWNWFNITN*LWYIKIFIMIVGG LVGLRIVFAVLSIVNRVRQGYSPLSFQTHLPTPRGPDRPEGIEEEGGERDRDRSIRLVNG SLALIWDDLRSLCLFSYHRLRDLLLIVTRIVELLGRRGWEALKYWWNLLQYWSQELKNSA VSLLNATAIAVAEGTDRVIEVVQGACRAIRHIPRRIRQGLERILL*				
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)
XPPPNPEGTRQARRNRRRRWRERQRQIHSISERILSTYLGRSAEPVPLQLPPLERLTLDC NEDCGTSGTQGVGSPQILVESPTVLESGTKE*				

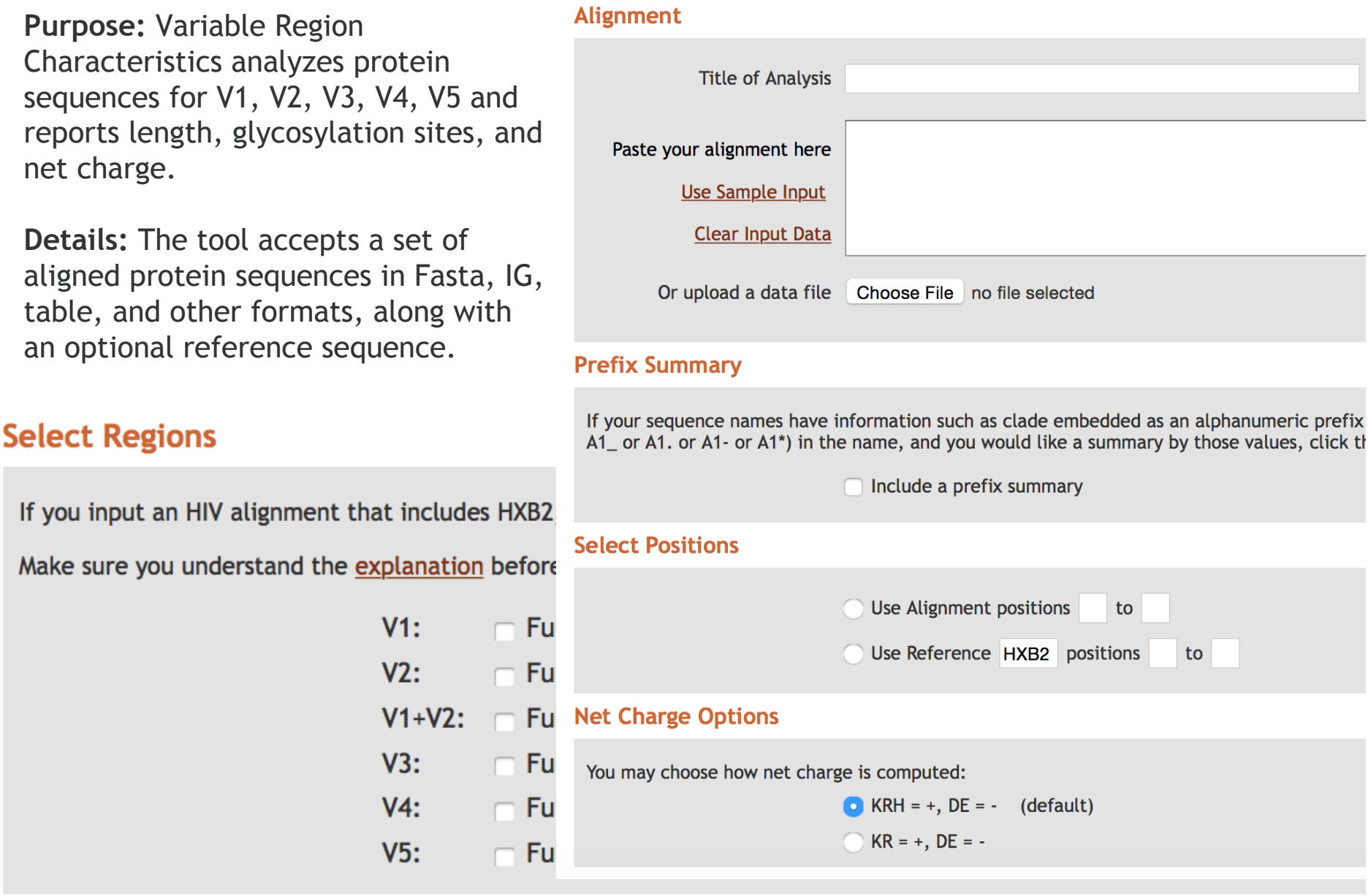
Alignment of the query sequence to HXB2 (Similarity 97.8%):

```
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
```

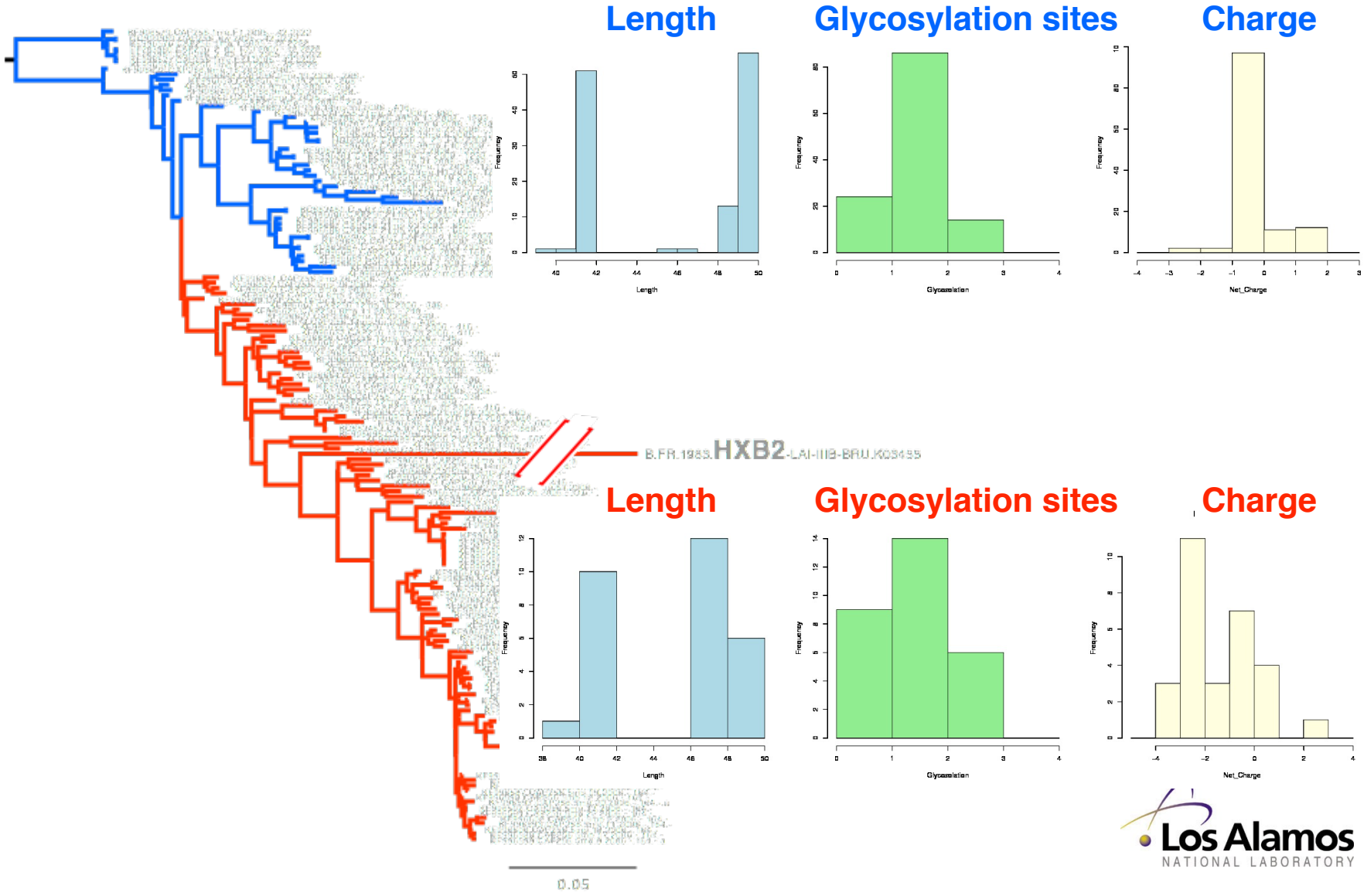
Variable Region Characteristics

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.



Variable Region Characteristics



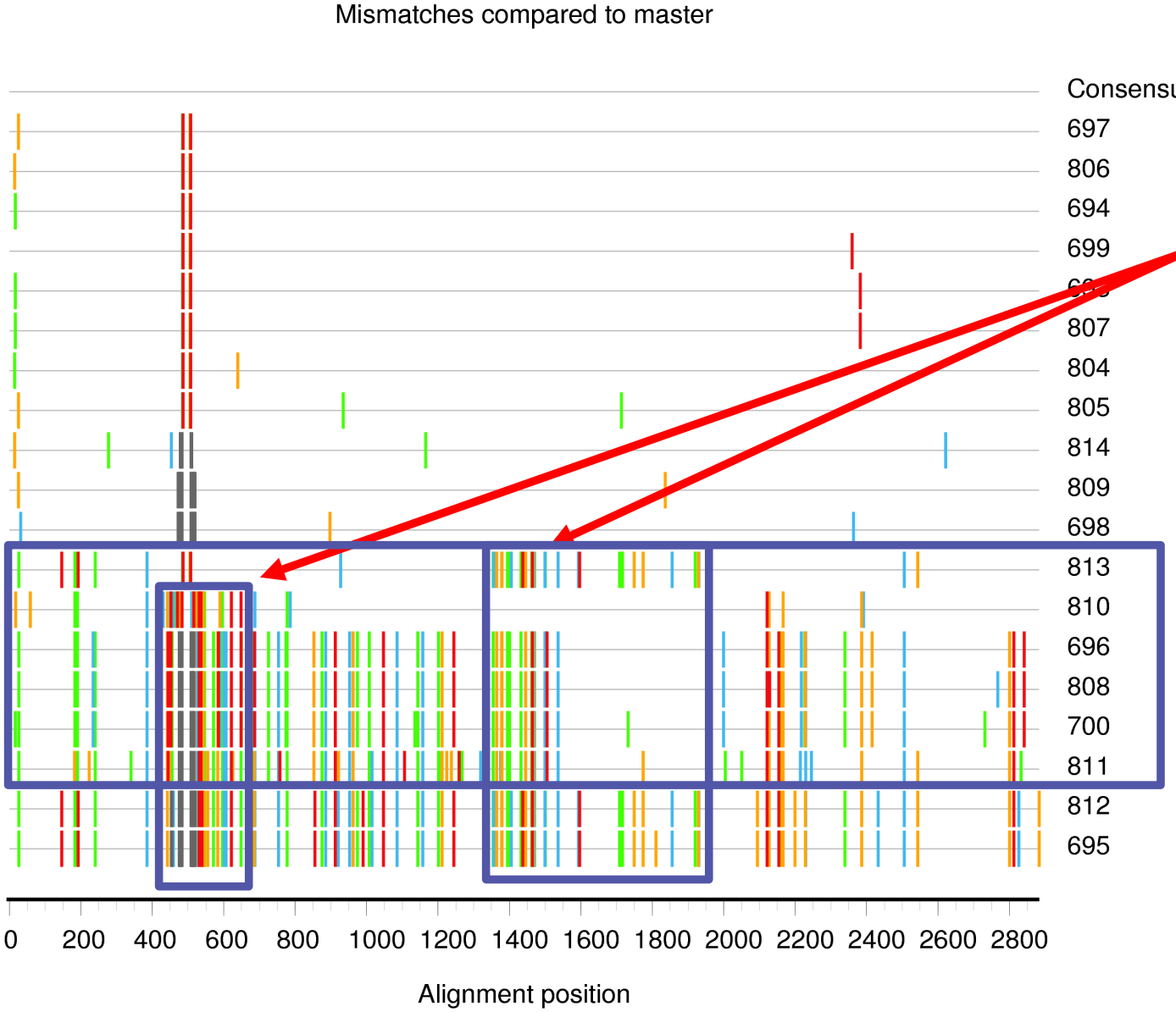
Los Alamos
NATIONAL LABORATORY

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

Los Alamos
NATIONAL LABORATORY

Highlighter sample data

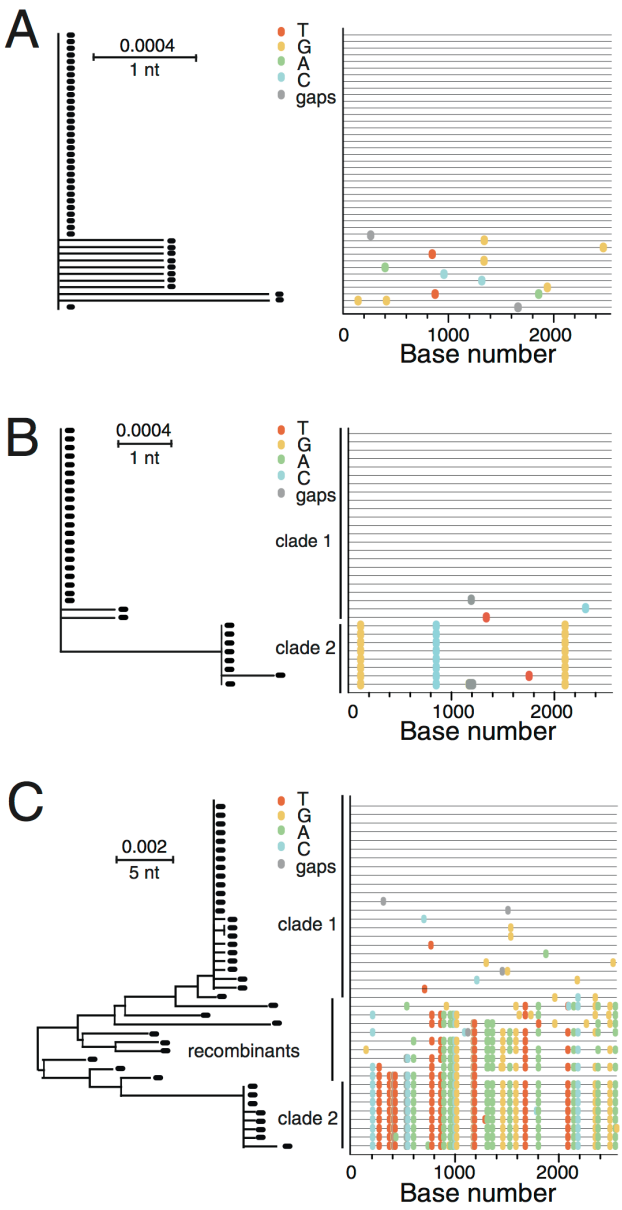


Nonrandom distribution of mutations evident.

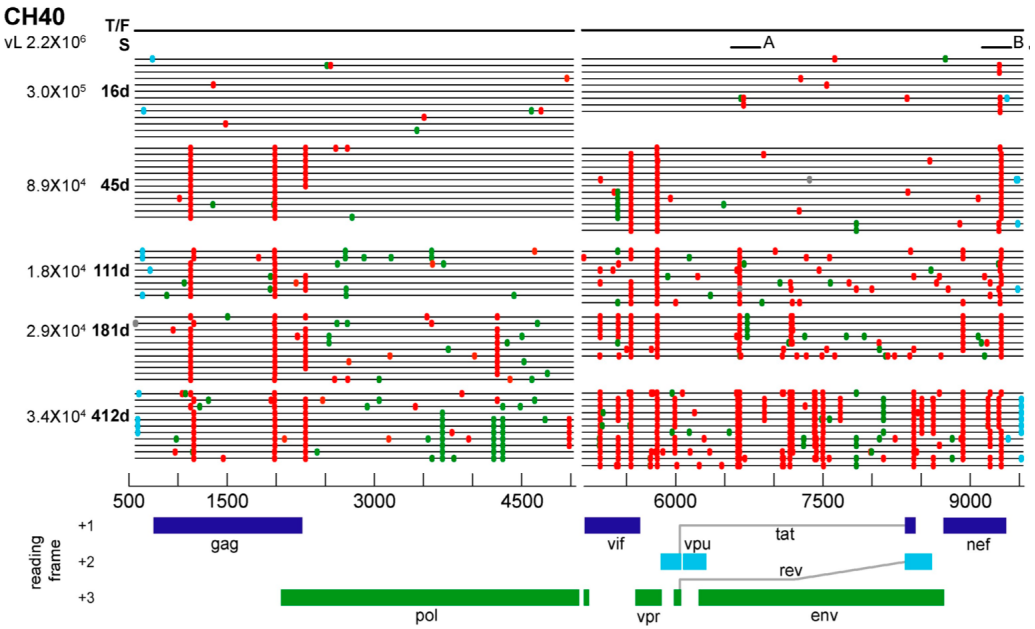
Sample Set is from a likely dual infection, with intra-subtype recombinants present.

Los Alamos NATIONAL LABORATORY

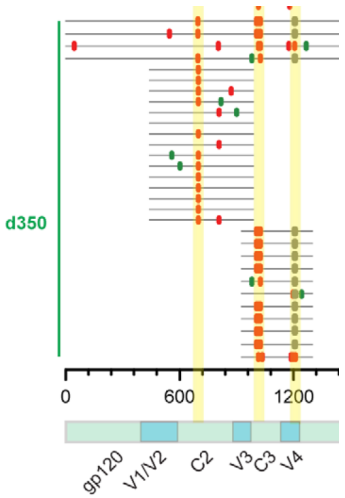
Highlighter examples



B.F. Keele et al. (2008) PNAS 105:7552–7557



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



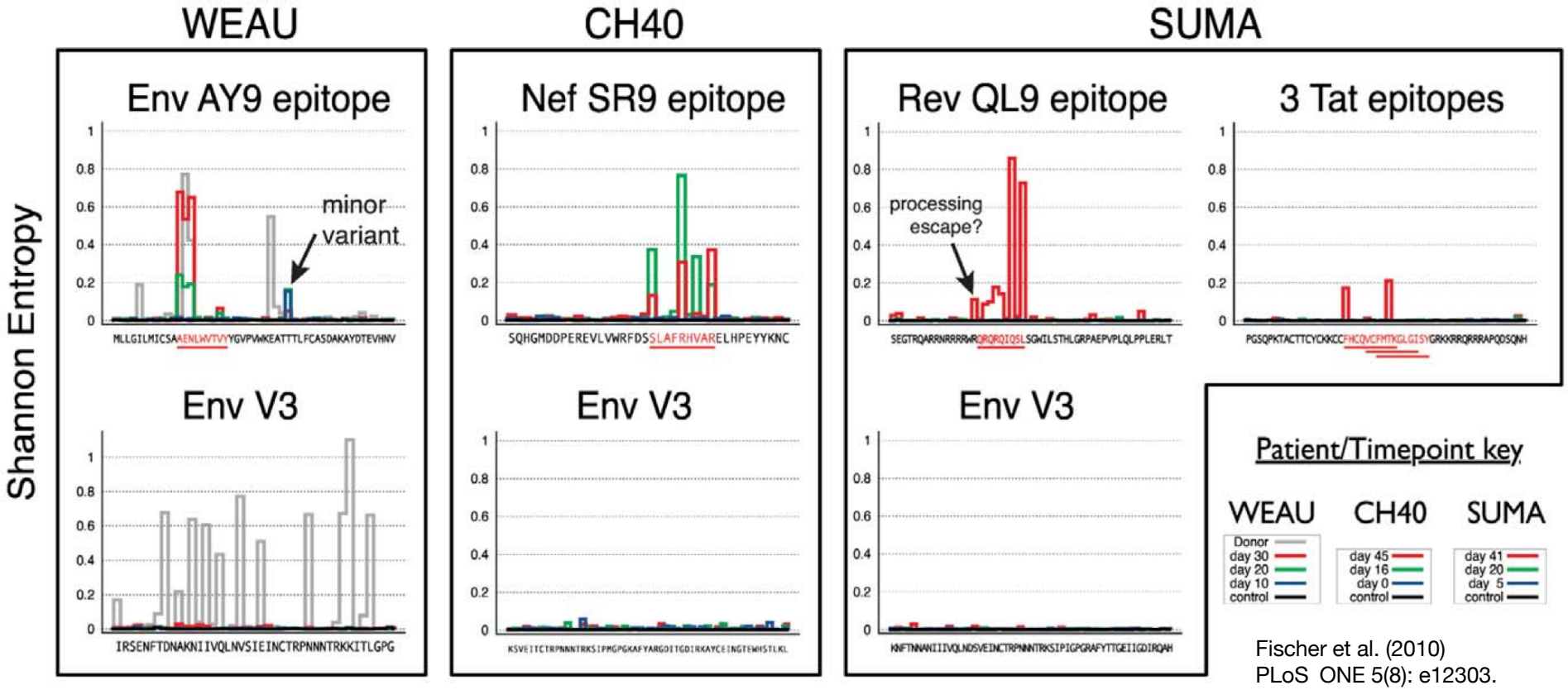
Bar et al. (2012) *PLoS Pathogens* e1002721

- Infection multiplicity
- CTL escape
- Antibody escape

Los Alamos
NATIONAL LABORATORY

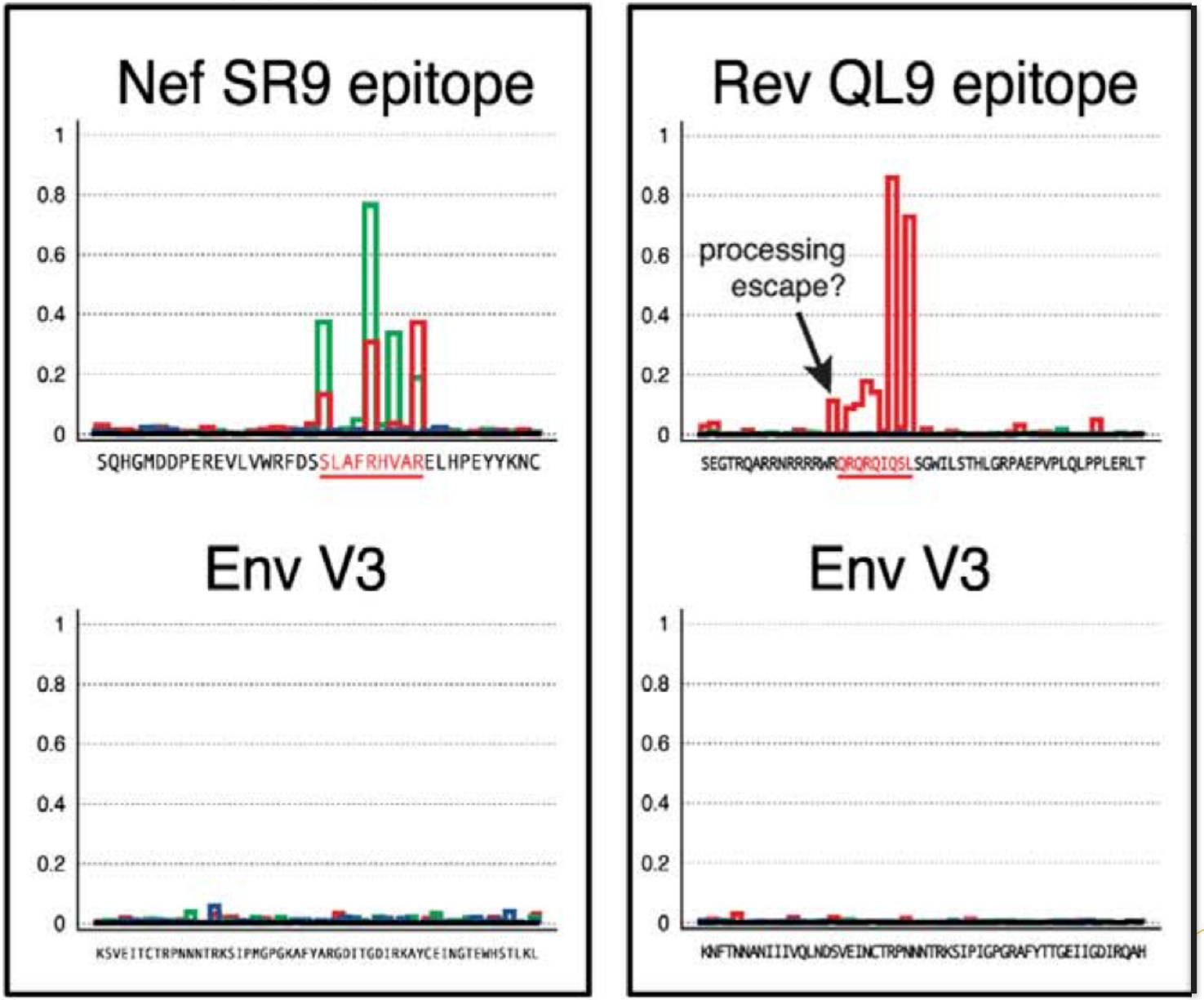
Entropy

- 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)



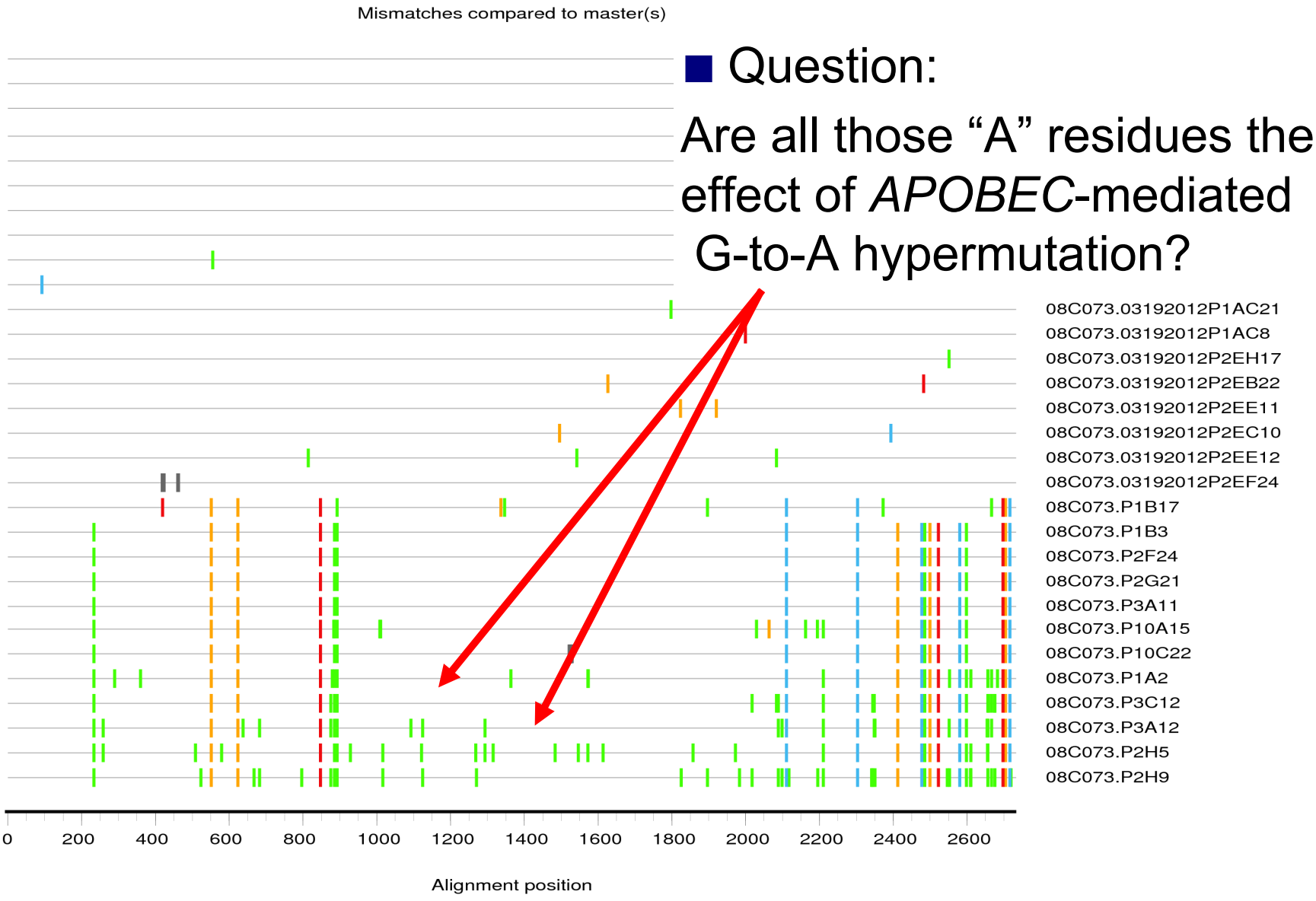
Fischer et al. (2010)
PLoS ONE 5(8): e12303.

Entropy



Los Alamos
NATIONAL LABORATORY

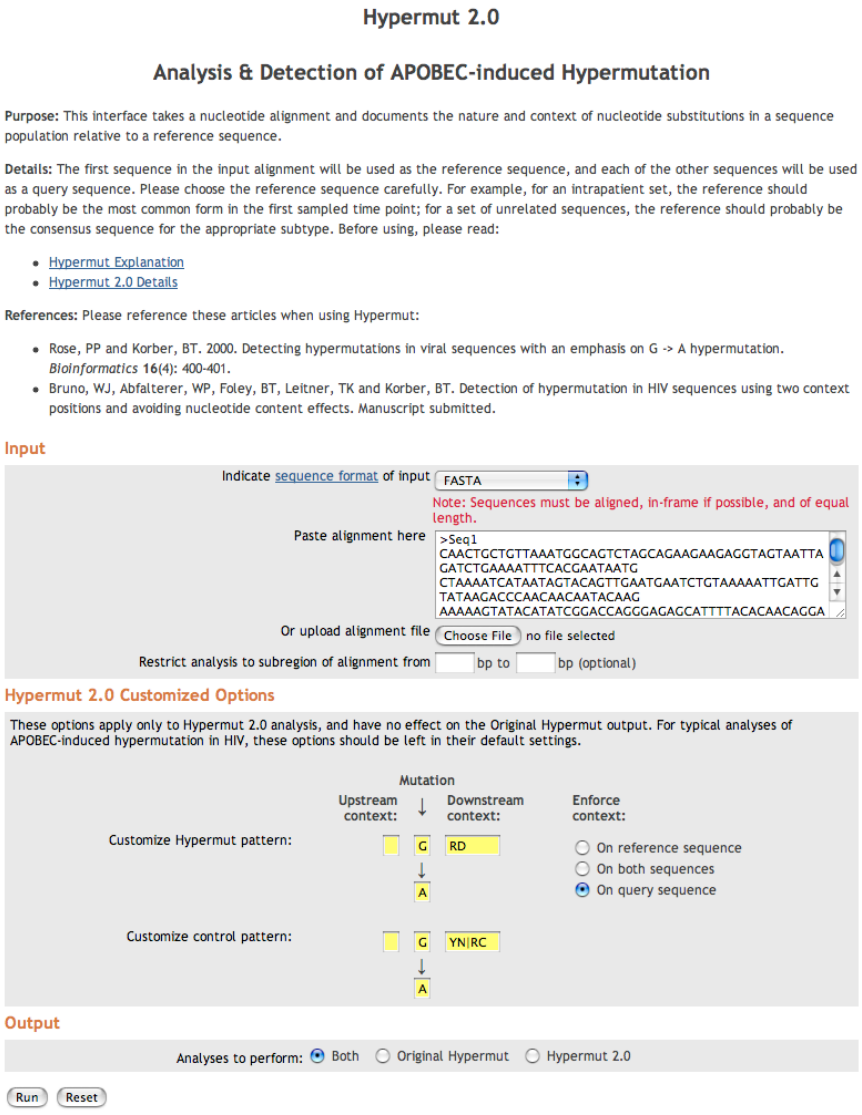
Hypermutation



- Question:

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

Hypermut Tool



Hypermut 2.0

Analysis & Detection of APOBEC-induced Hypermutation

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 intrapatient 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:

- Hypermut Explanation
- Hypermut 2.0 Details

References: Please reference these articles when using Hypermut:

- 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: FASTA

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

Paste alignment here:
>Seq1
CAACTGCTGTTAAATGGCAGTCTAGCAGAAGAAGAGGTAGTAATTA
GATCTGAAAATTTCACGAATAATG
CTAAAATCATAATAGTACAGTTGAATGAATCTGTAAAAATTGATTG
TATAAGACCCAACAACAATACAAG
AAAAAGTATACATATCGGACCAGGGAGAGCATTTTACACAACAGGA

Or upload alignment file: Choose File — no file selected

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

Hypermut 2.0 Customized Options

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

	Upstream context:	Mutation	Downstream context:	Enforce context:
Customize Hypermut pattern:		G → A	RD	On reference sequence / On both sequences / On query sequence
Customize control pattern:		G → A	YN|RC	

Output

Analyses to perform: Both / Original Hypermut / Hypermut 2.0

Run Reset

- 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.

Los Alamos NATIONAL LABORATORY

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: (Select for graphing)	Muts: (Match Sites)	Out of: (Potential Mut Sites)	Controls: (Control Muts)	Out of: (Potential Controls)	Rate Ratio: (A/B)/(C/D)	Fisher Exact P-value: (=P(Muts, Poten. Muts-Muts, Cntrls, Poten. Cntrls-Cntrls))
☑ Seq2	0	71	0	54	undef	1
☑ Seq5	4	69	1	52	3.01	0.282669
☑ Seq7	26	71	1	54	19.77	5.35061e-07
☑ 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.

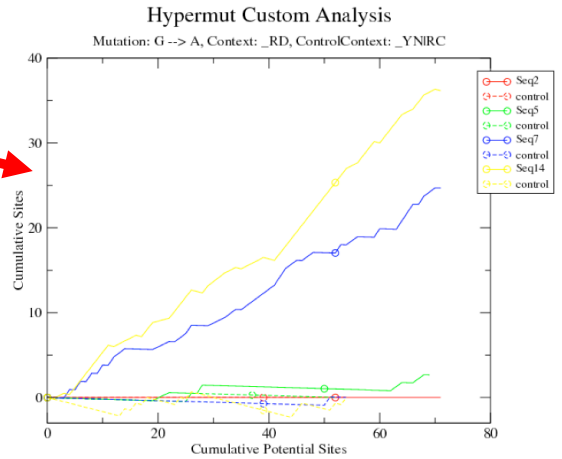
View Sites Along Sequence

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	Sequence names	Ratio	#diffs	perc_Gs	#A->G	#G->A	Dinuc Context: GG	GA	GC	GT	OBSERVED CHANGES
○	○	SEQ2	0/0	1	0.00	0	0	0	0	0	0	TC
○	○	SEQ5	5/2	33	3.97	2	5	2	3	0	0	GA TA GA CT GA GA AT CT AG A- T- A- G- T-

utput:
df version here.
ostscript version here.

Los Alamos NATIONAL LABORATORY

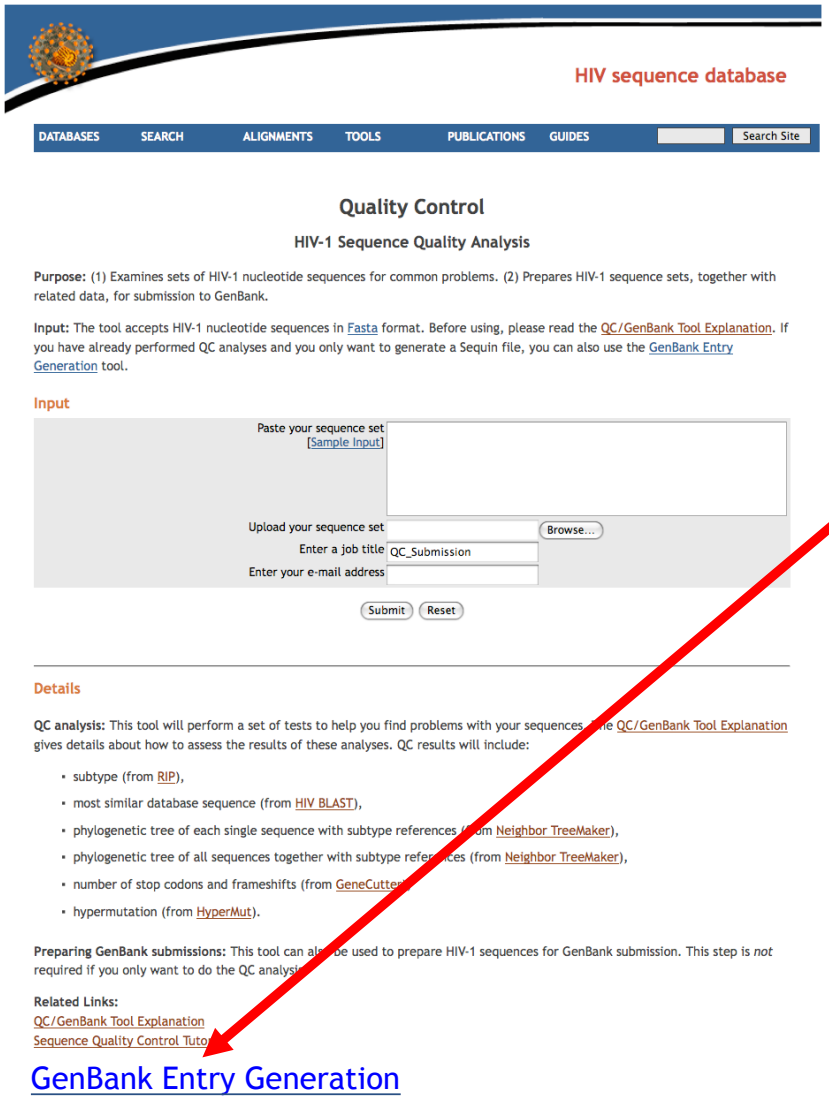
Quality Control Tool

- 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

Los Alamos
NATIONAL LABORATORY

Quality Control Tool

http://www.hiv.lanl.gov/content/sequence/QC/index



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

Los Alamos NATIONAL LABORATORY

Quality Control Tool

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

Click on each result link to see details of each analysis.

Optional: prepare results to submit to GenBank.

Los Alamos NATIONAL LABORATORY

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**

Los Alamos
NATIONAL LABORATORY

Thank you for attending!

Please fill out our evaluation form!

Your comments will help us with funding to provide future training.

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

Los Alamos
NATIONAL LABORATORY