Supplementary Materials For:

CAUSA 2.0: Accurate and Consistent Evolutionary Analysis of Proteins Using Codon and Amino Acid Unified Sequence Alignments

Xiaolong Wang[*], *Chao Yang*

Department of Biotechnology, Ocean University of China, Qingdao, 266003, China

**E-mail: Xiaolong@ouc.edu.cn*

1. Materials and Methods

1.1 Protein coding sequences and data analysis flowchart

Different strains of human and simian immunodeficiency virus were derived from the seed alignment of Pfam family pf00516. The coding sequences (CDSs) of the Envelope glycoprotein gp120 (Env) and core (Gag) proteins were retrieved from the HIV database. Thirty protein families and their standard phylogenetic trees in human and mammalian animals were randomly selected from TreeFam-A (http://www.treefam.org/).

The flowchart of data analysis is shown in Fig 1 in the main document, DNA or protein alignments were aligned by the multiple sequence alignment tools at EBI (http://www.ebi.ac.uk/), including ClustalW, MAFFT, MUSCLE, T-COFFEE and PRANK. Codon alignments were aligned by PRANK using "align translated codons" option (PRANK-Codon), and a codon alignment tool (CAT) provided by the HIV database at the Los Alamos National Laboratory (http://www.hiv.lanl.gov/). All programs were run with their default settings.

1.2 Simulation of Coding DNA sequences

Coding sequences were simulated using *indel-seq-gen v2.1.03*. Thirty samples of 8 or 16 individuals with 500 codons were simulated guided by an asymmetric tree (Fig S5B) in mode HKY. The simulated coding sequences were aligned by different alignment programs. Phylogeny trees were constructed from each method and compared with the guide tree input into the program. Two datasets were simulated using the following indel options:

(1) The low-level-indel dataset: the maximum indel size is 30 nucleotides, and with an indel occurs once for every 100 base substitutions. And the --invar option is set to be 0.99, so that 99 percent of the sites is invariable;

(2) The high-level-indel dataset: the maximum indel size is 30 nucleotides, and with an indel occurs once for every 20 base substitutions. And the --invar option is set to be 0.9, so that 90 percent of the sites is invariable.

1.3 DNA, Protein, Back-translated and Codon Alignments

DNA or protein alignments were aligned by multiple sequence alignment tools at EBI (http://www.ebi.ac.uk/), including ClustalW, MAFFT, MUSCLE, T-COFFEE and PRANK. Back-translated codon alignments were constructed by MEGA5 (ClustalW-BT) or PRANK (PRANK-BT) using "align translated proteins" option. Codon (64-state) alignments were constructed by PRANK using "align translated codons" option (PRANK-Codon), and a codon alignment tool (CAT) provided by the HIV database at the Los Alamos National Laboratory (http://www.hiv.lanl.gov/). All programs were run with their default settings.

1.4 Converting CDSs into unified sequences

As shown in Fig 1 in the main document, coding DNA sequences (CDSs) were translated into amino acids (AA) and converted into *Codon-AA Unified Sequences* using CAUSA 2.0, in which the one-letter code of every AA is immediately followed by its encoding triplet codon, forming a *codon-AA 4-tuple*. Since nucleotides and AAs both contain code A, C, G, U and T, in every 4-tuple, nucleotides and AAs are written respectively in lower and upper case.

1.5 Construction of a combined DNA-protein scoring matrix

A combined DNA-protein scoring matrix (CDPSM) is a 24 x 24 array derived by combining a nucleotide substitution matrix with an amino acid substitution matrix, such as Gonnet250, Blosum62 and PAM250 scoring matrices. As shown in Table S1, in the default matrix, CDP-Gon250, the first line of the matrix is the symbol set, lower case letters (a, c, g, u, t) stands for DNA/RNA bases, upper case letters (A/B, C/X, D, E, F, G/Z, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) stand for amino acids. In CDPSM, substitutions between any pair of amino acids, or any pair of nucleotide, are allowed, but high penalties (-999) is given to prohibited mismatches between any nucleotide

and amino acid. The different scales of the nucleotide and amino acid substitution matrices were reconciled, the optimal scores for matched (20) and mismatched (-10) base, and the penalties for opening (20) and extension (0.4) gaps, were chosen heuristically by testing scores on a set of ENV CDSs (For match, mismatch scores and gap opening penalty ranged from -50 to +50 stepped by 5, and gap extension penalty ranged from 0.1 to 1.0 stepped by 0.1). Fortunately, it seems that scores optimized on this set of sequences works as well in other sequences also. The scoring matrix could be changed by users if needed, and several combined scoring matrices that are suitable for different divergent sequences were provided in the software packages.

1.6 Aligning codon-AA unified sequences.

The unified sequences were subjected to progressive alignment by calling ClustalW using a user-defined *combined DNA-Protein* (CDP) scoring matrix, such as *CDP-Gon250* matrix (Table S1). Technically, CAUSA can be used in any MSA algorithm. Only ClustalW was selected because the other MSAs do not support user defined substitution matrices. A set of heuristically chosen optimal parameters for ClustalW to align the CAUSs were listed in Table S2. ClustalW does not support case-sensitive alphabet, in order to avoid the conflict in bases and amino acid codes, CAUSA changes the codes input to ClustalW temporarily: replacing base t with u, and replacing amino acid A with B, C with X, and G with Z.

A unified alignment output from ClustalW is input back into CAUSA for final processing: the codes were changed back into ordinary codes by replacing base u with t, amino acid B with A, X with C, and Z with G, and then the alignment was displayed in unified, protein, and/or DNA views (Fig S1), in which bases were written in lowercase and amino acids in uppercase. In addition, every triplet codon and their encoded amino acid, together with the gaps inserted among them, were colored with 64 different colors, thus synonymous and non-synonymous substitutions, insertion and deletions, can be easily distinguished visually. The alignment statistics were computed, including total numbers of sites, substitutions, insertions, deletions, split codons, their average length (codons), and ratio.

At present, for CAUSA 2.0, two programs have been released as open source software under GNU/GPL license, and are downloadable free of charge from website www.dnapluspro.com. The first is an executable for Microsoft Windows with a user-friendly graphic interface programmed in Microsoft Visual C#. The second is coded in Java and run cross-platform in Linux, Windows or MAC OS.

1.7 Construction of phylogenetic trees

Phylogenetic trees were drawn using MEGA v5.05. Phylogenetic trees for a family of protein homologs and their CDSs were reconstructed respectively from protein alignments given by ClustalW, MAFFT, MUSCLE, T-COFFEE, PRANK, and codon alignments given by PRANK, CAT and CAUSA. Evolutionary histories were inferred using Neighbor joining (NJ), or maximum likelihood (ML) method based on Tamura-Nei model. The percentage of trees in which the associated taxa clustered together is shown next to the branches.

In ML method, the tree with the highest log likelihood is shown. Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was < 100 or less than one fourth of the total number of sites, the maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. For each alignment, two modes of gaps/missing data treatment were compared: (1) *Complete deletion* (CD): every site containing gaps and missing data was eliminated; (2) *Use all sites* (AS): all sites were used, but all ambiguous positions were removed for each sequence pair.

1.8 Testing the computation time of different alignment methods

1.8.1 Testing environment:

Intel(R) Xeon(R) CPU E7- 4820 @ 2.00GHz, 64 Cores; 1TB shared memory; Operating system: Red Hat Enterprise Linux Server release 6.2 (Santiago), Linux version 2.6.32-220.el6.x86_64.

1.8.2 Testing sequence data:

As shown in Table S6, the sequence data is the same to that were used in Table S3. The number of sequences ranges from 6~30, the total length ranges from 5~61 kb.

1.8.3 Testing methods:

Download and install the latest Linux version of all of the alignment software on the local computer. A set of Perl scripts was written to call the programs to align the testing sequence data, and measure the computation time for each program on each data set.

2. *Supplementary Tables*

Table S1. The combined DNA-Protein scoring matrix (CDP-Gon250).

	a	c	g	u	B (A)	X (C)	D	E	F	Z (G)	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
a	20	-10	-10	-10	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	**a**
c	-10	20	-10	-10	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	**c**
G	-10	-10	20	-10	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	**G**
u	-10	-10	-10	20	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	-999	**u**
B(A)	-999	-999	-999	-999	24	5	-3	0	-23	5	-8	-8	-4	-12	-7	-3	3	-2	-6	11	6	1	-36	-22	**B(A)**
X(C)	-999	-999	-999	-999	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	**X(C)**
D	-999	-999	-999	-999	-3	-32	47	27	-45	1	4	-38	5	-40	-30	22	-7	9	-3	5	0	-29	-52	-28	**D**
E	-999	-999	-999	-999	0	-30	27	36	-39	-8	4	-27	12	-28	-20	9	-5	17	4	2	-1	-19	-43	-27	**E**
F	-999	-999	-999	-999	-23	-8	-45	-39	70	-52	-1	10	-33	20	16	-31	-38	-26	-32	-28	-22	1	36	51	**F**
Z(G)	-999	-999	-999	-999	5	-20	1	-8	-52	66	-14	-45	-11	-44	-35	4	-16	-10	-10	4	-11	-33	-40	-40	**Z(G)**
H	-999	-999	-999	-999	-8	-13	4	4	-1	-14	60	-22	6	-19	-13	12	-11	12	6	-2	-3	-20	-8	22	**H**
I	-999	-999	-999	-999	-8	-11	-38	-27	10	-45	-22	40	-21	28	25	-28	-26	-19	-24	-18	-6	31	-18	-7	**I**
K	-999	-999	-999	-999	-4	-28	5	12	-33	-11	6	-21	32	-21	-14	8	-6	15	27	1	1	-17	-35	-21	**K**
L	-999	-999	-999	-999	-12	-15	-40	-28	20	-44	-19	28	-21	40	28	-30	-23	-16	-22	-21	-13	18	-7	0	**L**
M	-999	-999	-999	-999	-7	-9	-30	-20	16	-35	-13	25	-14	28	43	-22	-24	-10	-17	-14	-6	16	-10	-2	**M**
N	-999	-999	-999	-999	-3	-18	22	9	-31	4	12	-28	8	-30	-22	38	-9	7	3	9	5	-22	-36	-14	**N**
P	-999	-999	-999	-999	3	-31	-7	-5	-38	-16	-11	-26	-6	-23	-24	-9	76	-2	-9	4	1	-18	-50	-31	**P**
Q	-999	-999	-999	-999	-2	-24	9	17	-26	-10	12	-19	15	-16	-10	7	-2	27	15	2	0	-15	-27	-17	**Q**
R	-999	-999	-999	-999	-6	-22	-3	4	-32	-10	6	-24	27	-22	-17	3	-9	15	47	-2	-2	-20	-16	-18	**R**
S	-999	-999	-999	-999	11	1	5	2	-28	4	-2	-18	1	-21	-14	9	4	2	-2	22	15	-10	-33	-19	**S**
T	-999	-999	-999	-999	6	-5	0	-1	-22	-11	-3	-6	1	-13	-6	5	1	0	-2	15	25	0	-35	-19	**T**
V	-999	-999	-999	-999	1	0	-29	-19	1	-33	-20	31	-17	18	16	-22	-18	-15	-20	-10	0	34	-26	-11	**V**
W	-999	-999	-999	-999	-36	-10	-52	-43	36	-40	-8	-18	-35	-7	-10	-36	-50	-27	-16	-33	-35	-26	142	41	**W**
Y	-999	-999	-999	-999	-22	-5	-28	-27	51	-40	22	-7	-21	0	-2	-14	-31	-17	-18	-19	-19	-11	41	78	**Y**
	a	**c**	**g**	**u**	**B (A)**	**X (C)**	**D**	**E**	**F**	**Z (G)**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	

Table S2. Optimal parameters for ClustalW to align Unified sequences.

Option	Parameter	Description
-INFILE	Input.fasta	Input merged DNA-Protein sequence in FASTA format
-TYPE	PROTEIN	protein alignment
-NEGATIVE	ON	protein alignment with negative values in matrix
-OUTFILE	Output.fasta	Output sequence alignment file name
-OUTPUT	PIR	Output alignment file in FASTA format
-PWMATRIX	CDP-Gon250.txt	User-defined pairwise alignments scoring matrix (Table S1)
-PWGAPOPEN	20~40	Pairwise alignments gap opening penalty
-PWGAPEXT	0.1~0.2	Pairwise alignments gap opening penalty
-MATRIX	CDP-Gon250.txt	User-defined multiple alignments scoring matrix (Table S1)
-GAPOPEN	20~40	Multiple alignments gap opening penalty
-GAPEXT	02~0.4	Multiple alignments gap opening penalty
-ENDGAPS	NO	No end gap separation pen.
-GAPDIST	4	Gap separation pen. range
-NOPGAP		Residue-specific gaps off
-NOHGAP		Hydrophilic gaps off
-ITERATION	NONE	Perform iteration at each step to improve the alignment.

Table S3. The number of consistent branches (NCB), bootstrap percentages (BSP) and TOPD/FMTS split distances (SD) in variable protein families

No	TreeFam ID	Total Branches	ClustalW-DNA-as			ClustalW-DNA-cd			ClustalW-codon-AS			ClustalW-codon-cd			prank-codon-as			prank-codon-cd			prank-DNA-as			prank-DNA-cd			CAUSA-DNA-AS			CAUSA-DNA-cd		
			NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD	NCB	BSP	SD
1	101201	10	9	65.40	0.1429	9	65.80	0.1429	9	60.50	0.1429	9	59.10	0.1429	8	65.40	0.1429	8	63.60	0.2857	8	63.90	0.1429	8	64.20	0.1429	9	63.00	0.1429	8	65.00	0.2857
2	105101	12	11	70.25	0.0000	9	66.83	0.2222	10	69.33	0.1111	11	68.25	0.1111	10	68.42	0.3333	9	63.50	0.3333	9	64.42	0.3333	10	66.33	0.3333	11	68.33	0.1111	11	69.67	0.1111
3	105301	12	12	82.58	0.0000	12	82.67	0.0000	12	82.08	0.0000	12	82.83	0.0000	11	80.42	0.1111	11	78.33	0.0000	12	82.75	0.0000	12	81.83	0.0000	12	81.75	0.0000	12	81.00	0.0000
4	105401	11	10	81.27	0.1250	10	80.18	0.1250	10	78.73	0.1250	11	78.00	0.0000	9	75.45	0.2500	9	67.55	0.1250	10	81.18	0.1250	10	79.55	0.1250	10	81.36	0.1250	10	79.45	0.1250
5	105501	19	16	86.63	0.1875	16	83.11	0.2500	17	83.32	0.1250	17	86.21	0.1250	15	71.95	0.3750	14	57.16	0.5000	16	86.26	0.1875	16	81.58	0.1875	17	83.05	0.1250	17	79.74	0.1875
6	105601	19	15	78.68	0.2500	16	73.21	0.2500	15	77.53	0.2500	15	74.05	0.3125	15	69.84	0.3125	14	56.63	0.5000	15	77.89	0.3125	14	64.00	0.3750	17	77.37	0.1875	15	66.53	0.3125
7	105701	22	19	75.32	0.2632	17	61.41	0.4211	19	76.64	0.2632	18	55.18	0.3684	19	70.45	0.2105	17	44.95	0.4211	18	71.68	0.2632	18	55.05	0.3684	19	73.82	0.2632	18	56.73	0.3684
8	105801	18	16	79.67	0.1429	12	67.67	0.5000	16	72.00	0.1429	12	59.00	0.5000	16	77.28	0.1429	12	52.06	0.5000	16	83.22	0.1429	16	60.44	0.2143	16	80.22	0.1429	15	66.22	0.2857
9	105901	14	13	74.14	0.0909	13	70.93	0.0909	13	75.86	0.0909	13	69.79	0.0909	13	74.79	0.0909	13	61.93	0.0909	13	79.36	0.0909	13	74.79	0.0909	13	73.29	0.0909	13	69.86	0.0909
10	106301	15	14	75.60	0.0833	13	68.67	0.1667	13	63.07	0.1667	14	67.13	0.0833	13	71.93	0.1667	11	54.07	0.5000	14	75.93	0.0833	12	56.60	0.3333	13	74.07	0.1667	13	63.40	0.1667
11	101202	9	9	71.89	0.0000	8	67.44	0.1667	9	74.11	0.0000	9	65.56	0.0000	9	71.56	0.0000	8	66.78	0.1667	8	75.44	0.1667	9	59.44	0.0000	9	69.89	0.0000	8	52.56	0.1667
12	101602	20	19	81.15	0.0588	19	83.25	0.0588	19	83.30	0.0588	19	82.75	0.0588	19	80.80	0.0588	19	77.50	0.0588	17	81.35	0.1765	17	78.30	0.1765	19	85.85	0.0588	19	85.25	0.0588
13	102802	16	14	70.81	0.2308	14	70.63	0.2308	14	70.00	0.2308	14	64.75	0.2308	15	75.06	0.1538	13	64.31	0.3077	15	72.44	0.1538	14	72.00	0.2308	14	73.13	0.2308	14	72.75	0.2308
14	102202	9	8	74.89	0.1667	8	74.89	0.1667	8	75.89	0.1667	8	75.67	0.1667	8	74.78	0.1667	8	75.22	0.1667	8	75.33	0.1667	8	74.78	0.1667	8	75.67	0.1667	8	75.22	0.1667
15	300002	17	14	80.53	0.2857	14	69.88	0.2857	15	79.41	0.2143	14	69.82	0.2857	15	79.88	0.2143	13	70.59	0.4286	13	77.00	0.3571	15	72.94	0.2143	15	73.59	0.2143	15	69.47	0.2143
16	312802	12	11	73.92	0.1111	11	73.83	0.1111	11	76.50	0.1111	11	69.33	0.1111	10	83.25	0.2222	9	70.50	0.3333	10	80.25	0.1111	10	77.17	0.2222	11	71.58	0.1111	11	68.08	0.0000
17	312902	12	11	73.75	0.1111	11	71.17	0.1111	11	74.25	0.1111	10	71.83	0.2222	9	73.00	0.3333	9	66.17	0.3333	10	74.50	0.2222	10	72.42	0.2222	11	73.67	0.1111	11	67.42	0.1111
18	313002	6	6	66.67	0.0000	6	66.67	0.0000	6	66.67	0.0000	6	66.33	0.0000	6	60.50	0.0000	6	55.17	0.0000	6	66.33	0.0000	6	66.50	0.0000	6	60.67	0.0000	6	59.83	0.0000
19	314302	17	13	73.12	0.3571	14	70.47	0.2857	14	68.12	0.2857	14	67.06	0.2857	13	71.24	0.3571	13	60.41	0.3571	13	73.59	0.3571	13	74.94	0.3571	14	70.76	0.2143	14	69.12	0.2857
20	314902	17	14	81.94	0.2857	13	70.06	0.4286	15	75.24	0.2143	14	71.65	0.2857	14	76.53	0.2857	12	40.65	0.5000	14	85.35	0.3571	14	66.12	0.3571	15	75.59	0.2143	14	63.29	0.2857
21	101203	15	12	72.93	0.3333	12	69.07	0.2500	13	79.40	0.2500	13	76.13	0.2500	13	54.80	0.1667	11	35.93	0.6667	13	79.13	0.2500	13	68.87	0.2500	14	78.80	0.1667	13	72.07	0.2500
22	105403	6	6	66.17	0.0000	6	62.83	0.0000	6	66.67	0.0000	6	66.50	0.0000	6	64.17	0.0000	6	60.00	0.0000	6	66.50	0.0000	6	66.50	0.0000	6	66.67	0.0000	6	66.33	0.0000
23	109903	19	17	82.89	0.1875	17	76.63	0.1875	17	81.74	0.1875	18	71.63	0.1250	17	74.58	0.1875	14	50.42	0.5625	17	78.37	0.1875	15	56.26	0.4375	17	82.16	0.1875	14	57.84	0.5000
24	105603	8	6	58.50	0.4000	8	52.13	0.0000	6	60.50	0.4000	6	58.13	0.4000	6	59.50	0.4000	5	55.13	0.6000	6	60.75	0.2000	6	50.00	0.2000	6	62.75	0.2000	6	56.75	0.4000
25	314703	9	8	66.44	0.1667	8	65.67	0.1667	9	61.78	0.0000	9	62.22	0.0000	8	60.11	0.1667	6	44.67	0.5000	8	71.11	0.1667	8	66.89	0.1667	9	64.56	0.0000	8	66.89	0.1667
26	315303	11	10	81.18	0.1250	10	77.55	0.1250	10	81.45	0.1250	10	77.27	0.1250	10	80.73	0.1250	10	65.91	0.1250	10	81.55	0.1250	10	75.91	0.1250	10	81.36	0.1250	9	75.45	0.2500
27	316903	8	8	74.88	0.0000	7	65.50	0.2000	8	75.00	0.0000	8	74.50	0.0000	7	74.25	0.2000	7	70.88	0.2000	7	68.88	0.2000	7	72.50	0.2000	8	69.75	0.0000	8	64.00	0.0000
28	331103	12	11	64.67	0.2222	10	60.50	0.2222	11	66.33	0.2222	11	59.17	0.1111	11	61.92	0.1111	10	53.25	0.2222	11	64.25	0.1111	11	57.17	0.1111	11	72.17	0.1111	10	55.42	0.3333
29	331903	10	10	77.60	0.0000	10	74.80	0.0000	10	75.90	0.0000	9	69.20	0.1429	9	70.80	0.1429	9	50.60	0.1429	10	78.50	0.0000	10	75.60	0.0000	10	78.00	0.0000	9	70.20	0.1429
30	332003	9	8	73.22	0.1667	8	77.44	0.1667	9	76.67	0.0000	9	72.56	0.0000	8	74.89	0.1667	8	70.56	0.1667	8	74.89	0.1667	9	73.00	0.0000	9	74.00	0.0000	9	73.44	0.0000
	Average	13.13	11.67	74.56	0.15	11.37	70.70	0.18	11.83	73.60	0.13	11.67	69.72	0.15	11.40	71.61	0.19	10.47	60.15	0.30	11.37	75.07	0.17	11.33	68.72	0.19	11.97	73.90	0.12	11.47	67.97	0.18

Note: The protein families were randomly selected from TreeFam-A.

Table S4. The Baliscore of different alignments for BaliBase BB11001

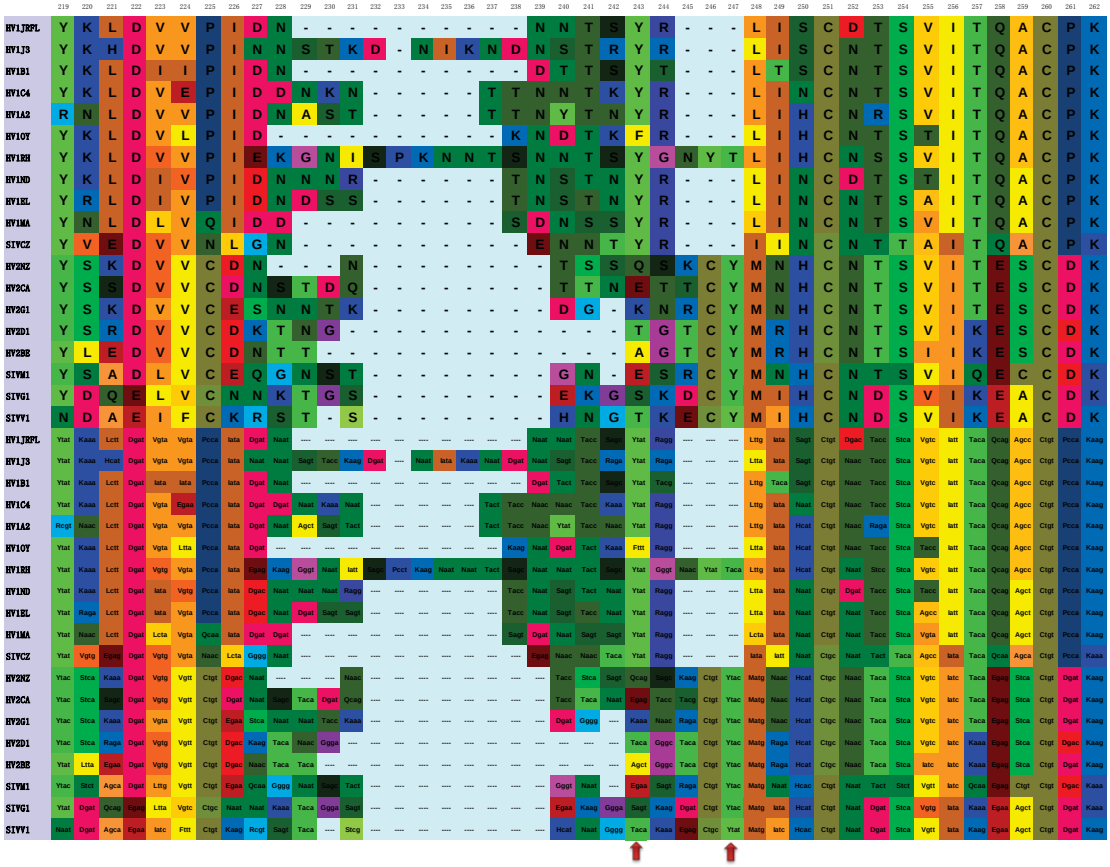
Alignment method	SP	TC
ClustalW	0.947	0.895
muscle	0.904	0.855
T-Coffee	0.943	0.895
prank	0.770	0.605
prank-codon	0.026	0
CAUSA	0.432	0.289

Table S5. The computation time of different alignments

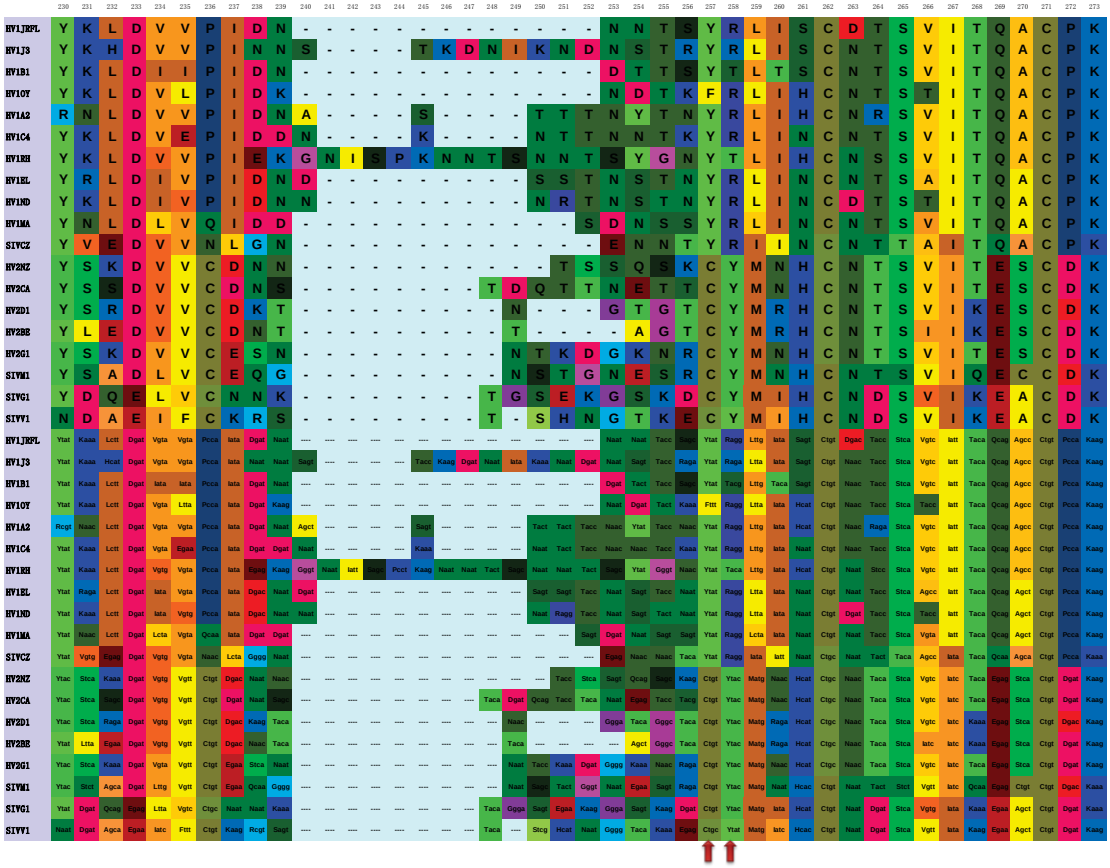
No.	Number of Sequences	Total Length (kb)	Computation time (*Sec*)					
			PRANK	T-coffee	MUSCLE	MAFFT	ClustalW	CAUSA
1	6	17	918	18	32	6	25	24
2	8	12	350	10	13	1	12	13
3	9	9	577	27	6	2	21	4
4	9	10	220	7	7	2	11	11
5	9	5	25	2	1	0	1	1
6	9	5	134	3	2	1	3	2
7	10	26	1676	42	87	8	82	54
8	11	27	1770	34	127	7	63	59
9	12	45	4560	75	342	20	182	171
10	12	59	7782	244	641	34	321	324
11	12	34	2690	43	96	7	94	107
12	12	22	1301	27	54	5	40	39
13	14	22	12112	72	32	4	277	33
14	14	13	266	13	6	2	11	14
15	14	61	11717	415	309	26	631	227
16	15	19	628	21	26	4	26	23
17	16	39	4753	118	64	9	127	118
18	16	11	169	11	5	2	10	7
19	17	40	3449	100	177	14	146	109
20	17	36	2215	47	100	7	101	70
21	17	16	534	20	13	3	25	20
22	18	19	1072	22	19	5	45	22
23	19	21	570	25	19	2	26	20
24	19	70	11563	162	618	31	410	267
25	19	22	1021	29	19	3	63	31
26	19	21	784	21	14	3	40	32
27	20	21	2706	23	18	3	36	44
28	22	26	1052	36	24	4	45	43
29	27	19	958	19	8	3	41	30
30	30	28	3054	70	37	4	68	42
Total		775	80626	1756	2916	222	2983	1961
Average			104.0335	2.2658	3.7626	0.2865	3.8490	2.5303

Supplementary Figures

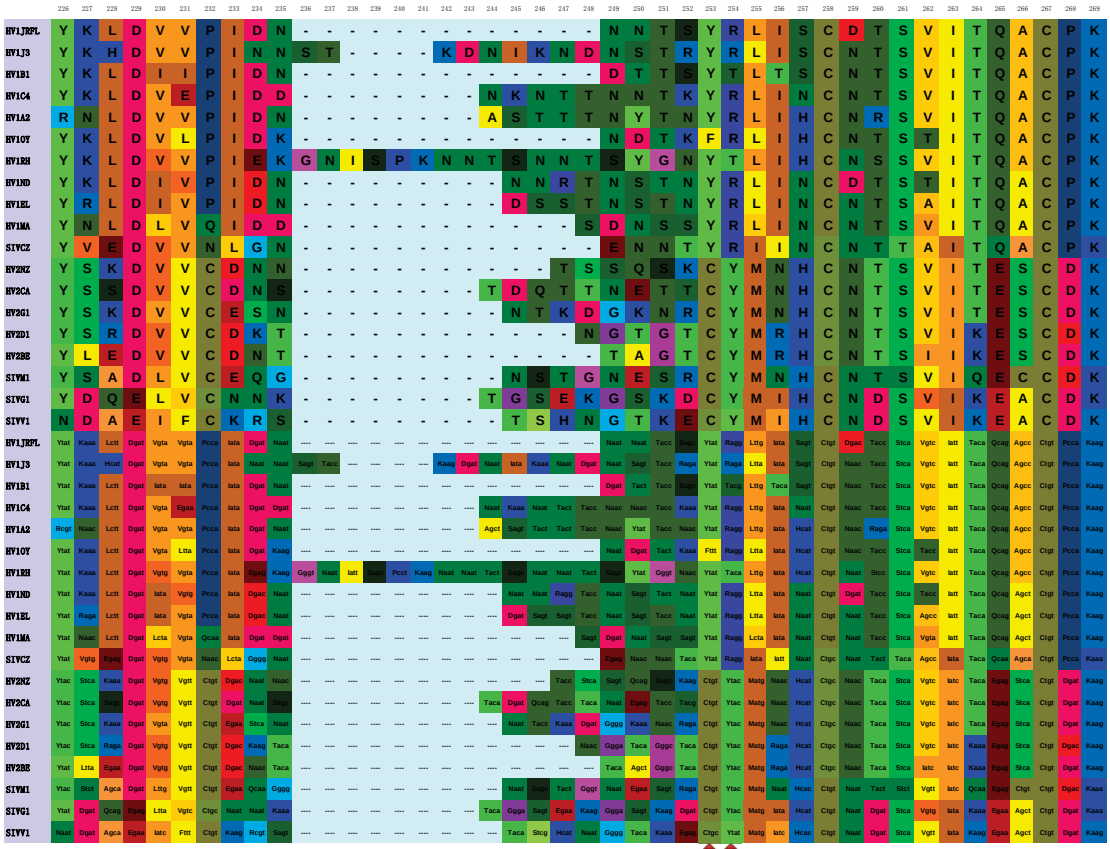
A
Clustalw



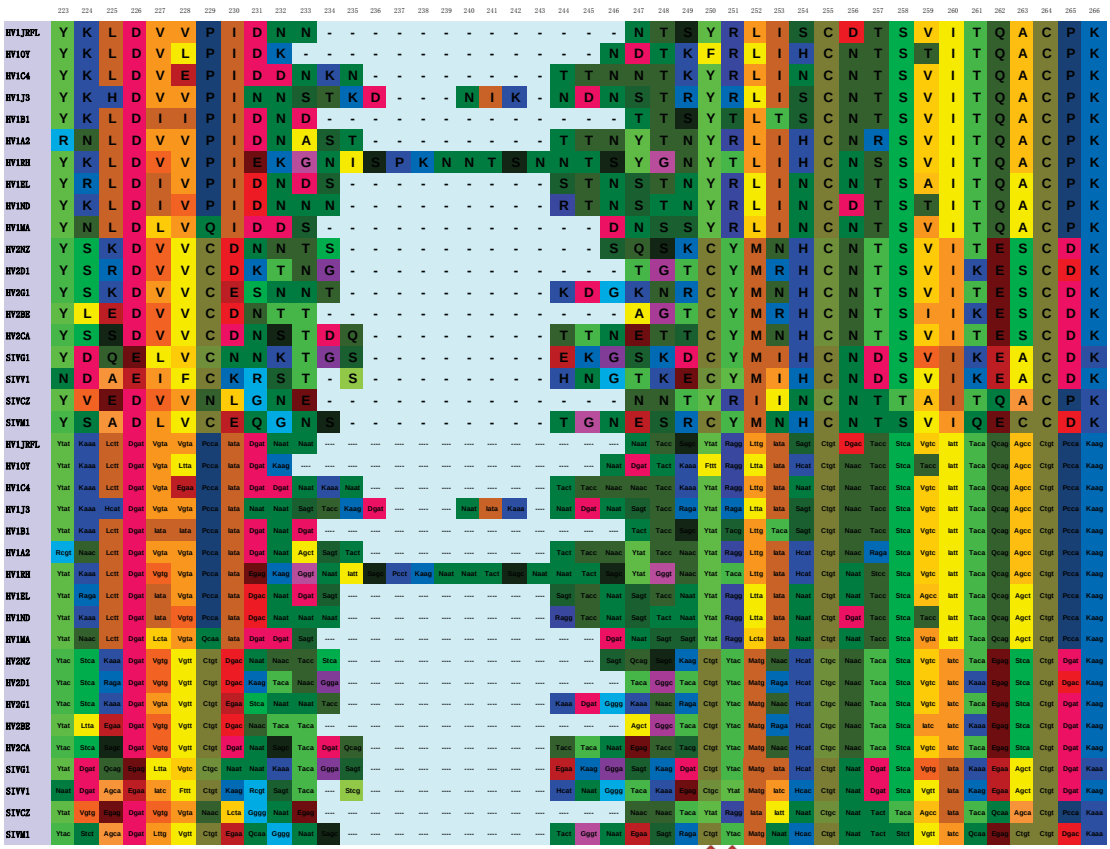
B
Mafft



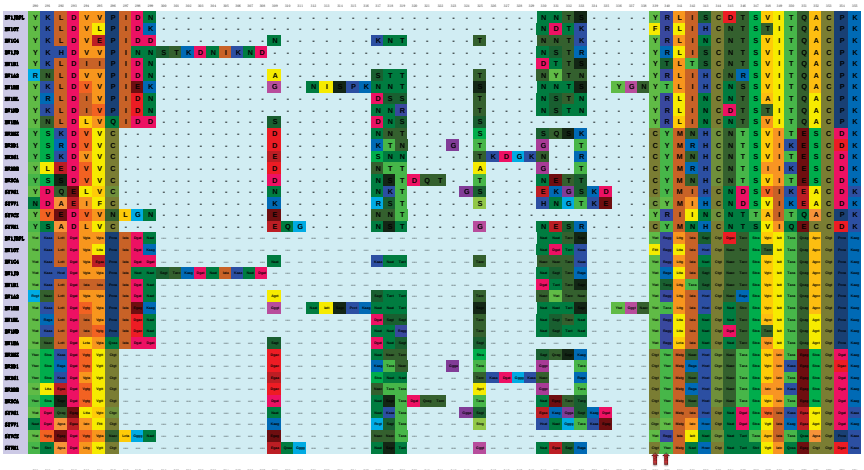
C
MUSCLE



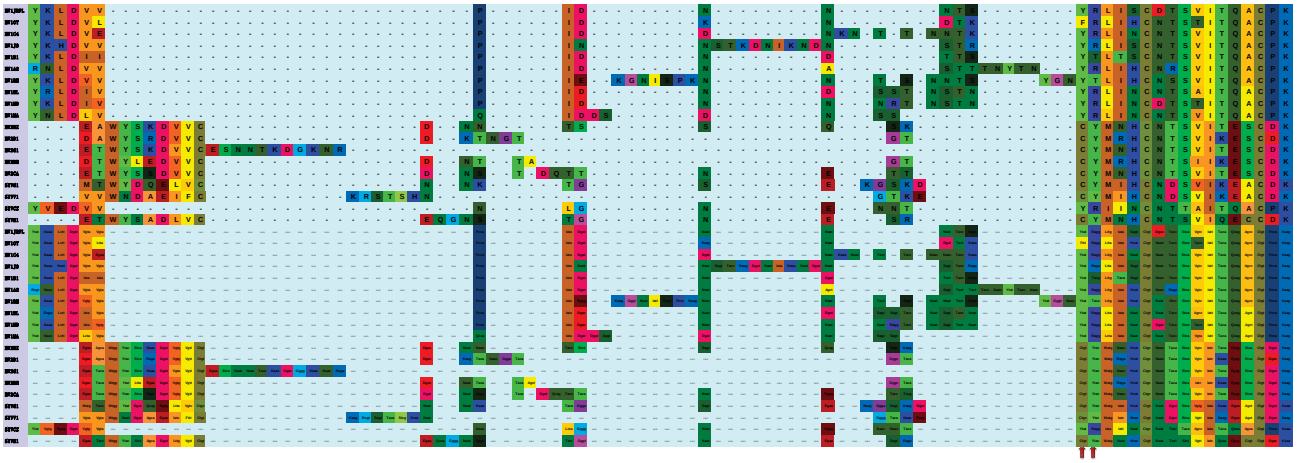
D
T-coffee



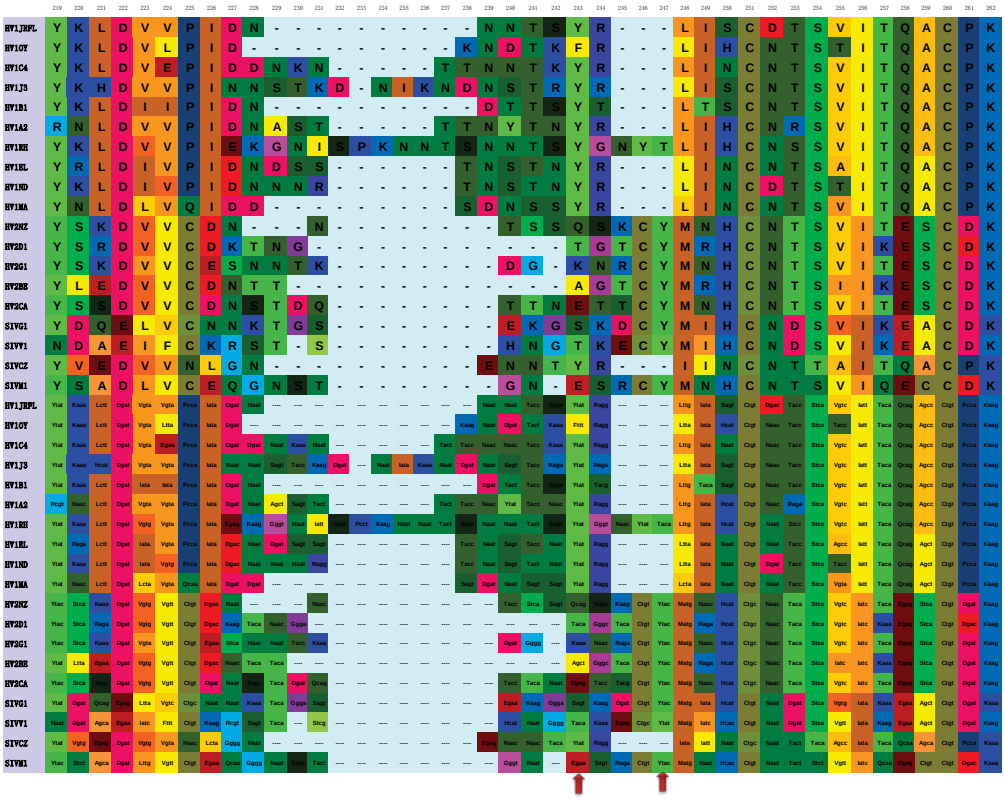
E
Prank



F
Prank-codon



G
CAT



H
CAUSA

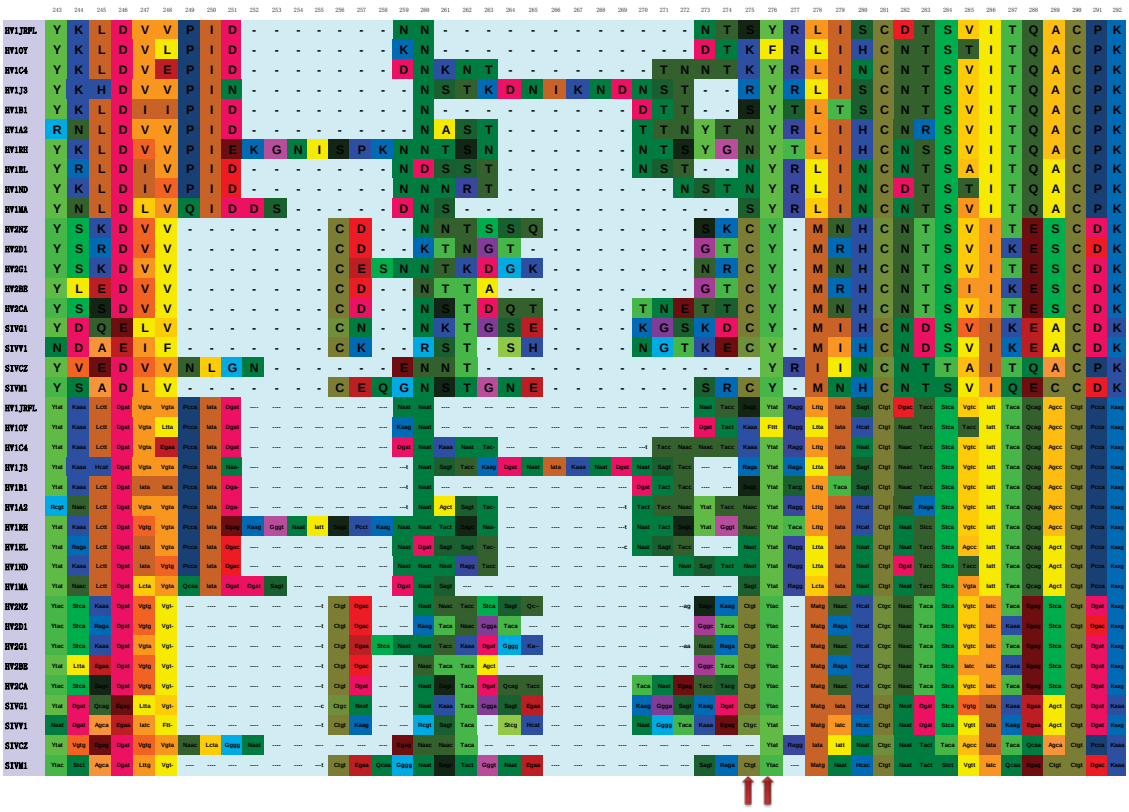


Fig. S1. Comparison of alignments for HIV Env generated by different programs.
(A) ClustalW, (B) MAFFT, (C) MUSLE, (D) T-coffee, (E) PRANK, (F) PRANK-Codon, (G) CAT, (H) CAUSA. HIV or SIV strains were derived from the seed alignment of Pfam gp120 protein family (pf00516). DNA and protein sequences are written respectively in lowercase and uppercase letters.

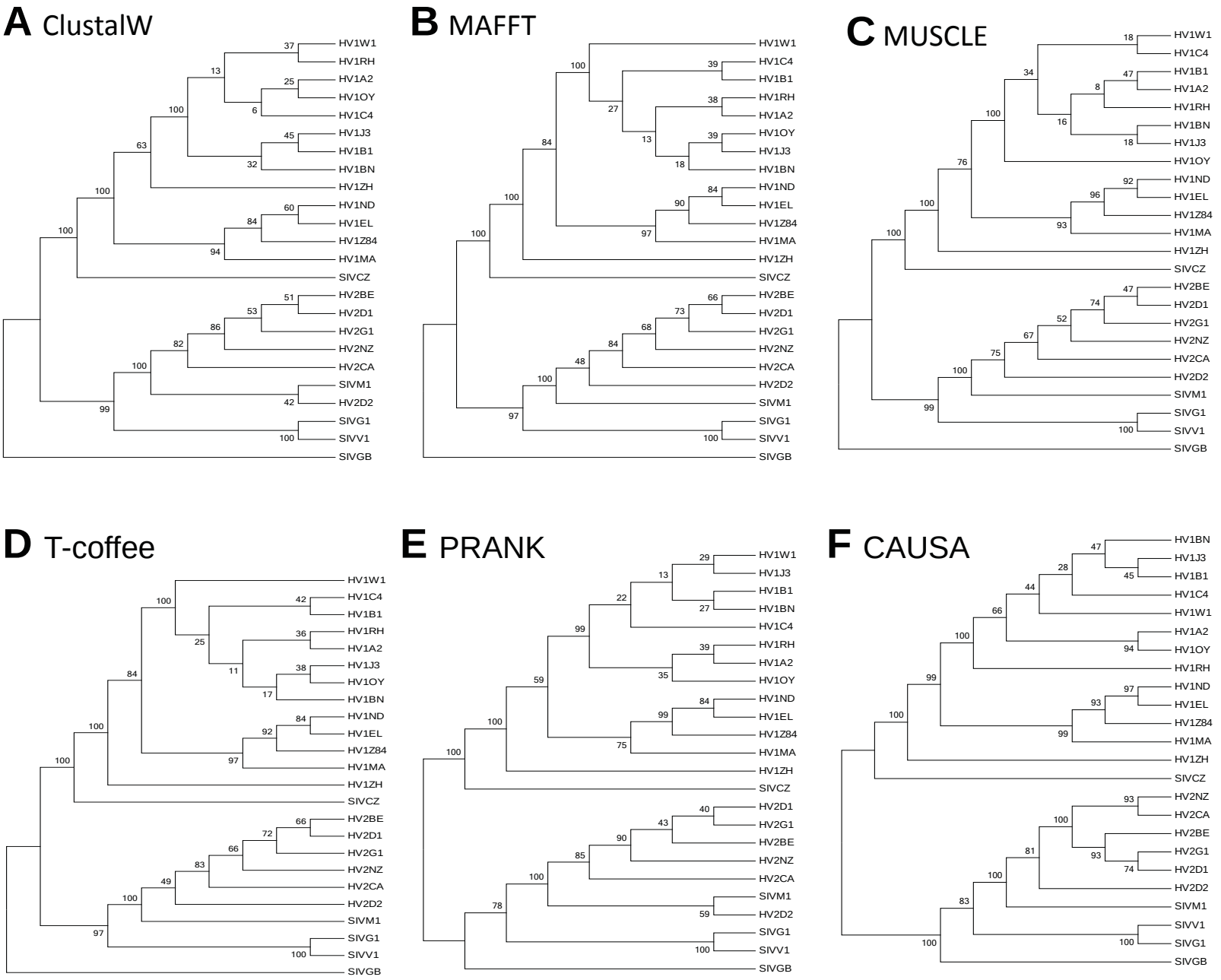


Fig. S2. Comparison of phylogenetic trees for env built from different alignments.
(A) ClustalW, (B) MAFFT, (C) MUSLE, (D) T-coffee, (E) PRANK, (F) CAUSA.

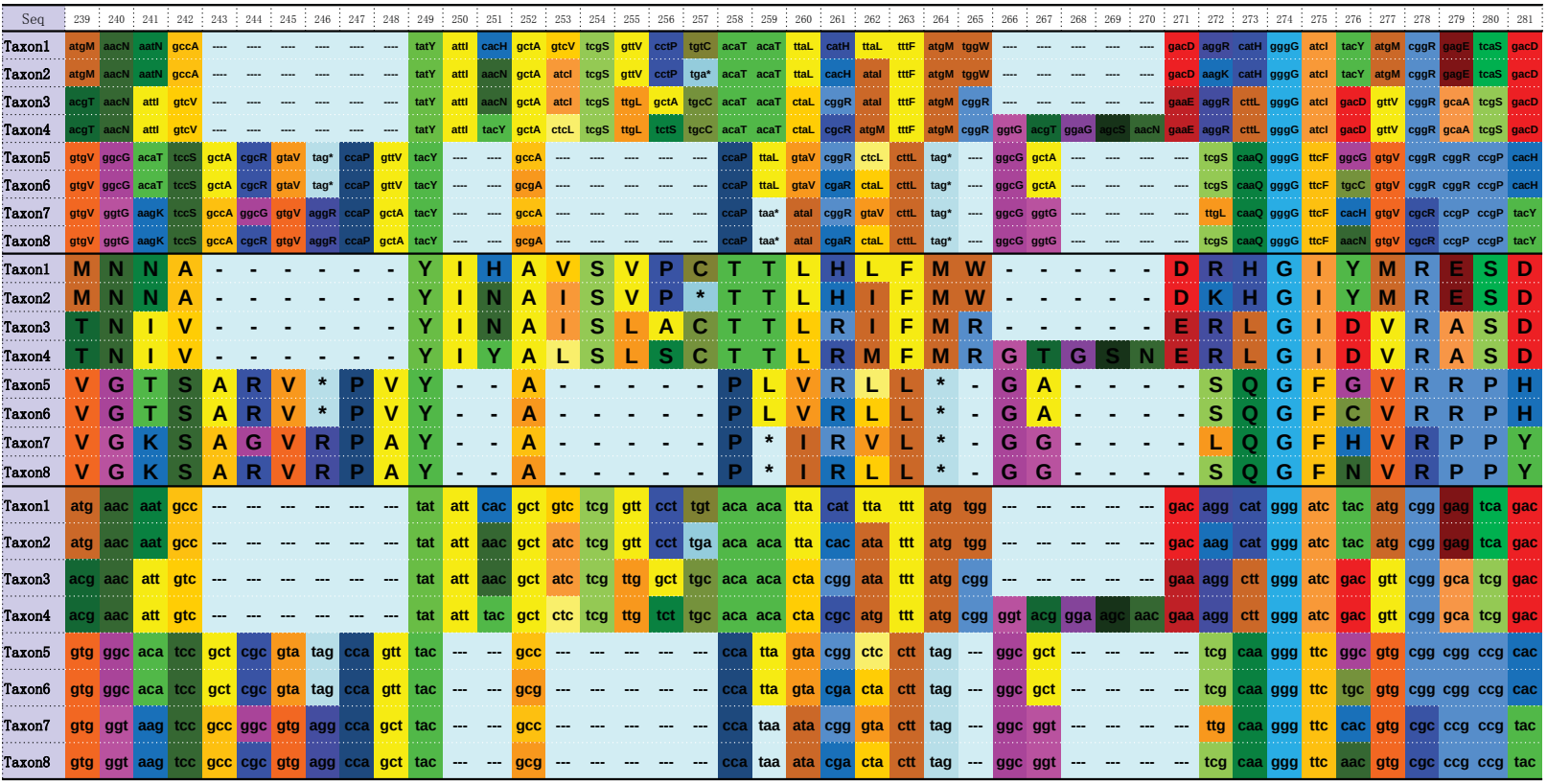


Fig. S3. The CAUSA alignments for a set of CDSs simulated by indel-seq-gen.