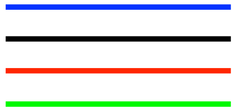
Multiple Sequence Alignments

Mark Stenglein, MIP 280A4

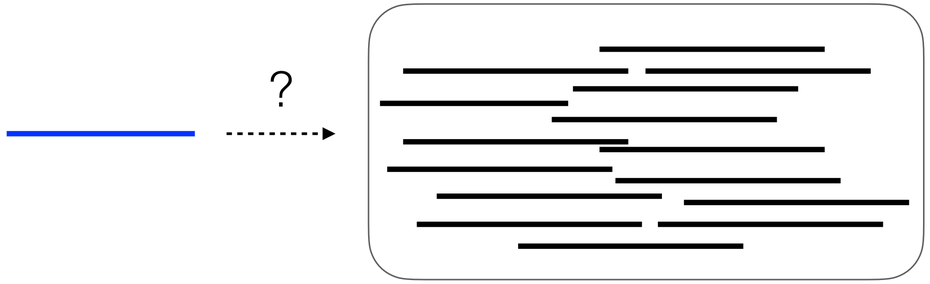
Today we will learn about alignment-based search (really: BLAST)

Pairwise alignment

Multiple sequence alignment

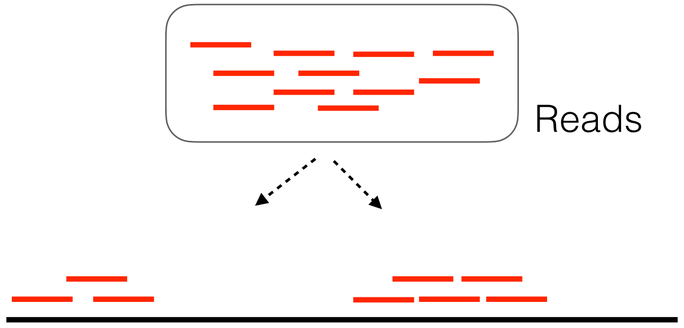


Alignment-based database searches



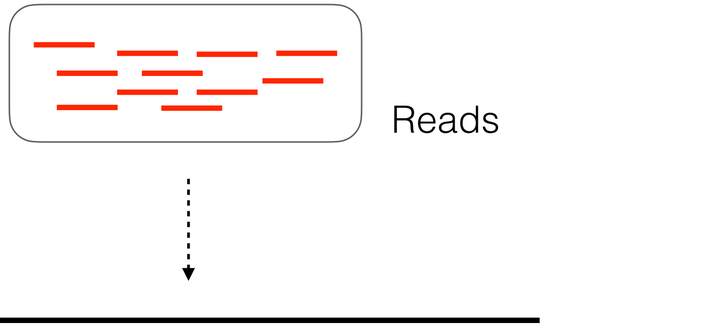
Database

Alignment to reference (mapping)



Existing Reference sequence

Assembly



New Reference sequence

What are multiple sequence alignments?

Alignments of > 2 sequences.
Can be nucleotide or protein sequences

ring-tailed cat

kinkajou

raccoon

```
ringtail  GAGGTCACACGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
raccoon   GAGGTCACGCGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
kinkajou  GAGGTCACACGCATGGTCGTCATCATGGTCATCGCATTCCTGATTTGCTGGTTGCCCT
```

Multiple sequence alignments assume that the sequences being aligned are evolutionarily related

ring-tailed cat

kinkajou

raccoon

```
ringtail   GAGGTCACACGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
raccoon    GAGGTCACGCGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
kinkajou   GAGGTCACACGCATGGTCGTCATCATGGTCATCGCATTCCTGATTTGCTGGTTGCCCT
```

This is a legitimate alignment because the sequences are homologs:
They all derive from a common ancestor

Multiple sequence alignments can identify variation in related sequences

ring-tailed cat

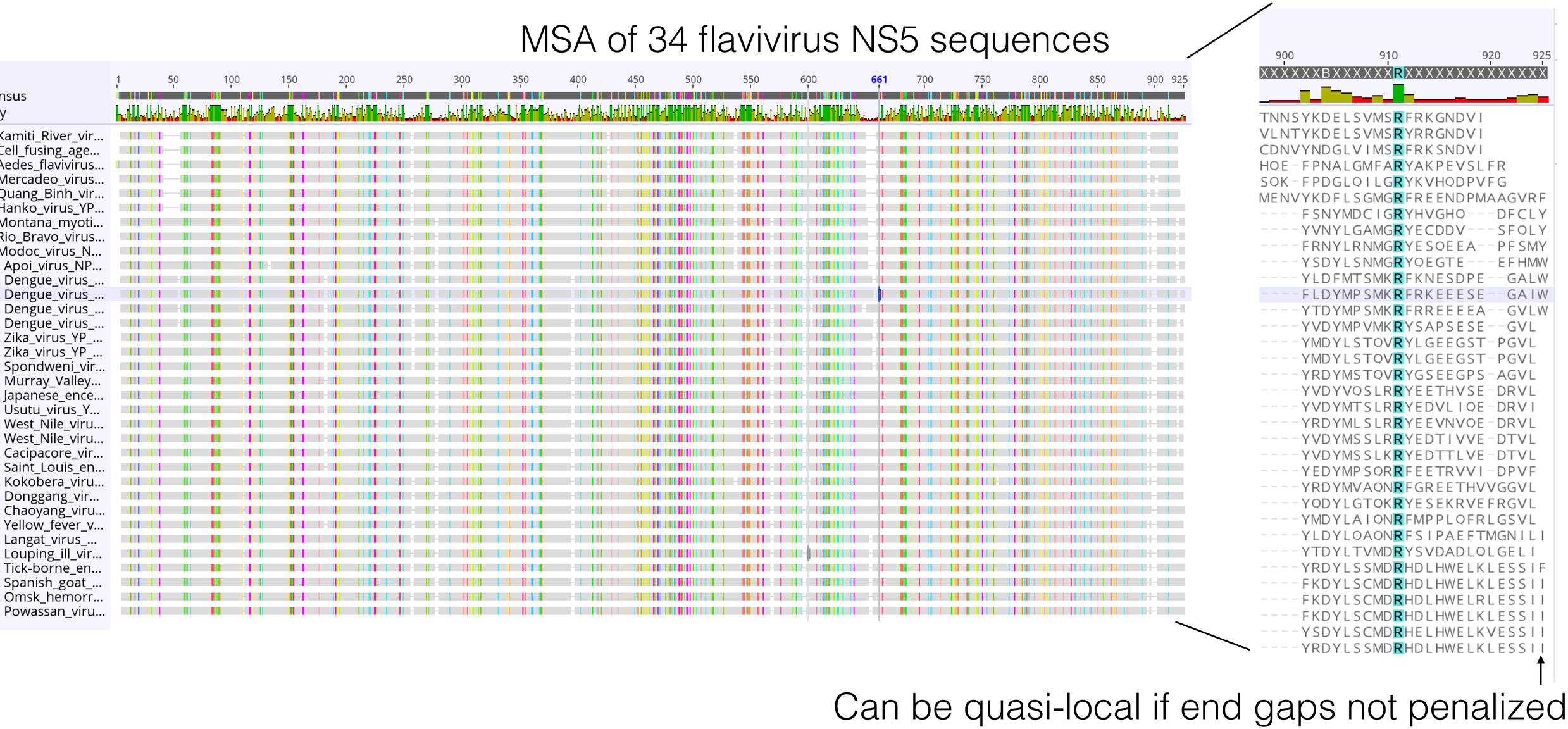
kinkajou

raccoon

```
ringtail  GAGGTCACACGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
raccoon   GAGGTCACGCGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
kinkajou  GAGGTCACACGCATGGTCGTCATCATGGTCATCGCATTCCTGATTTGCTGGTTGCCCT
```

These are single nucleotide variants (SNVs, aka single nucleotide polymorphisms, SNPS) that distinguish these 3 species

Multiple sequence alignments are end-to-end alignments (global not local)



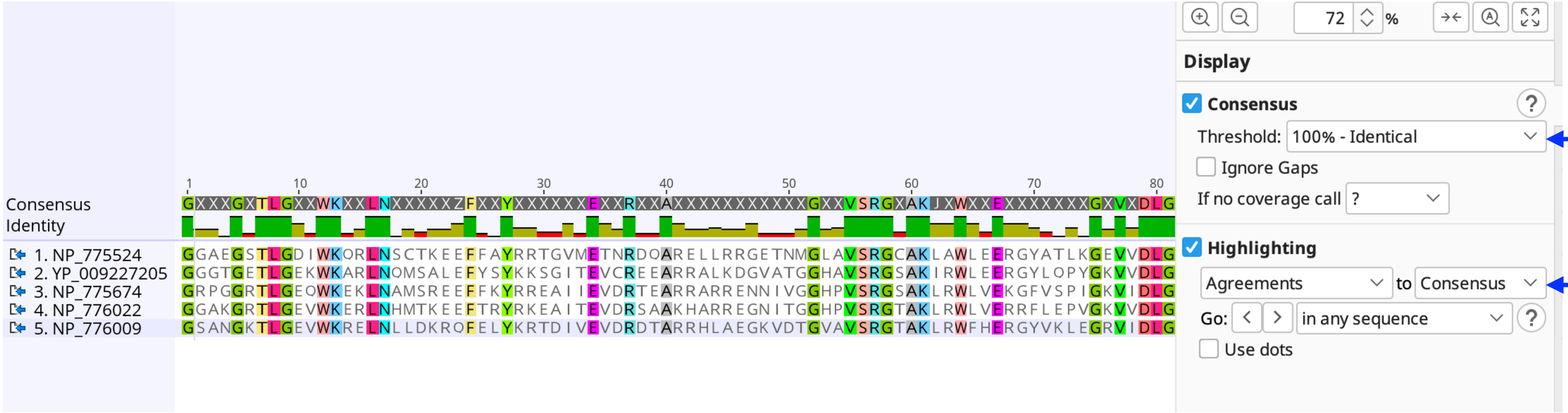
You write multiple sequence alignments as sequences: one per line (No | like in pairwise alignments)

MSA of 3 protein sequences

Pairwise alignment

```
>AB462167.1 Procyon lotor mitochondrial gene for 16S ribosomal RNA, complete
sequence, specimen_voucher: personal:Tomoharu Tokutomi:NDMC-PL-MIE18010
Length=1586

 Score = 433 bits (234),  Expect = 1e-117
 Identities = 234/234 (100%), Gaps = 0/234 (0%)
 Strand=Plus/Plus

Query  1     CAAAAACATCACCTCTAGCATTAACAGTATTAGAGGCACTGCCTGCCCAGTGACATTAGT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  840   CAAAAACATCACCTCTAGCATTAACAGTATTAGAGGCACTGCCTGCCCAGTGACATTAGT  899

Query  61    TAAACGGCCGCGGTATCCTGACCGTGCAAAGGTAGCATAATCATTTGTTCTCTAAATAGG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  900   TAAACGGCCGCGGTATCCTGACCGTGCAAAGGTAGCATAATCATTTGTTCTCTAAATAGG  959
I
Query  121   GACTTGTATGAATGGCCACACGAGGGTTTGACTGTCTCTTACTTCCAACCAGTGAAATTG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  960   GACTTGTATGAATGGCCACACGAGGGTTTGACTGTCTCTTACTTCCAACCAGTGAAATTG  1019

Query  181   ACCTTCCCGTGAAGAGGCGGGAATAAGAAAATAAGACGAGAAGACCCTATGGAG  234
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1020  ACCTTCCCGTGAAGAGGCGGGAATAAGAAAATAAGACGAGAAGACCCTATGGAG  1073
```

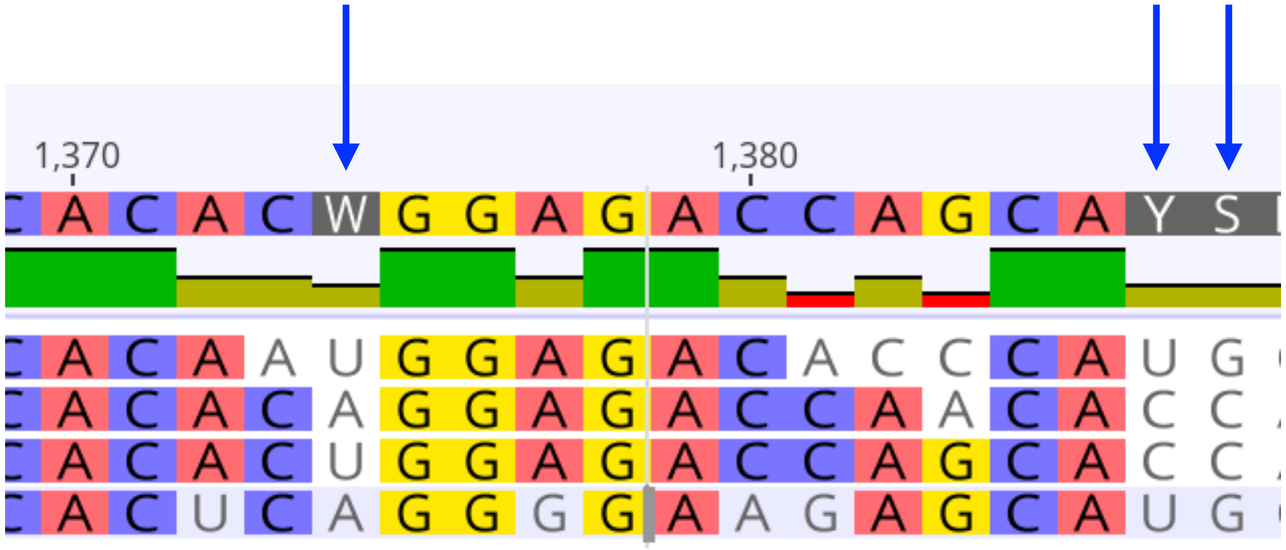
You can calculate the consensus of a multiple alignment and highlight residues that agree or disagree with the consensus



The consensus threshold defines the % of sequences that must agree for that base or amino acid to be in the consensus

You can calculate the consensus of a multiple alignment and highlight residues that agree or disagree with the consensus

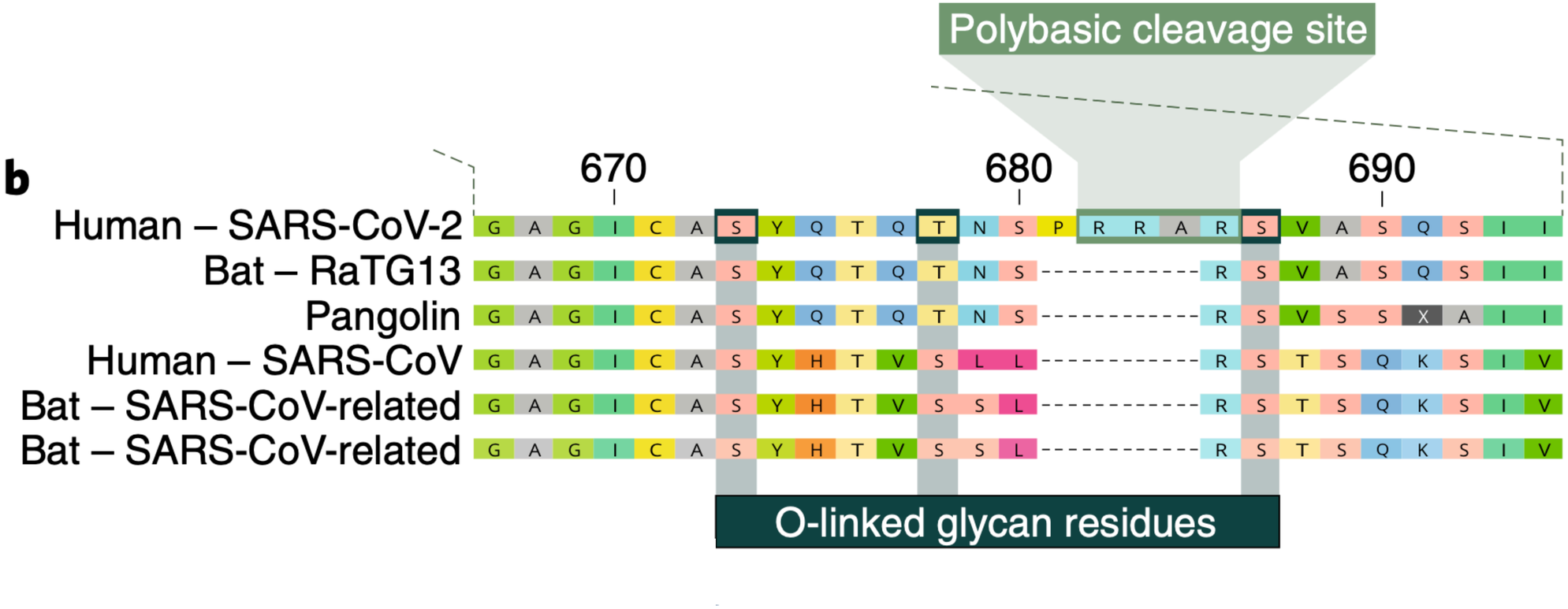
Nucleotide consensus sequences can contain ambiguous bases



IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Multiple sequence alignments can identify variation in closely related sequences

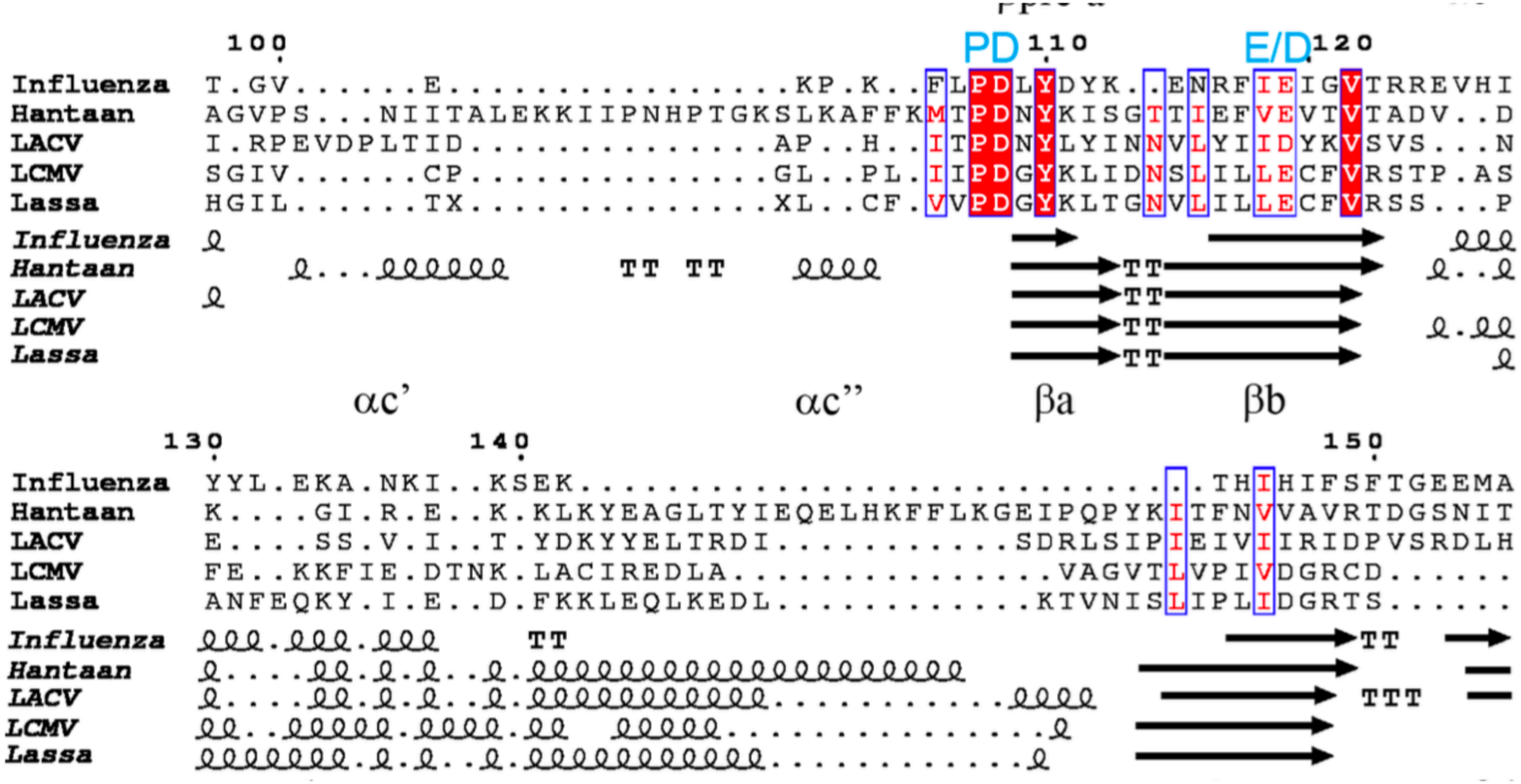
Insertion of a protease cleavage site in the spike protein of SARS-CoV-2



Andersen et al (2020) Nature

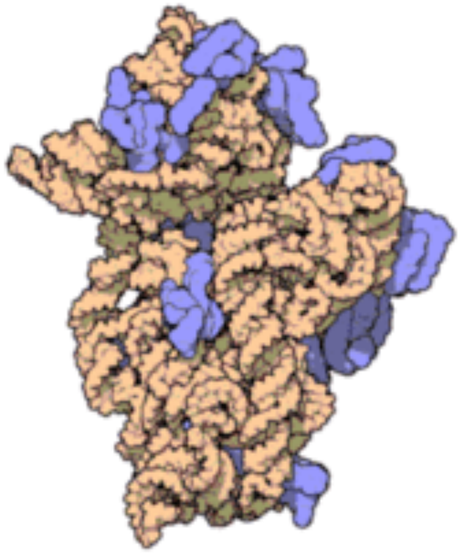
Multiple sequence alignments can also identify *similarities*, such as highly conserved amino acids, in more distantly related sequences

Highly conserved amino acids in virus nuclease sequences
(Conserved means the same base or amino acid in a high proportion of sequences aligned)



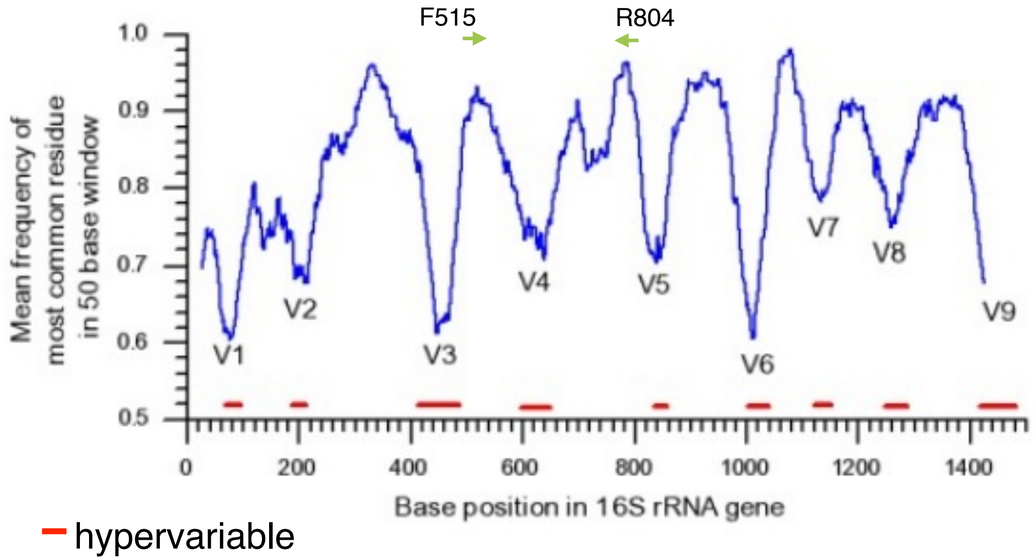
Reguera et al (2016) PLoS Path.

16S rRNA gene sequences contain conserved regions that flank variable regions



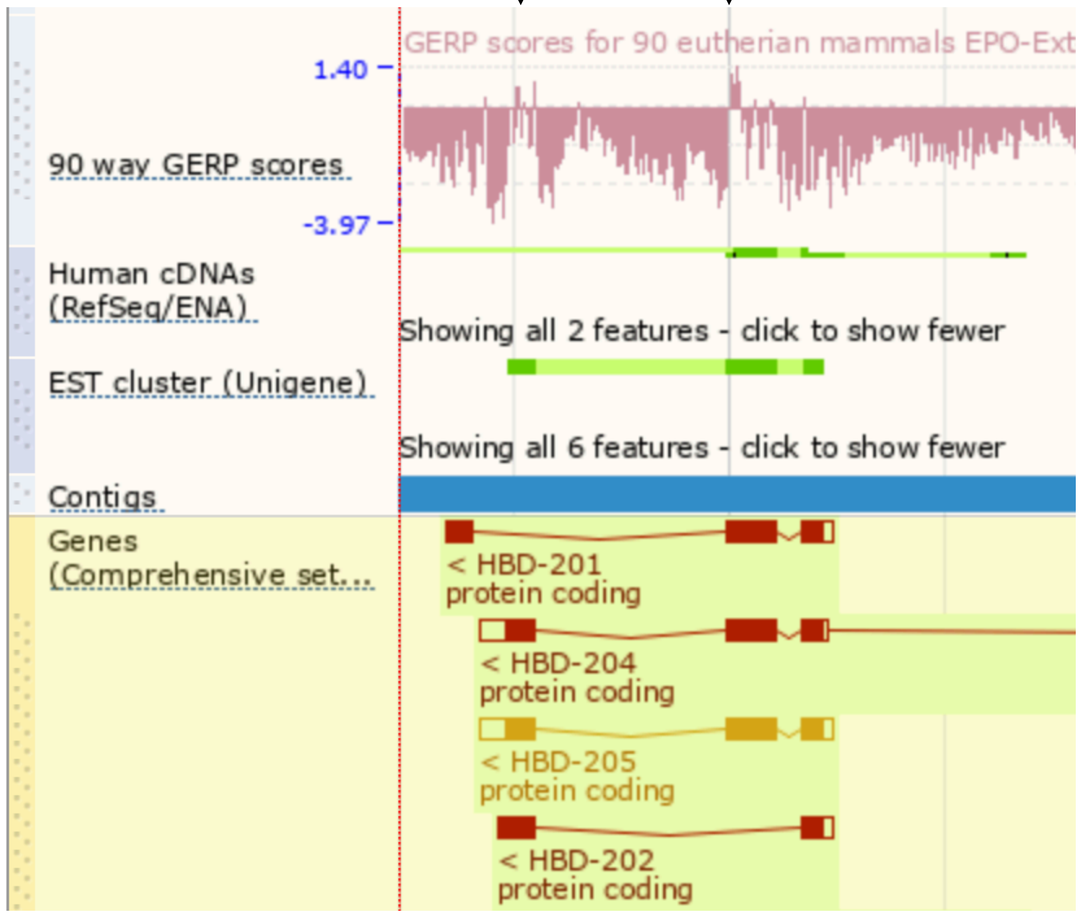
bacterial 30S ribosomal subunit
16S rRNA is in orange
(purple: ribosomal proteins)
image: wikipedia

PCR using primers that target the conserved regions amplifies the variable sequence in between. (Basis for 16S microbiome sequencing)



Conservation on a genome-wide level

conserved regions in protein coding axons



How conserved are positions in multiple sequence alignment of sequences from 90 mammals

Human hemoglobin delta gene

Ensembl.org

Multiple sequence alignments are used to produce substitution matrices like BLOSUM62

Many multiple sequence alignments of real protein sequences that share <= 62% identity

Quantifies how often amino acids are substituted for each other in real protein sequences

BLOSUM62 matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Multiple sequence alignments can be used to describe sequence motifs, like the TATA box transcription factor binding site

jaspar.genereg.net/matrix/MA0108.1/

Cart 0 | JASPAR Blog

Detailed information of matrix profile **MA0108.1**

Home > Matrix > MA0108.1

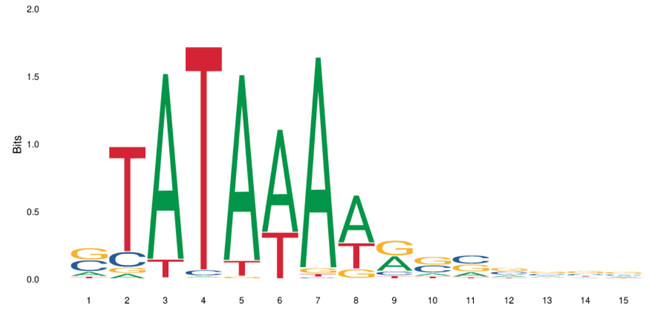
Profile summary

Add

Name:	TBP
Matrix ID:	MA0108.1
Class:	TATA-binding proteins
Family:	TBP-related factors
Collection:	CORE
Taxon:	Vertebrates
Species:	
Data Type:	
Validation:	2329577
Uniprot ID:	
Source:	
Comment:	

Sequence logo

Download SVG



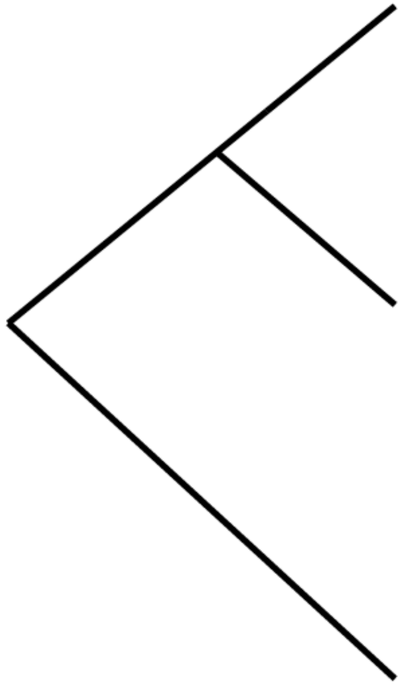
Frequency matrix

JASPAR | TRANSFAC | MEME | RAW PFM | Reverse comp.

A [	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77	]
C [	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101	]
G [	152	18	2	2	5	0	10	44	157	150	128	128	128	139	140	]
T [	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71	]

Multiple sequence alignments are the input to tree-building algorithms

GAGGTCACACGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
GAGGTCACGCGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
GAGGTCACACGCATGGTCGTCATCATGGTCATCGCATTCCTGATTTGCTGGTTGCCCT

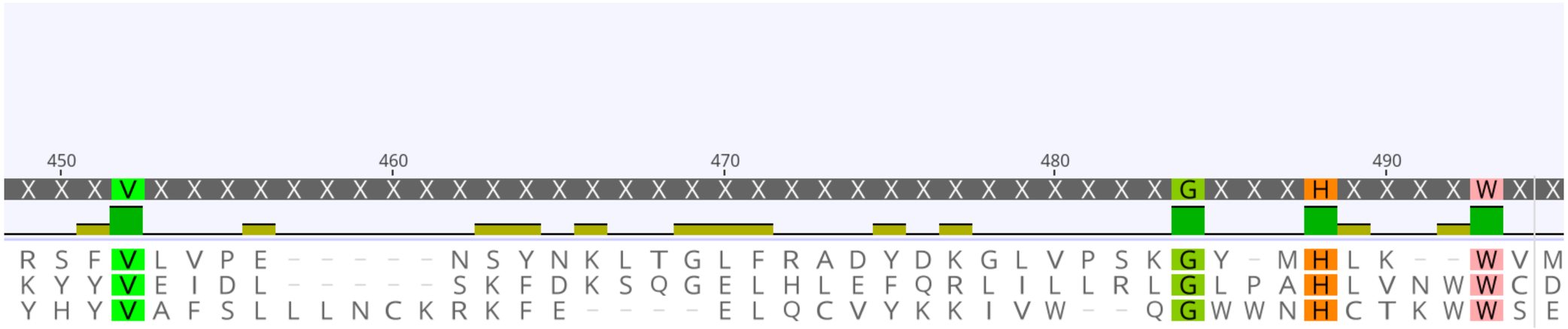


ring-tailed cat

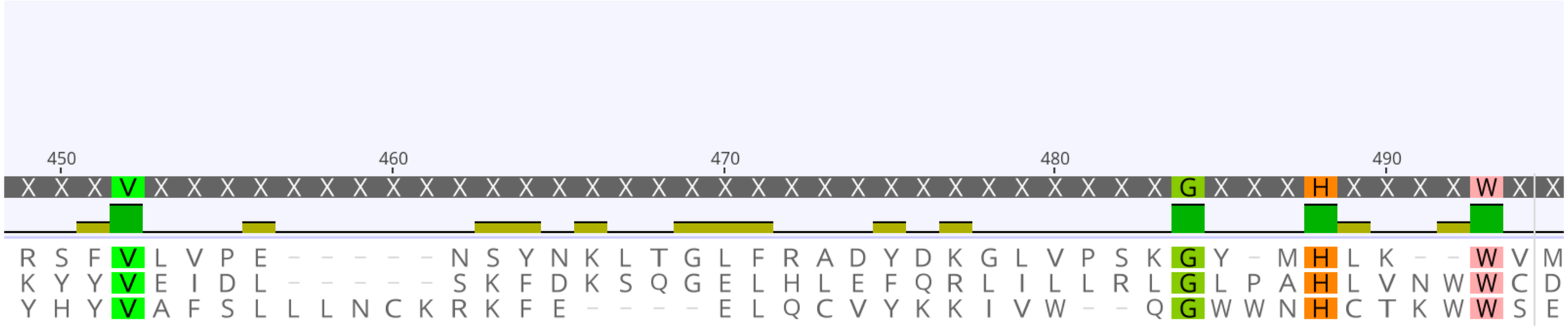
raccoon

kinkajou

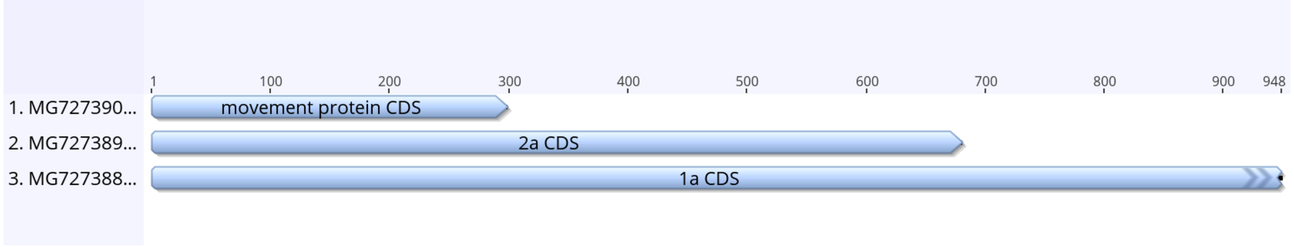
Is this a legitimate multiple sequence alignment?



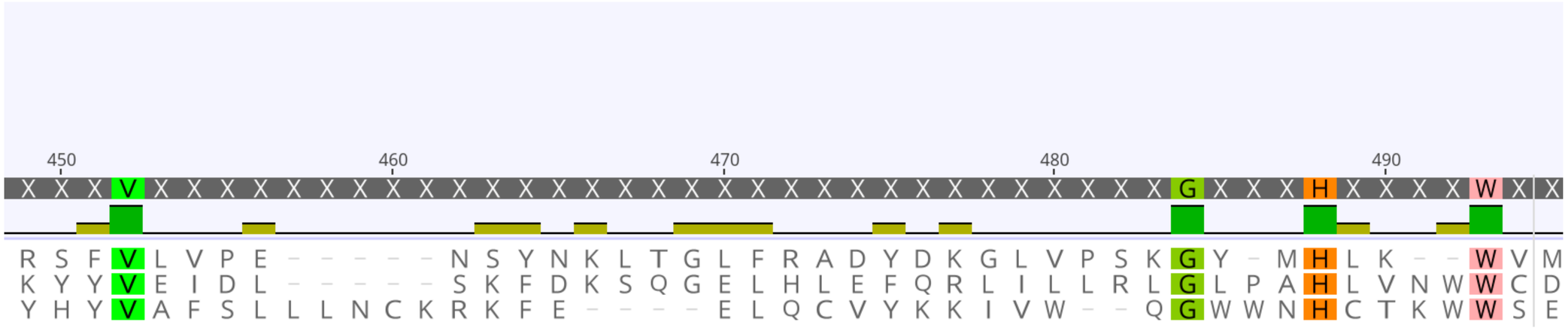
Is this a legitimate multiple sequence alignment?



Multiple sequence alignment of the 3 proteins encoded by hop yellow virus



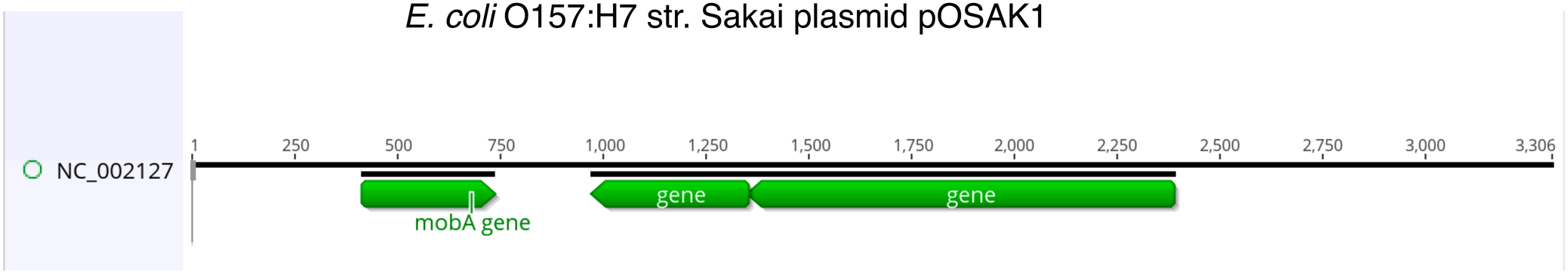
It is possible to generate biologically meaningless multiple sequence alignments



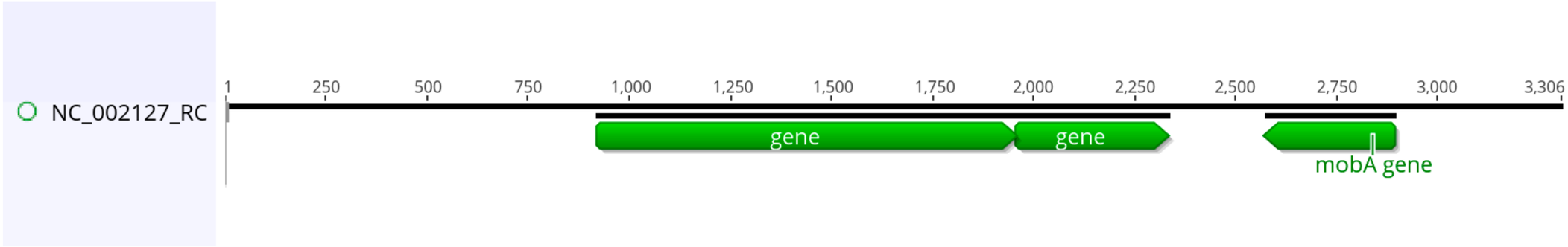
Sequence alignment algorithms are optimized to align sequences
It doesn’t mean the alignments are meaningful
So you have to be careful that you are aligning legitimately homologous sequences

DNA sequences can be written in one of 2 equally legitimate orientations
You want to make sure to align sequences in the same orientation as each other

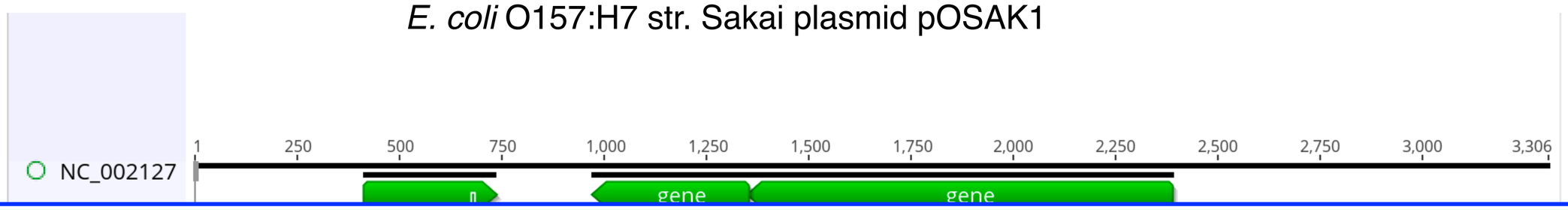
E. coli O157:H7 str. Sakai plasmid pOSAK1



E. coli O157:H7 str. Sakai plasmid pOSAK1 reverse complement

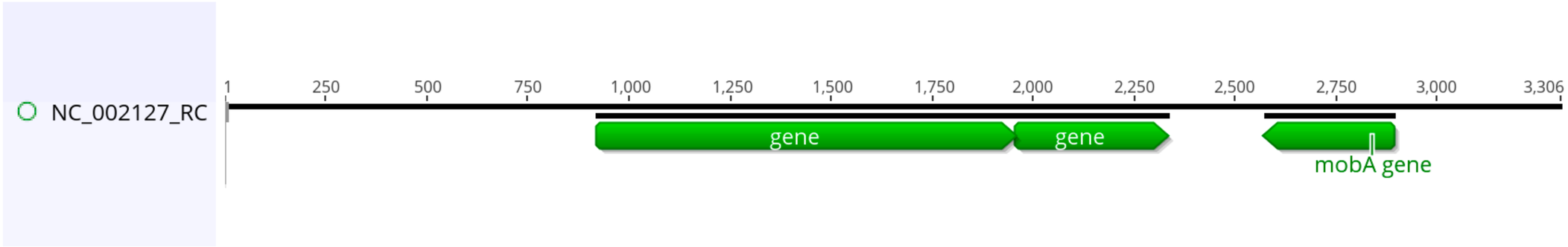


Good sequence aligners, like MAFFT, re-orient the sequences if necessary

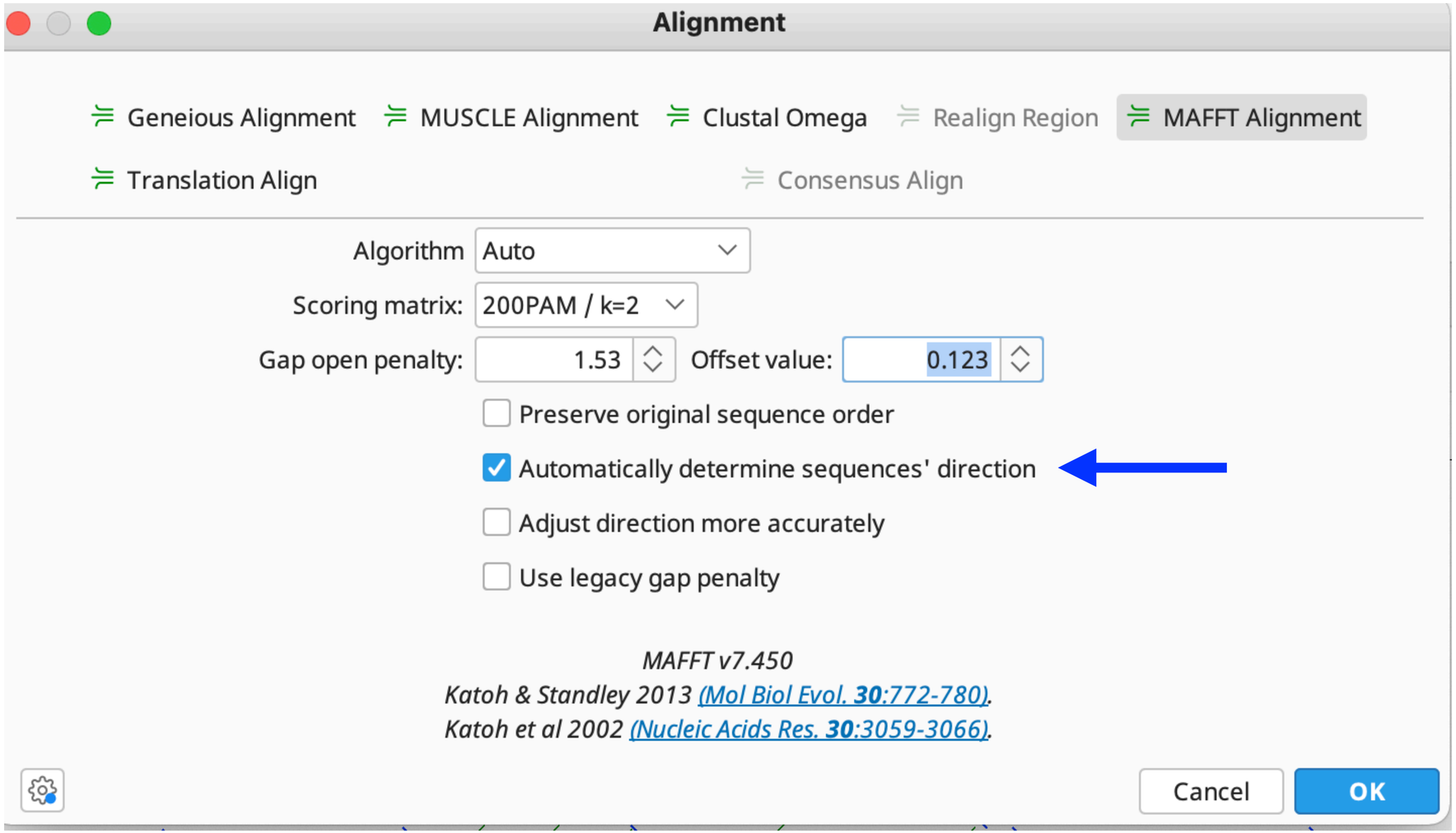


Some alignment tools will figure this out for you and will reverse complement some sequences to best match the other sequences being aligned

Protein sequences are always written in the N to C terminus orientation so this isn't an issue



The MAFFT aligner can automatically make sure DNA/RNA sequences are oriented in the same direction as each other



A simple recommendation: use MAFFT to make multiple seq. alignments

Nucleic Acids Research, 2002, Vol. 30 No. 14 **3059–3066**

MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Kazutaka Katoh, Kazuharu Misawa[1], Kei-ichi Kuma and Takashi Miyata*

Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan and
[1]Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA

Now on MAFFT v7

OXFORD

Briefings in Bioinformatics, 20(4), 2019, 1160–1166

doi: 10.1093/bib/bbx108
Advance Access Publication Date: 6 September 2017
Paper

MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization

Kazutaka Katoh, John Rozewicki and Kazunori D. Yamada

20 years of active development and improvement