HIV Recombination Overview

Presenter: Brian Foley

btf@lanl.gov

HIV Databases

Theoretical Biology and Biophysics,

Los Alamos National Laboratory

www.hiv.lanl.gov
seq-info@lanl.gov

Los Alamos
NATIONAL LABORATORY

Presenter: Brian Foley

btf@lanl.gov

HIV Databases

Los Alamos NATIONAL LABORATORY

Workshop Topics

HIV Intersubtype Recombination and Unique Circulating Recombinant Forms

General introduction:
HIV, like all retroviruses, is effectively diploid, packages two copies of the viral genome per virion.

Mechanism of Recombination:
Template switching during reverse transcription, no DNA damage-repair enzymes needed.

Tools for detecting Intersubtype Recombination:

Overview of HIV-1 M group subtypes and CRFs:

Examples of what recombinants "look like":

Los Alamos
NATIONAL LABORATORY

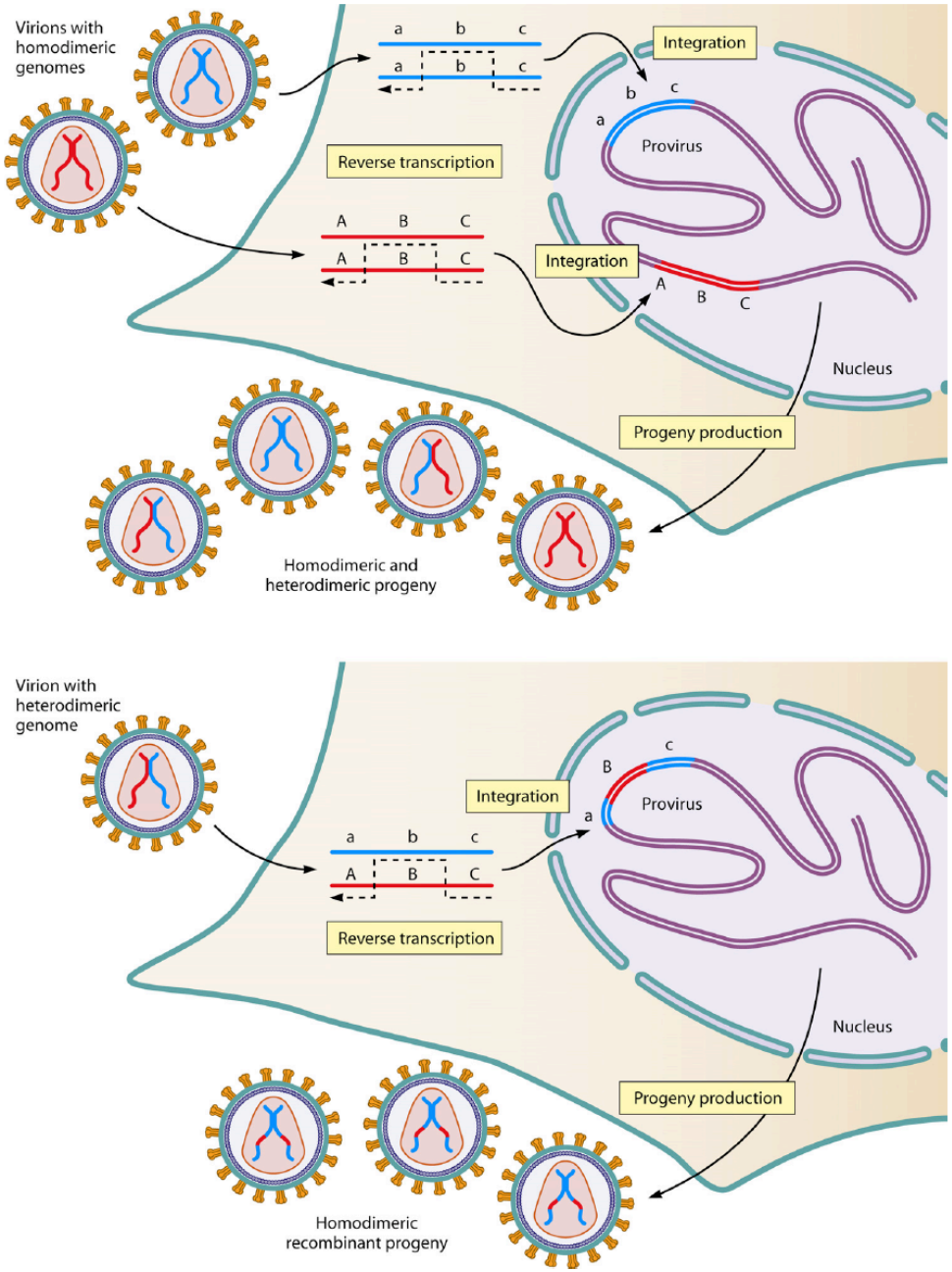


FIG. 8. Requirements for generating recombinant genomes. (A) Coinfection with two genetically distinct viruses does not yield recombinants. However, a producer cell must be coinfected with two genetically distinct viruses (shown here as viral particles with two blue or two red RNAs) to produce viral particles with heterodimeric gRNAs. (B) Recombination is observable in cells infected with heterodimeric virions (particle containing one red and one blue RNA strand). Template switching during reverse transcription can generate a recombinant provirus.

Recombinant viruses can be formed when one cell is infected with 2 viruses.

Distance or diversity between the two viruses can be large (intersubtype) or small (near identity intrapatient).

Recombination occurs by template switching during reverse transcription of heterodimeric viruses

Los Alamos NATIONAL LABORATORY

Virus Recombination Detection Tools:

RIP: HIV-databases https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html

Pros: Adjustable window size, prebuilt test set, allows user input test set, adjustable significance threshold, contains consensus for each subtype.

Cons: Does not output numerical breakpoint locations.

jpHMMer: Gobics http://jphmm.gobics.de/submission_hiv.html

Pros: Statistical support of precise breakpoints, outputs table of breakpoints.

Cons: Does not have HMM models for CRFs, weak models for rare subtypes, does not include sububtypes (A3, A4, A6, F2), window size not adjustable.

REGA Genotyping: SANBI, REGA http://regatools.med.kuleuven.be/typing/v1/subtyping.html

Pros: Accepts input of many sequences, phylogeny as well as other methods, CRFs updated reasonably often.

Cons: Window size not adjustable,

RDP3, RDP4: http://web.cbio.uct.ac.za/~darren/rdp.html

NCBI Genotyping: NCBI https://www.ncbi.nlm.nih.gov/projects/genotyping/formpagex.cgi

SimPlot: Stuart Ray https://sray.med.som.jhmi.edu/SCRoftware/simplot/

Highlighter: HIV Databases
https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html

Los Alamos
NATIONAL LABORATORY

Many more recombination detection tools:

http://bioinf.man.ac.uk/robertson/recombination/programs.shtml

Links to Recombinant Sequence Analysis/Detection Programs

Welcome to the comprehensive list of recombination analysis software maintained by the Robertson Lab.

Filter list by method or algorithm

Methods

Algorithms

3seq

- **Description** 3SEQ is a software program for identifying mosaic structure or recombination in nucleotide sequence data. 3SEQ takes as input a data set with a minimum of three aligned sequences, and it tests whether any sequence in the data set is a recombinant or mosaic of any other two sequences in the data set.
- **Implemented algorithms**
  - Recombination detection using hyper-geometric random walks
- **Reference** Boni, M.F., Posada, D. & Feldman, M.W. (2007) [PubMed][doi]

4SIS

- **Description** Two types of informative sites were distinguished, corresponding to the clustering of the putative recombinant with either of the parental representatives. The optimal breakpoint was located by maximizing a chi-square value.
- **Methods**
  - Phylogenetic methods
- **Additional notes** Temporarily unavailable.
- **Reference** van Cuyck, H., Fan, J., Robertson, D.L. & Roques, P. (2005) [PubMed][doi]

BARCE

- **Description** BARCE is a C++ program for detecting recombination breakpoints in four-sequence alignments using hidden Markov models, Bayesian principles and Markov chain Monte Carlo sampling.
- **Methods**
  - Phylogenetic methods
- **Implemented algorithms**
  - Phylogenetic hidden Markov model with Bayesian inference
  - Hidden Markov models
- **Additional notes** Implemented in TOPALi
- **Reference** McGuire, G., & Husmeier, D. (2003) [PubMed][doi]

Bellerophon

- **Description** Bellerophon is a program for detecting chimeric sequences in multiple sequence datasets by an adaption of partial treeing analysis.
- **Implemented algorithms**
  - Distance-matrix calculation across breakpoints
- **Reference** Huber, T., Faulkner, G. & Hugenholtz, P. (2004) [PubMed][doi]

BLAST Genotyping

- **Description** This tool helps identify the genotype of a viral sequence. A window is slid along the query sequence and each window is compared by BLAST to each of the reference sequences for a particular virus.
- **Implemented algorithms**
  - Similarity/distance plot

cBrother

- **Description** cBrother is software for inferring recombination when recombination is rare. This is a C version of the code originally written in Java available in DualBrothers.
- **Methods**
  - Phylogenetic methods
- **Implemented algorithms**
  - Bayesian multiple change-point modelling
- **Additional notes** Also implemented in DualBrothers.
- **Reference** Fang, F. (2005) [PubMed][doi]

DnaSP

- **Description** DnaSP, DNA Sequence Polymorphism, is a software package for the analysis of nucleotide polymorphism from aligned DNA sequence data. DnaSP can estimate several measures of DNA sequence variation within and between populations (in noncoding, synonymous or nonsynonymous sites, or in various sorts of codon positions), as well as linkage disequilibrium, recombination, gene flow and gene conversion parameters.
- **Methods**
  - Population genetics based
- **Reference** Librado, P. & Rozas, J. (2009) [PubMed][doi]

DualBrothers

- **Description** DualBrothers is a recombination detection software based on the dual Multiple Change-Point (MCP) model. This model allows for changes in topology and evolutionary rates across sites in a multiple sequence alignment.
- **Methods**
  - Phylogenetic methods
- **Implemented algorithms**
  - Bayesian multiple change-point modelling
- **Additional notes** Also implemented in cBrother.
- **Reference** Minin, V., Dorman, K., Fang, F. and Suchard, M. (2005) [PubMed][doi]

Frag-dist

- **Description** This program use the principle of "Recombinant has high similarity with its parental sequence' to visualize the location of breakpoints and gives out the potential parental sequences.
- **Methods**

Los Alamos NATIONAL LABORATORY

How to decide “which tool”?

The answer depends on many factors such as number of sequences which need to be screened, short or long sequences, diversity in the local population, etc…

No single tool does everything very well.

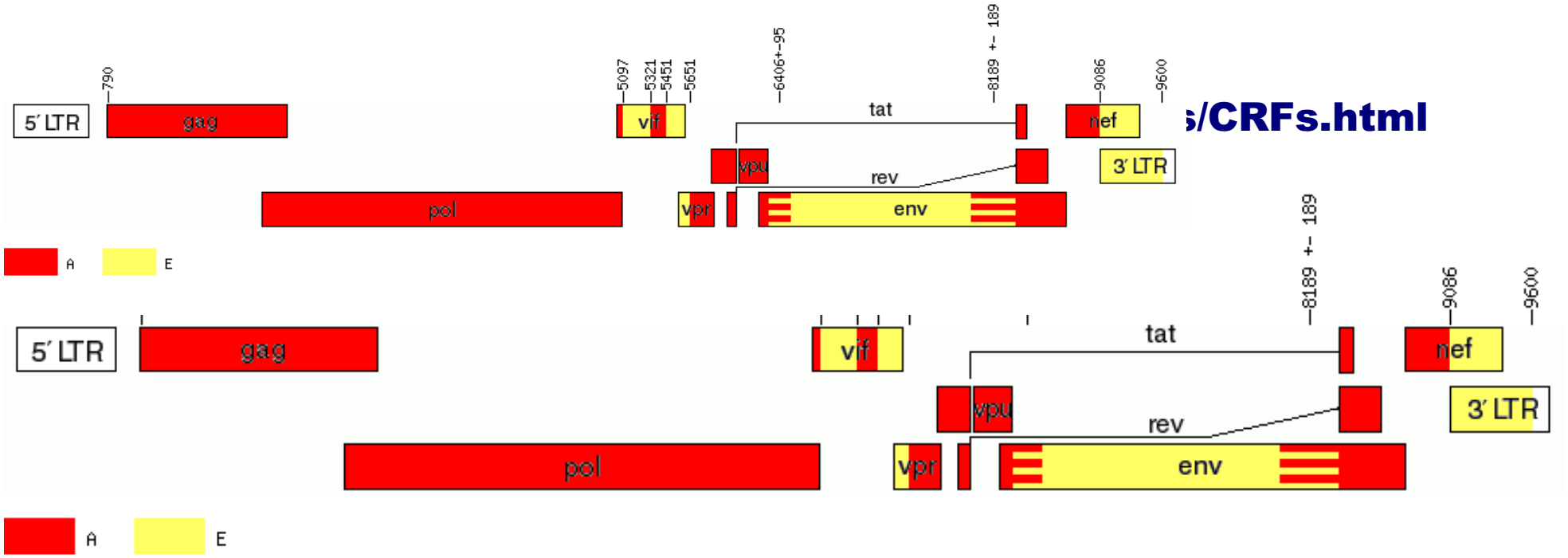
There are trade-offs in speed vs accuracy, rate of false-positive vs false negative results, etc…

Very fast methods can be simple, such as BLAST of each sequence against a small subtype reference set local database. Parsing and interpretting the results may be the most difficult part.

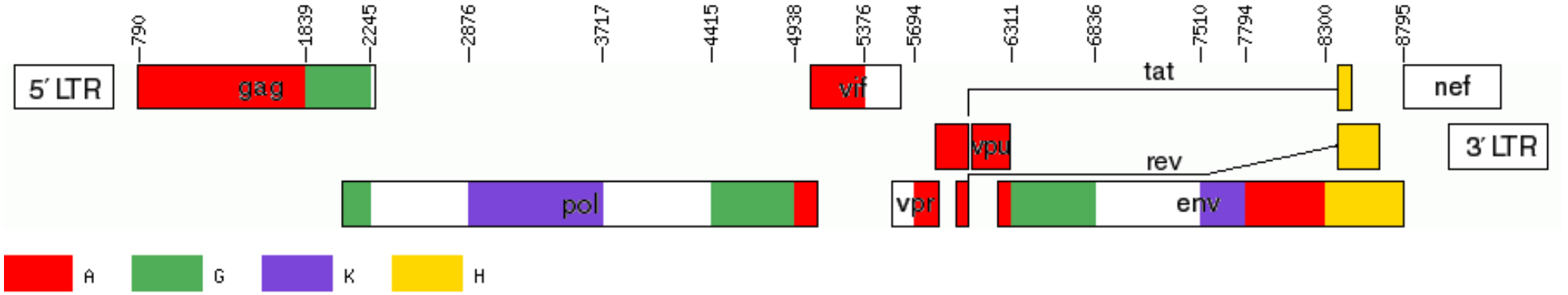
HIV-1 subtypes and CRFs

- Subtypes and subsubtypes: A1 – A6, B, C, D, F1 – F2, G, H, J, K
- Circulating Recombinant Forms: CRF01_AE – CRF98_06B
- Both require a minimum of 2 complete genomes and at least one more partial genome with sequences from regions that can confirm the structure of the first 2.

Los Alamos
NATIONAL LABORATORY



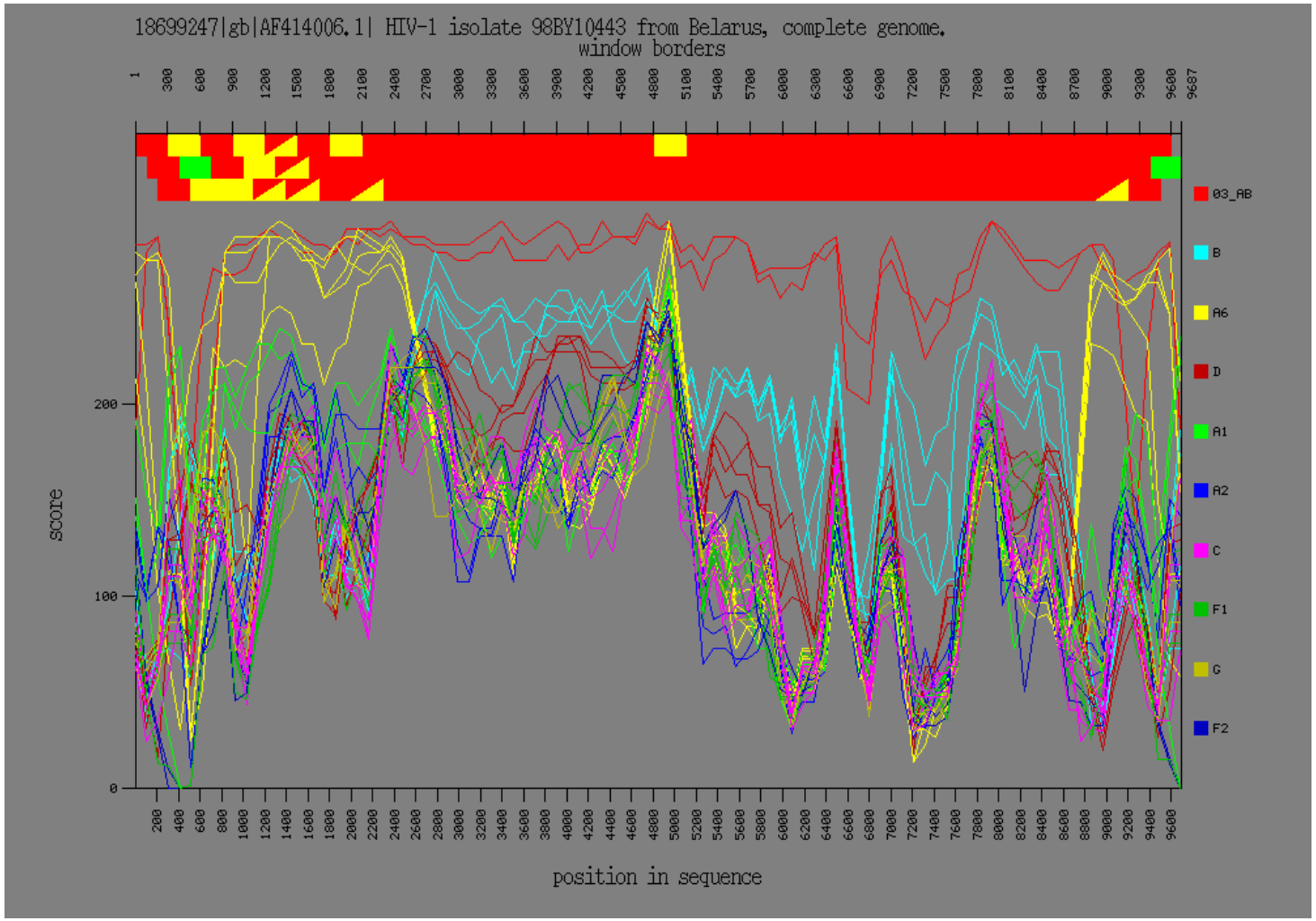
CRF04_cpx



Los Alamos NATIONAL LABORATORY

The NCBI “Window BLAST” genotyping tool

CRF03_AB



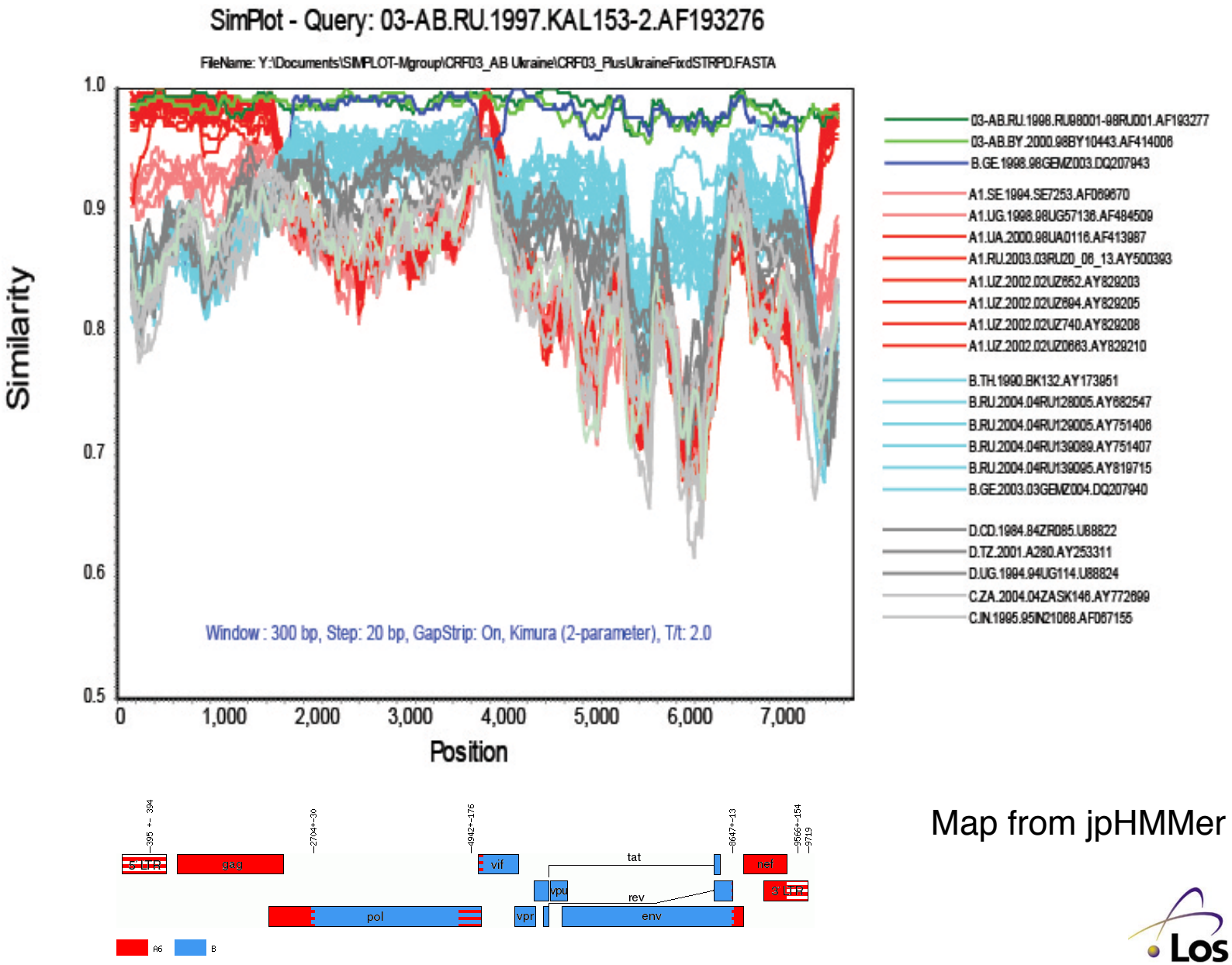
CRF03_AB is recombinant between A6, the subsubtype of A found in the former Soviet Union region and B.

In this plot 2 reference genomes of CRF03_AB are included, along with A1, A2, B, C, F1, F2, and G genomes.

Window size 150 bases and step of 50 bases

Los Alamos NATIONAL LABORATORY

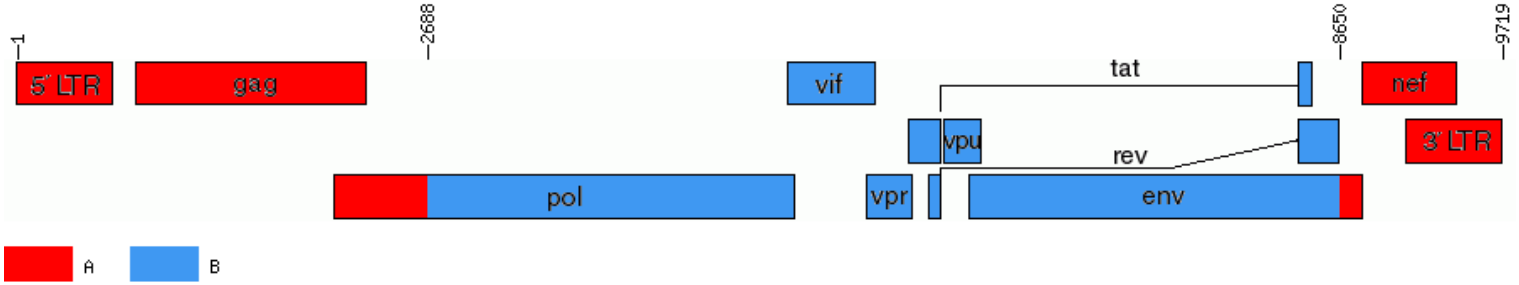
CRF03_AB SimPlot



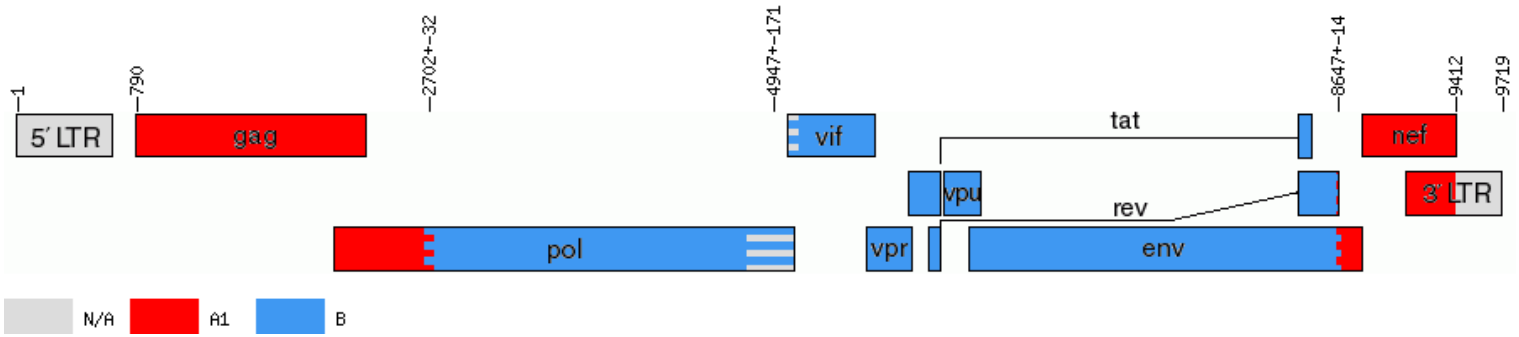
Map from jpHMMer

Los Alamos NATIONAL LABORATORY

CRF03_AB Maps



From publication describing CRF03_AB



From jpHMMer analysis at Gobics

Los Alamos
NATIONAL LABORATORY

AIDS Res. Hum. Retroviruses
Volume 16, Number 11, 2000, pp. 1047–1053

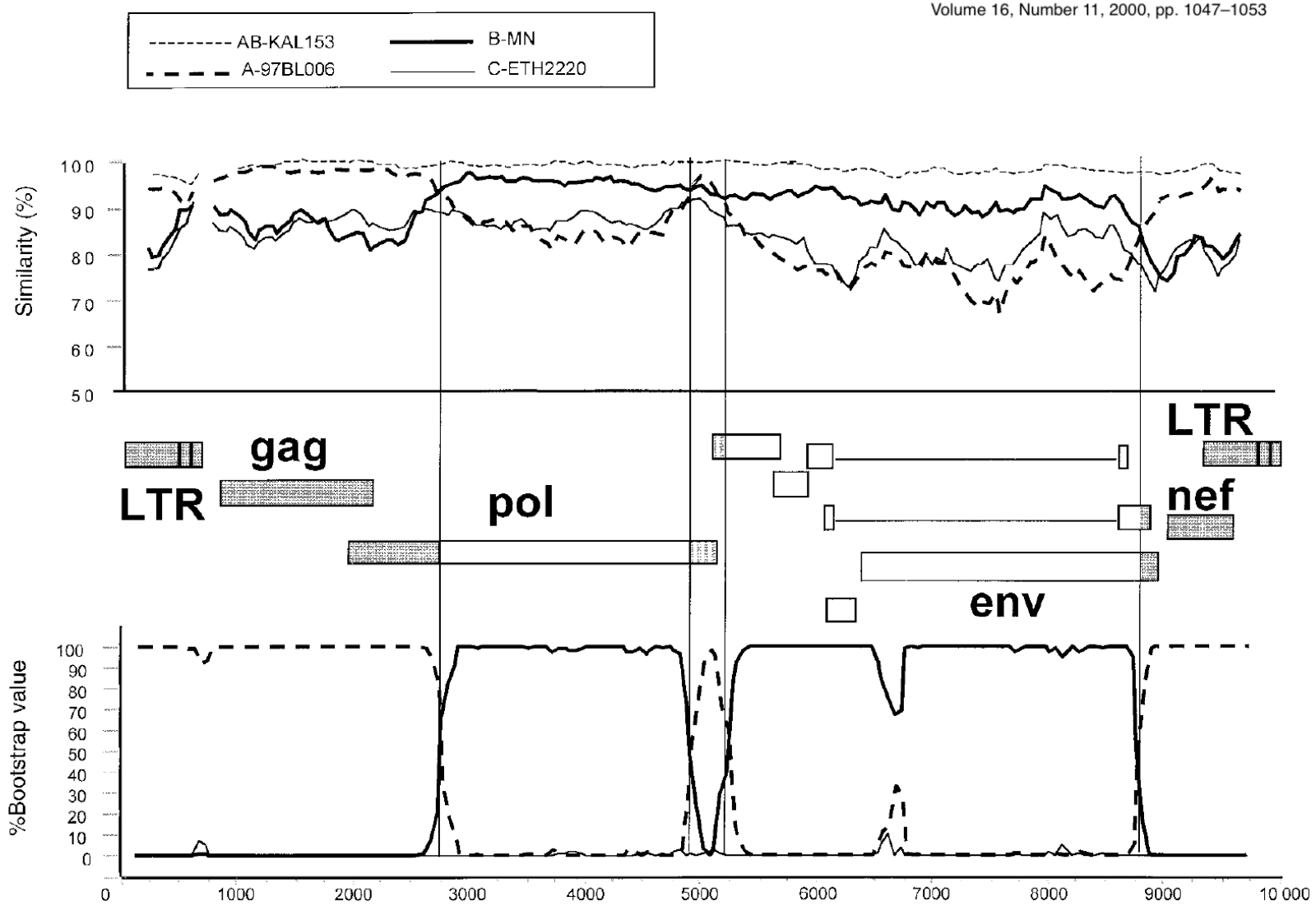


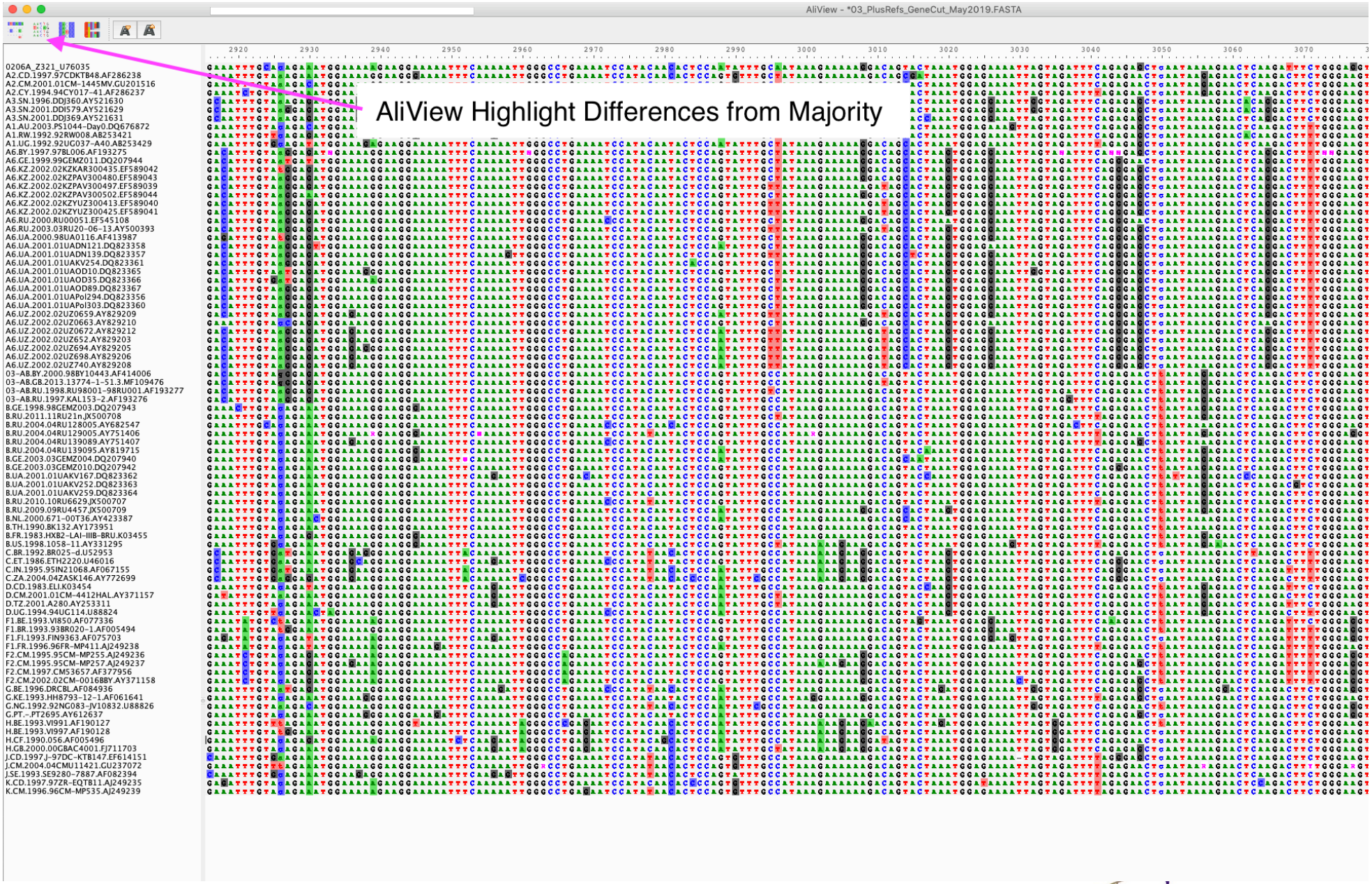
FIG. 2. Recombinant structure of the Kaliningrad IDU-associated HIV-1 strain. Similarity (*top*) and bootscanning (*bottom*) analyses were used to map the complete genome sequence of the AB-98RU001 HIV-1 isolate. In both analyses a window of 400 bases and an increment of 50 bases were used. Gap regions in the alignment were excluded from the analyses. The Kimura 2-parameter model with 100 replicates was used as the algorithm for bootscanning. Subtype C isolate ETH2220 was used as an outgroup. Similarity and bootstrap value are shown on the *y* axis and positions on the full genome alignment are shown on the *x* axis. Vertical lines indicate the recombination points. The subtype origin of the different genome regions is indicated (*middle*) as gray (subtype A) and white (subtype B) regions. The small region in the *pol–vif* region, which seems to be derived from subtype A, could not be reliably verified by separate phylogenetic analysis and is therefore shown as uncertain (striped).

AliView (free for Mac) or BioEdit (free for Windows) multiple sequence alignment editor view of CRF03_AB plus reference genomes.

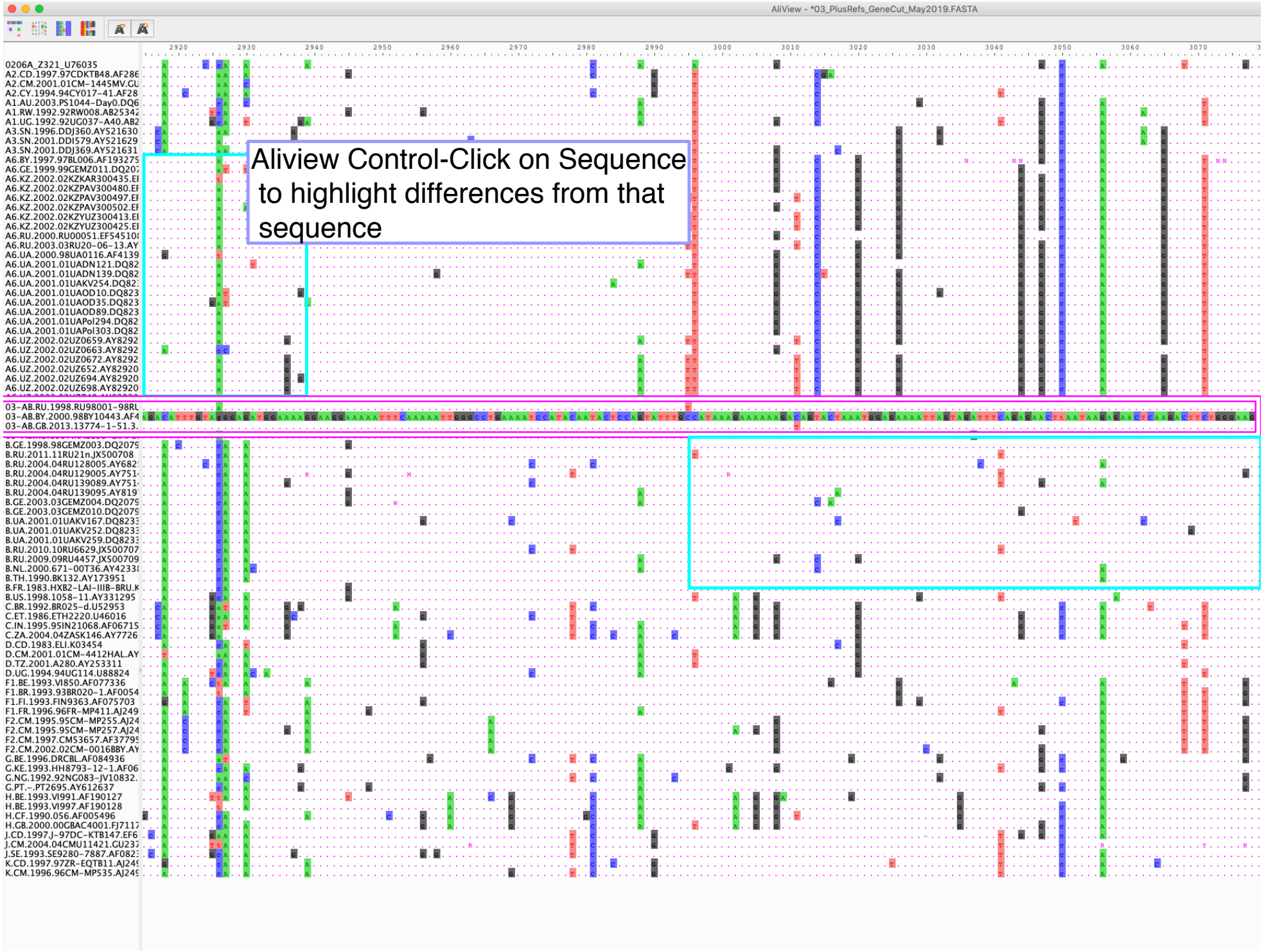
http://ormbunkar.se/aliview/

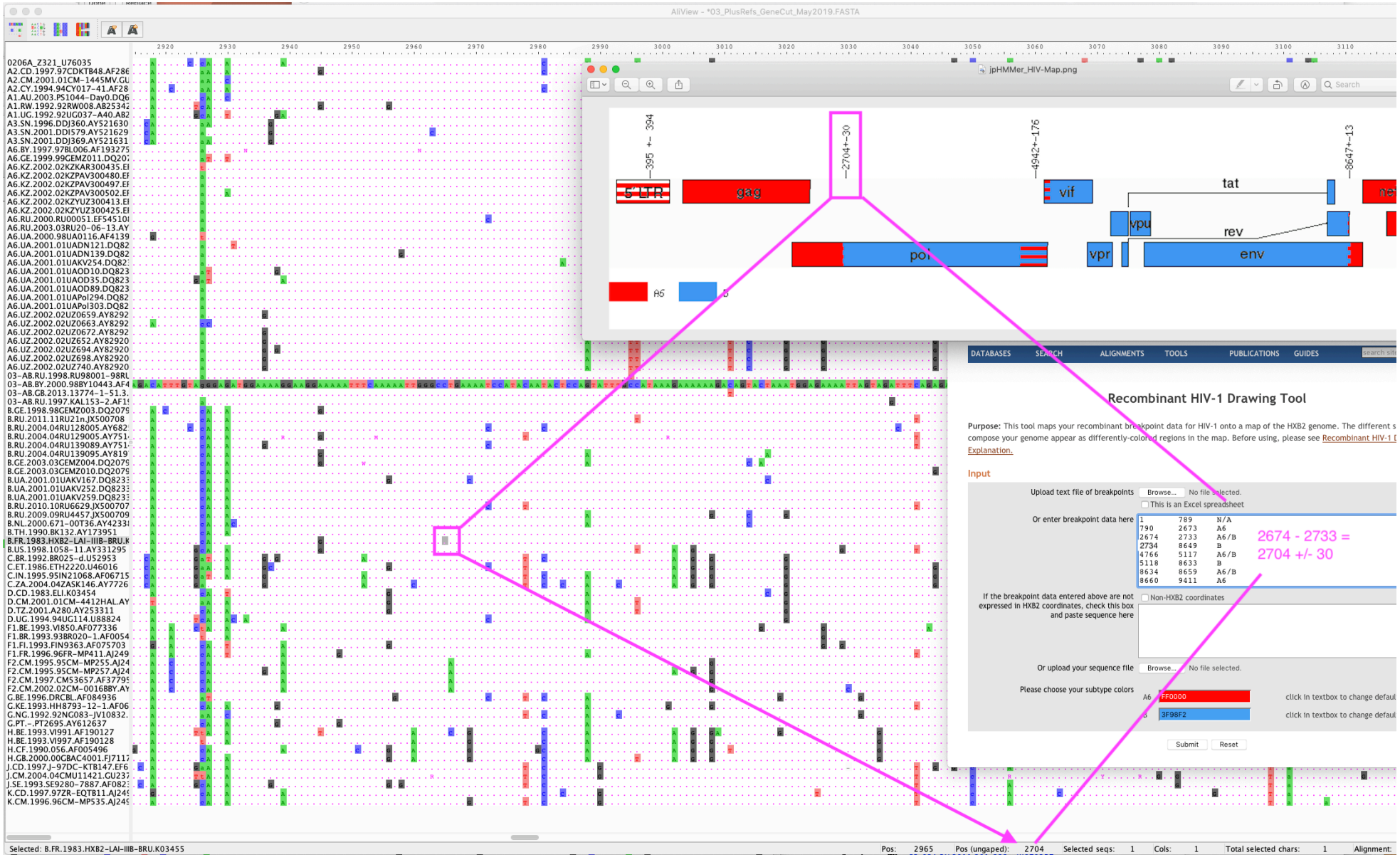
http://www.mbio.ncsu.edu/BioEdit/bioedit.html

Los Alamos NATIONAL LABORATORY



Los Alamos
NATIONAL LABORATORY


AliView - *03_PlusRefs_GeneCut_May2019.FASTA
Aliview Control-Click on Sequence to highlight differences from that sequence



Positions or sites in HIV-1 genomes are numbered using alignment to the HXB2 reference genome as the standard. The HIV Map drawing tool can be used to create maps of the genome colored by region.
https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html

Los Alamos
NATIONAL LABORATORY

jpHMM-HIV [Result]

Input data:

- sequence file

http://jphmm.gobics.de/submission_hiv

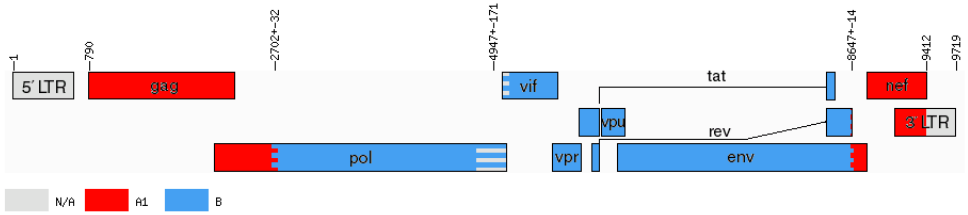
jpHMM result:

Sequence #1: >03-AB_BY_2000_98BY10443_AF414006

This sequence is related to subtype(s): **A1 B**

Fragment Start Position	Uncertainty Region Start - End	Breakpoint Interval Start - End	Fragment End Position	Fragment Subtype
Position in the original sequence [pred_recombination], [recombination_incl_UR_and_BPI], [UR_and_BPI]				
1	-	-	**798**	**N/A**
799	-	2673 - 2736	**2690**	**A1**
2691	4779 - 5120	8587 - 8614	**8603**	**B**
8604	-	-	**9368**	**A1**
9369	-	-	**9687**	**N/A**
Position based on HXB2 numbering [pred_recombination] [recombination_incl_UR_and_BPI] [UR_and_BPI]				
1	-	-	**789**	**N/A**
790	-	2670 - 2733	**2687**	**A1**
2688	4776 - 5117	8633 - 8660	**8649**	**B**
8650	-	-	**9411**	**A1**
9412	-	-	**9719**	**N/A**

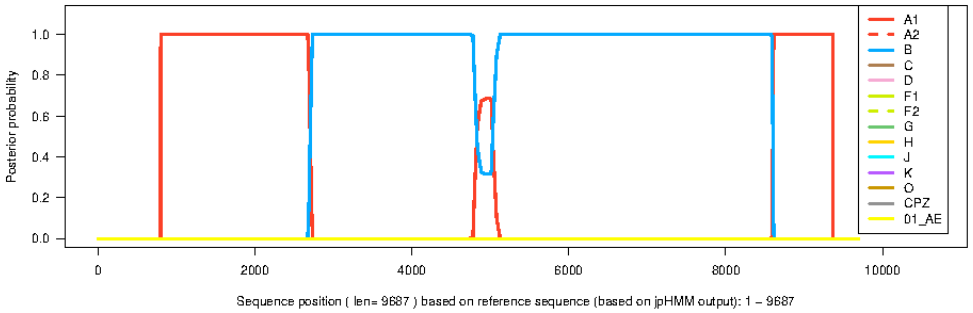
Genome map (based on HXB2 numbering)



Note:

- Numbers in the above figure denote intervals for recombination breakpoints based on HXB2 numbering.
- The uncolored regions denote missing information due to input fragment sequence.
- The gray regions denote missing infomation due to uninformative subtype models (subtype: N/A).
- The sequence regions of less than 10 nucleotides long are too short to be mapped onto the genome map.

Posterior probabilities of the subtypes (based on HXB2 numbering)

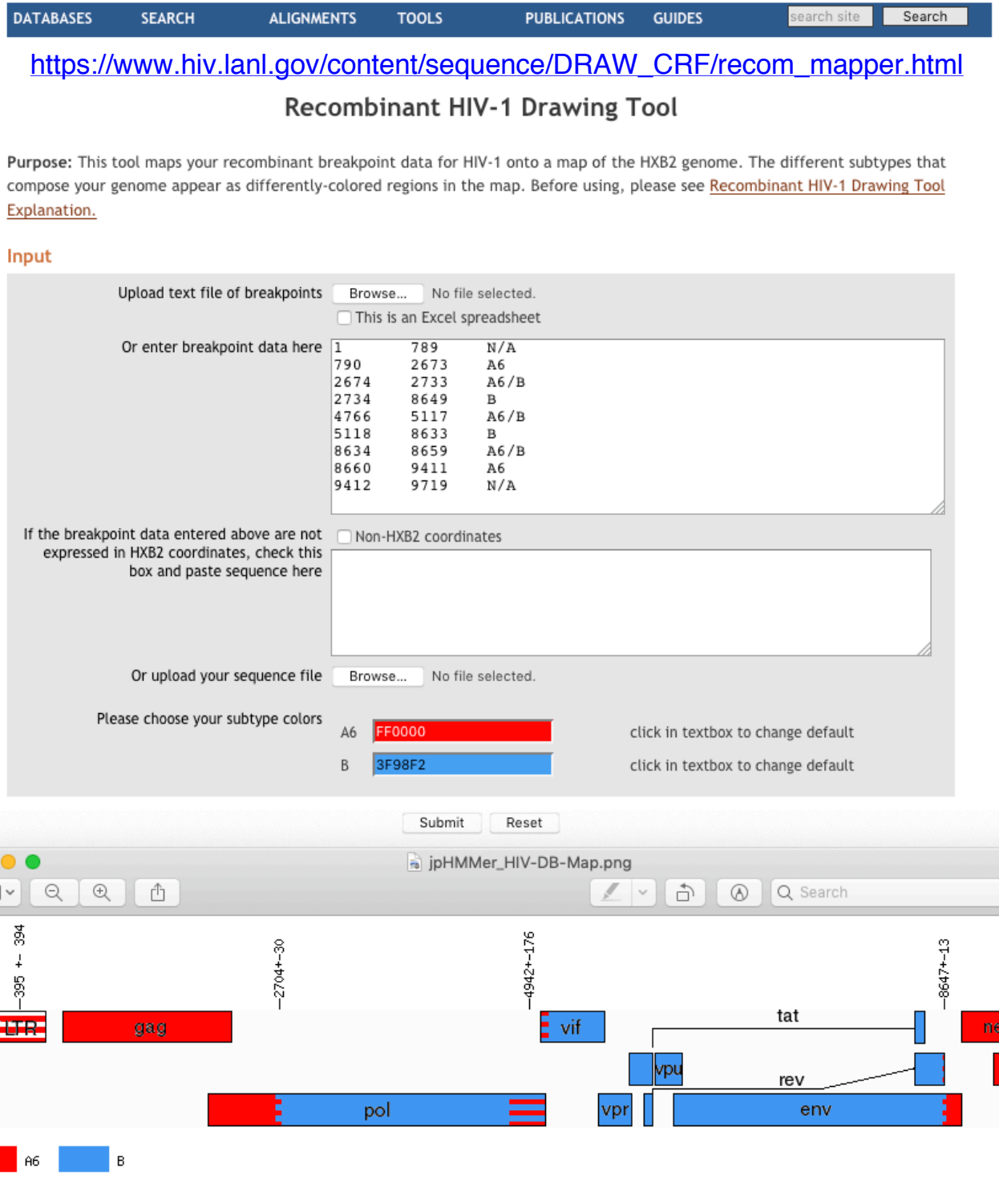


jpHMM-HIV at Gobics gives best recombination site location numbering.

vs input query sequence location

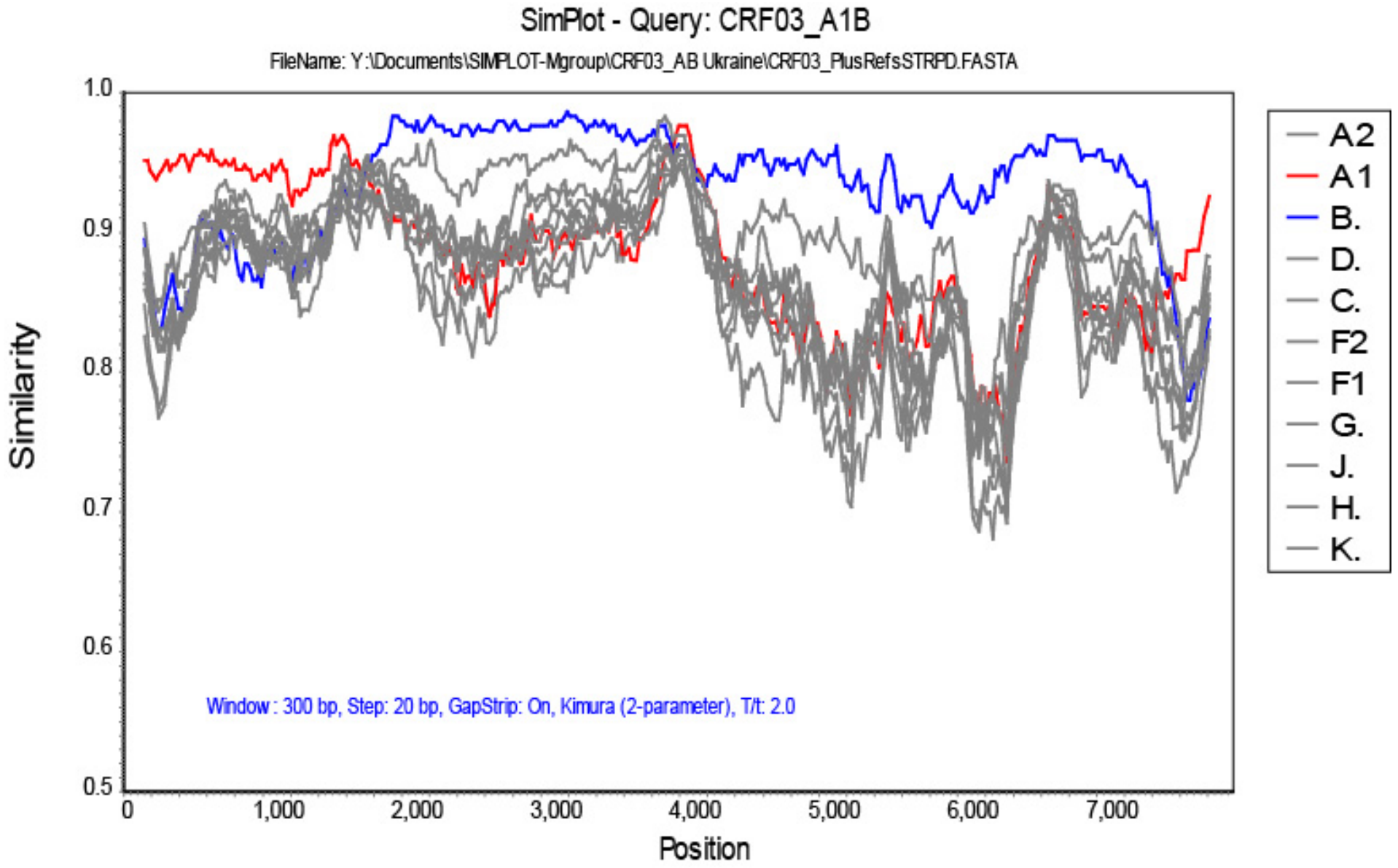
vs HXB2 standard sequence location

Los Alamos NATIONAL LABORATORY



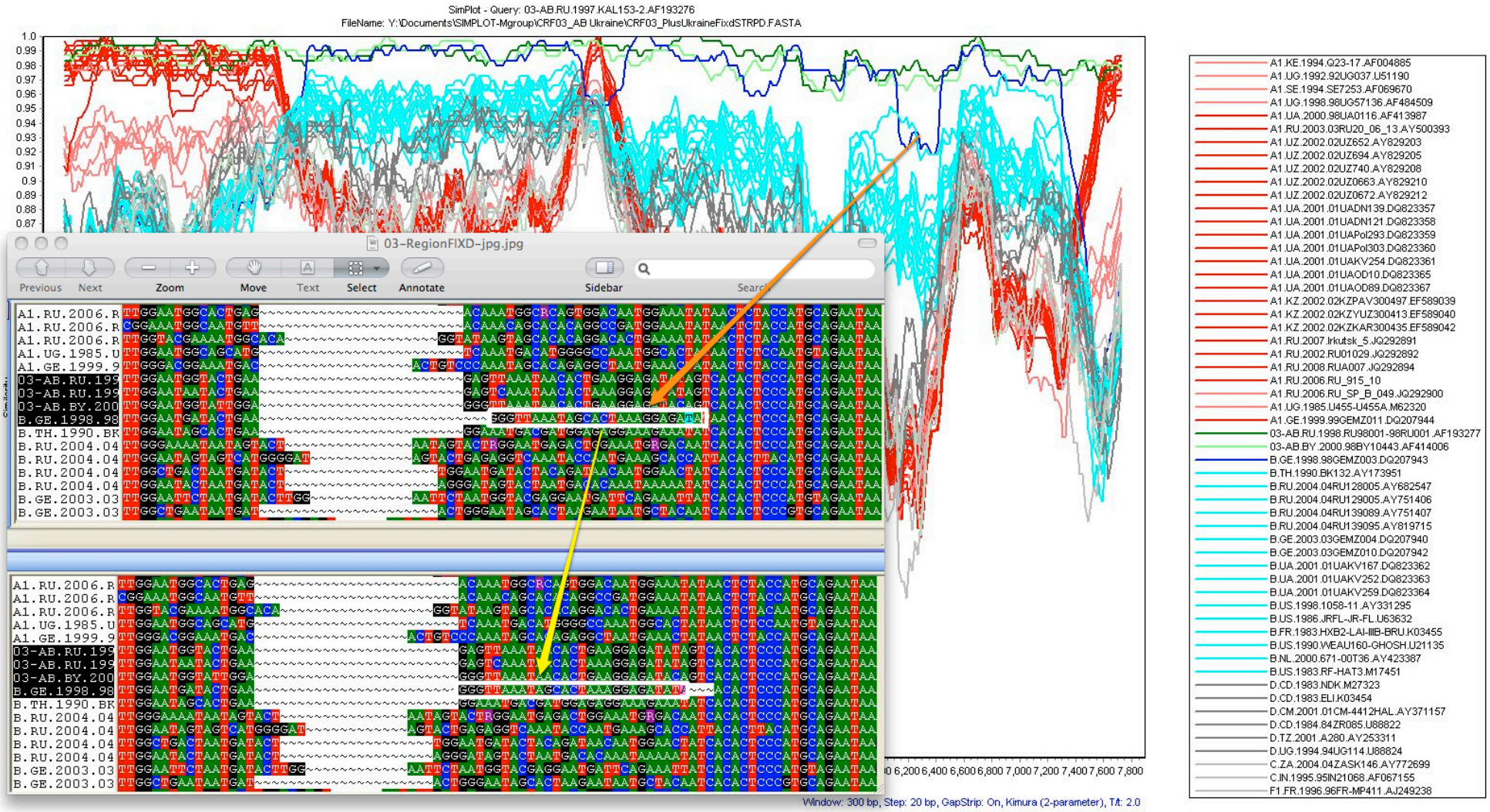
HIV recombination Map Drawing tool at HIV-DB customizable to change colors used, and input your own numbers in cases where jpHMM (or another tool) did not give the correct sites.

Los Alamos NATIONAL LABORATORY



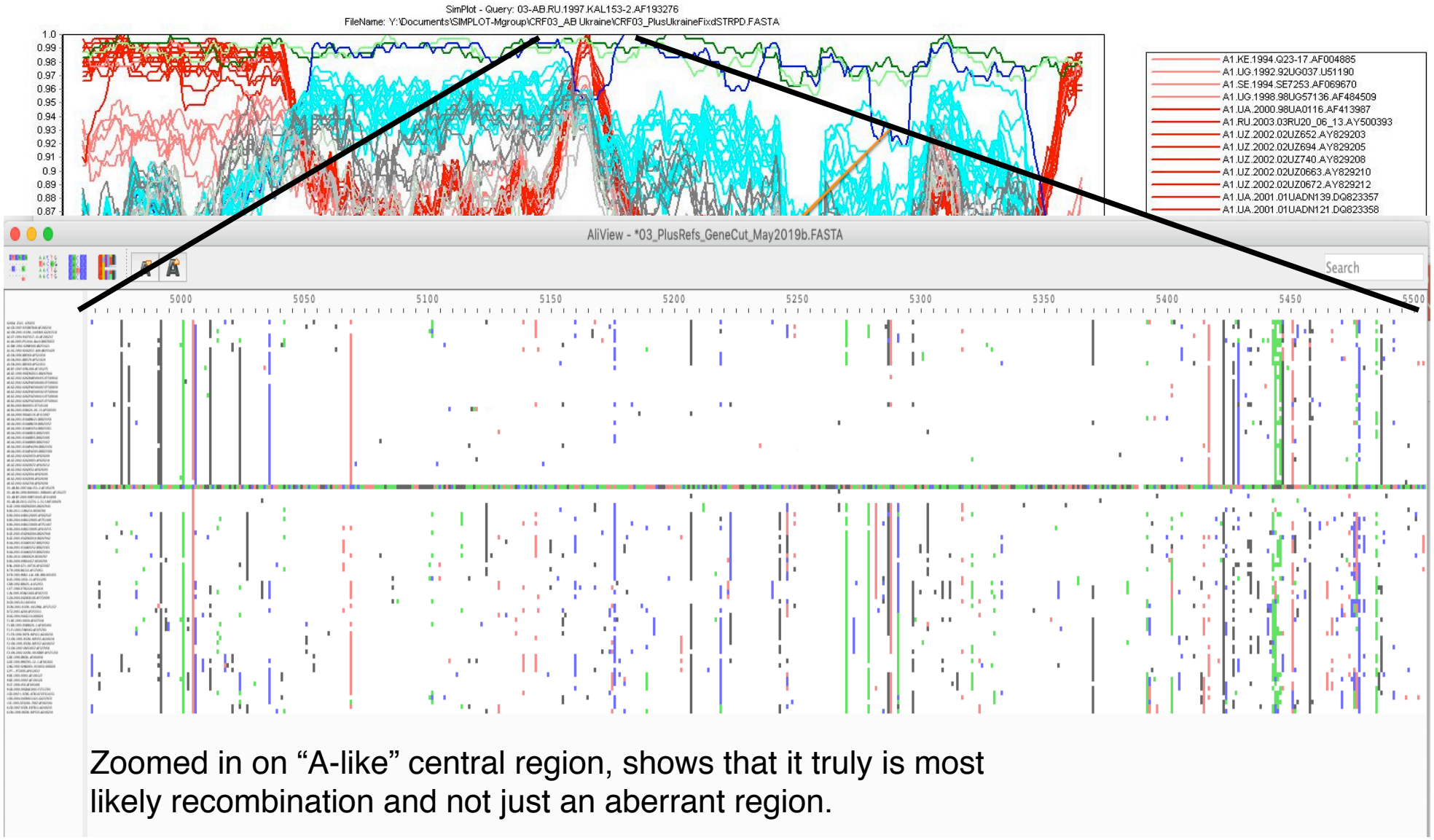
https://sray.med.som.jhmi.edu/SCRoftware/simplot/

Los Alamos
NATIONAL LABORATORY



Even a very small region of misalignment, hypermutation, or poor sequence quality can have a large impact on similarity plots, phylogenetic trees, and other analyses. Similarity plots can be quite useful for identifying sites in a multiple sequence alignment that should be scrutinized, and corrected if in error, as this example shows.

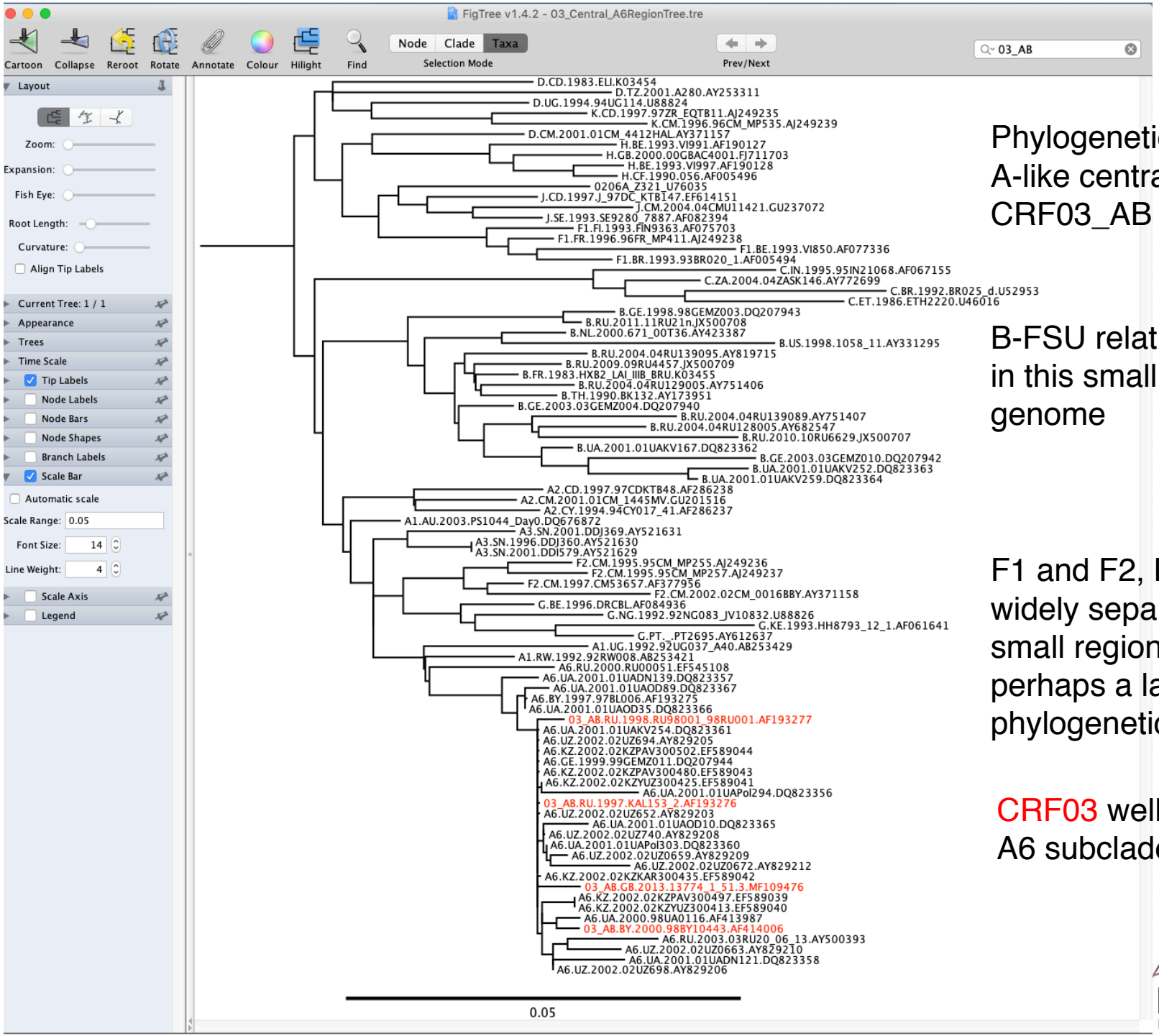
Los Alamos
NATIONAL LABORATORY



Zoomed in on “A-like” central region, shows that it truly is most likely recombination and not just an aberrant region.

Other factors should also be considered too. In this case for example we know that the recombinant was formed in a person dual-infected with A6 and B viruses, so a region of A6 is not at all unexpected.

Los Alamos
NATIONAL LABORATORY



Phylogenetic tree of A-like central region of CRF03_AB

B-FSU relatively diverse in this small region of the genome

F1 and F2, B and D widely separated in this small region, indicates perhaps a lack of solid phylogenetic information

CRF03 well within the A6 subclade

Los Alamos
NATIONAL LABORATORY

KATHOLIEKE UNIVERSITEIT LEUVEN

REGA HIV Subtyping Tool

HIV BIOINFORMATICS BIOAFRICA

REGA HIV-1 & 2 Automated Subtyping Tool (Version 2.0)

This tool is designed to use phylogenetic methods in order to identify the subtype of a specific sequence. The sequence is analysed for recombination using bootscanning methods.

Note for batch analysis: *The REGA subtype tool accepts up to 1000 sequences at a time.*

Enter here your input data as FASTA format.

https://www.genomedetective.com/app/typingtool/hiv

Choose a mirror to subtype your sequences

```
>03-AB.RU.1998.RU98001-98RU001.AF193277
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------------------------------------------------------------------------------
----------aaatctctagcagtggcgcccg-----------------------aaca-gggacttg-
aaagcgaaagt---tccag---agaagatctcctcgacgc-aggactcggcttgct-----g-aggtgc-
acacagcaagaggc-gagagcg--gcgactggtgagtacgcctaa-aga------t-tTTTGACTAGCGGAG---
GCTAGAAGGAGAGAG--
```

clear Run

Submit sequences | How to cite | HIV Tutorials | HIV Decision Trees | HIV Subtyping Process | HIV Example Sequences | Contact us

---

Developed by: Tulio de Oliveira, Koen Deforche, Sharon Cassol, Andrew Rambaut and Anne-Mieke Vandamme.

Developed in cooperation with the Evolutionary Biology Group at University of Oxford, UK., the HIV-1 Pathogenesis and Immunotherapeutics Program at University of Pretoria, South Africa, and the REGA Institute at the Katholieke Universiteit Leuven, Belgium.

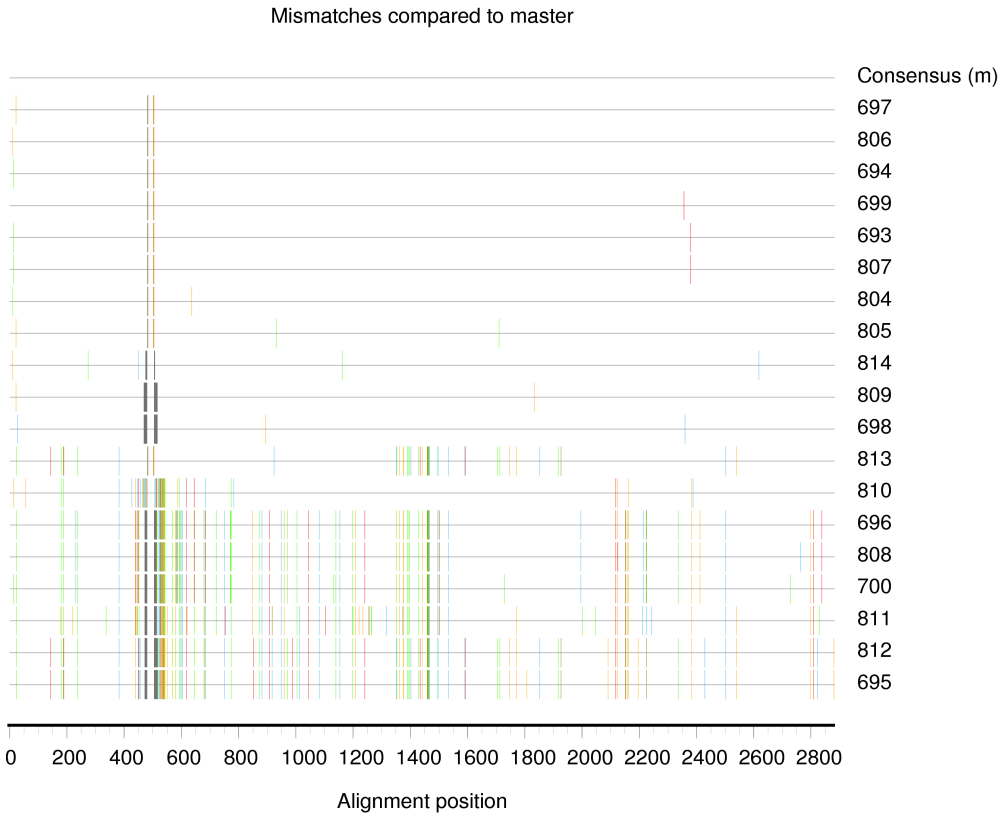
Funded by the Marie Curie Fellowship, Flanders Bilateral Cooperation Program and the Wellcome Trust (grant # 061238).

For questions, suggestions or problems please contact: Dr. Tulio de Oliveira.

Los Alamos NATIONAL LABORATORY

HighLighter for intra-patient recombination

https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html



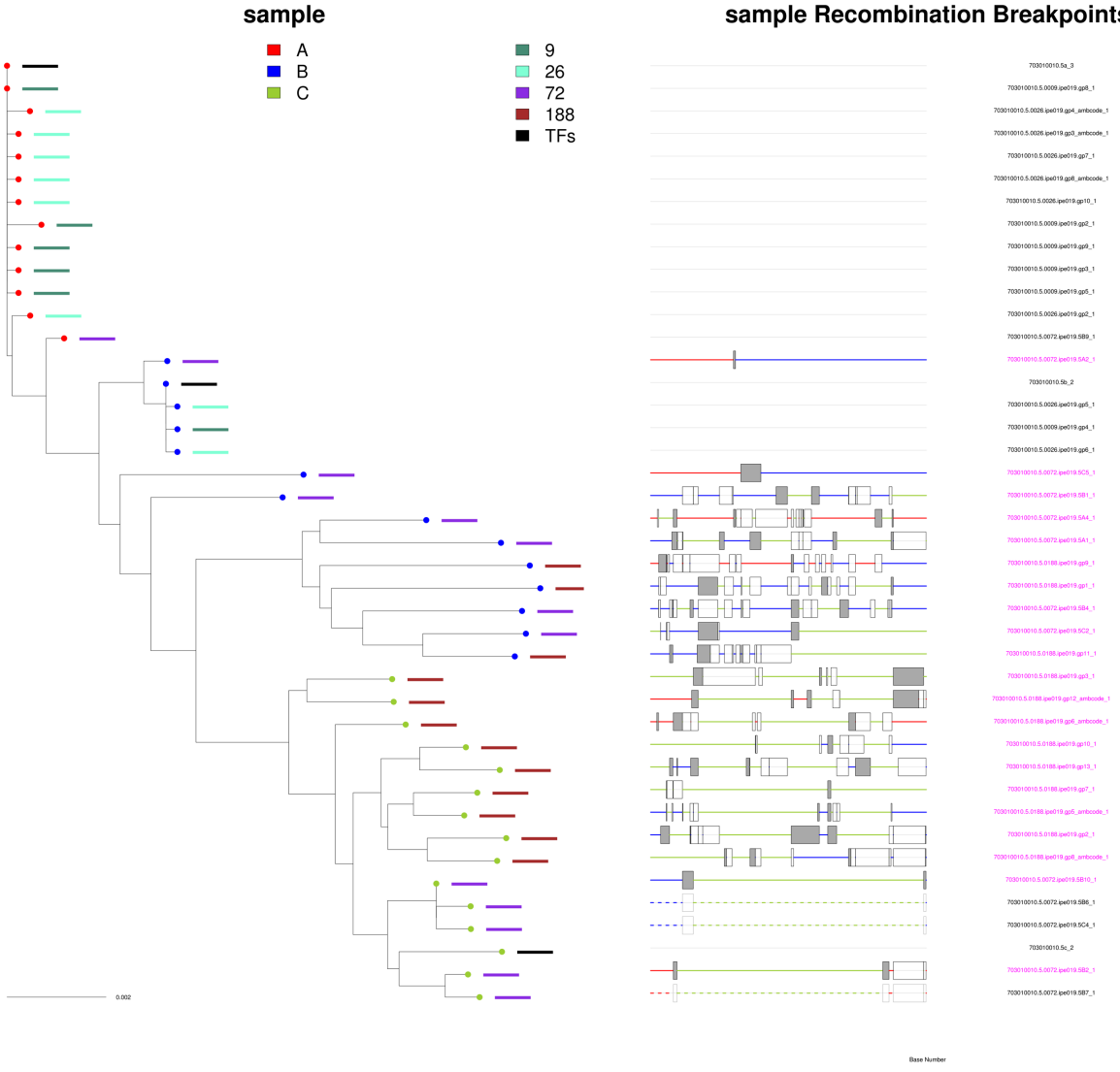
Dual-infected patient.

Infected with two strains of the same subtype.

Los Alamos NATIONAL LABORATORY

RAPR recombination analysis program

https://www.hiv.lanl.gov/content/sequence/RAP2017/



Song H, Giorgi EE, et al.
Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection.
Nat Commun. 2018
PubMed: 29765018

Los Alamos NATIONAL LABORATORY