Multiple Sequence Alignments

Genomics of Disease in Wildlife

Using Consensus Sequences from NGS

- **Where Are We In The Pipeline?**

  - You've sequenced your samples
  - You've mapped the reads to a reference or de novo assembly
  - You've created contigs, scaffolds, chromosome and whole genome assemblies
  - For each construct, you have a single consensus sequence
  - You can extract that sequence (under different QC requirements).
  - You now have a single sequence per sample, depending research project
  - You want to research genomes and genes of host and pathogen by **comparative, evolutionary and phylogenomic methods.**

Multiple Sequence Alignment (MSA)

Definition:

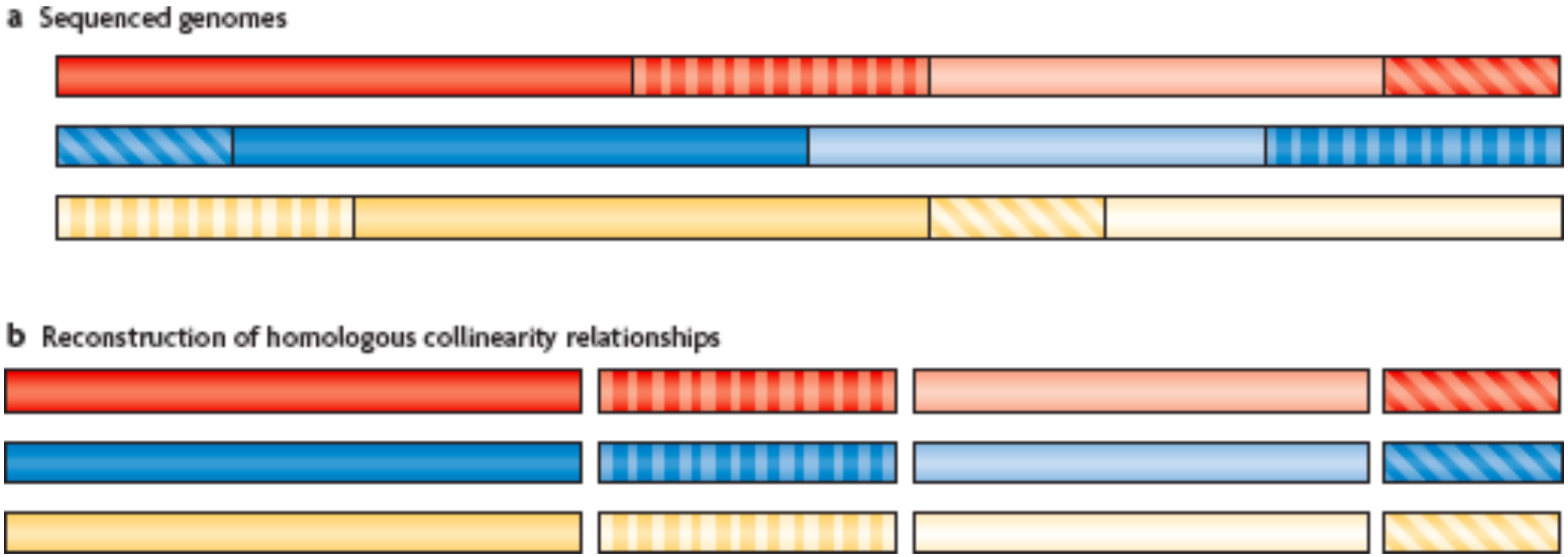
A multiple sequence alignment (**MSA**) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.

Importance:

Homology across genome sequences can be inferred from the MSA and used to assess gene or genome evolution, diversity, structure and function.

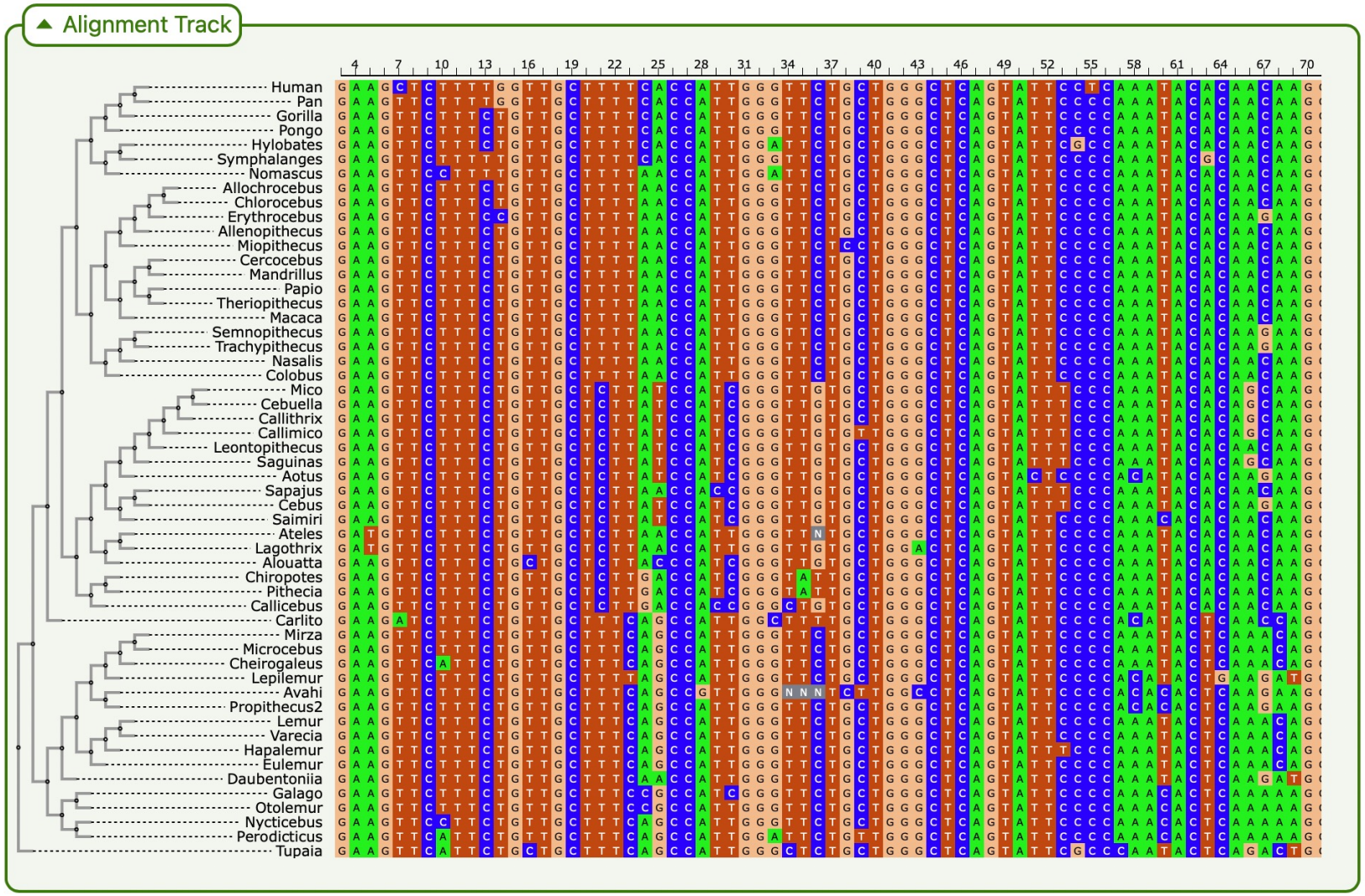
GDW
Genomics of Disease in Wildlife

Alignment Criteria: Optimal Collinearity of Homologous Regions



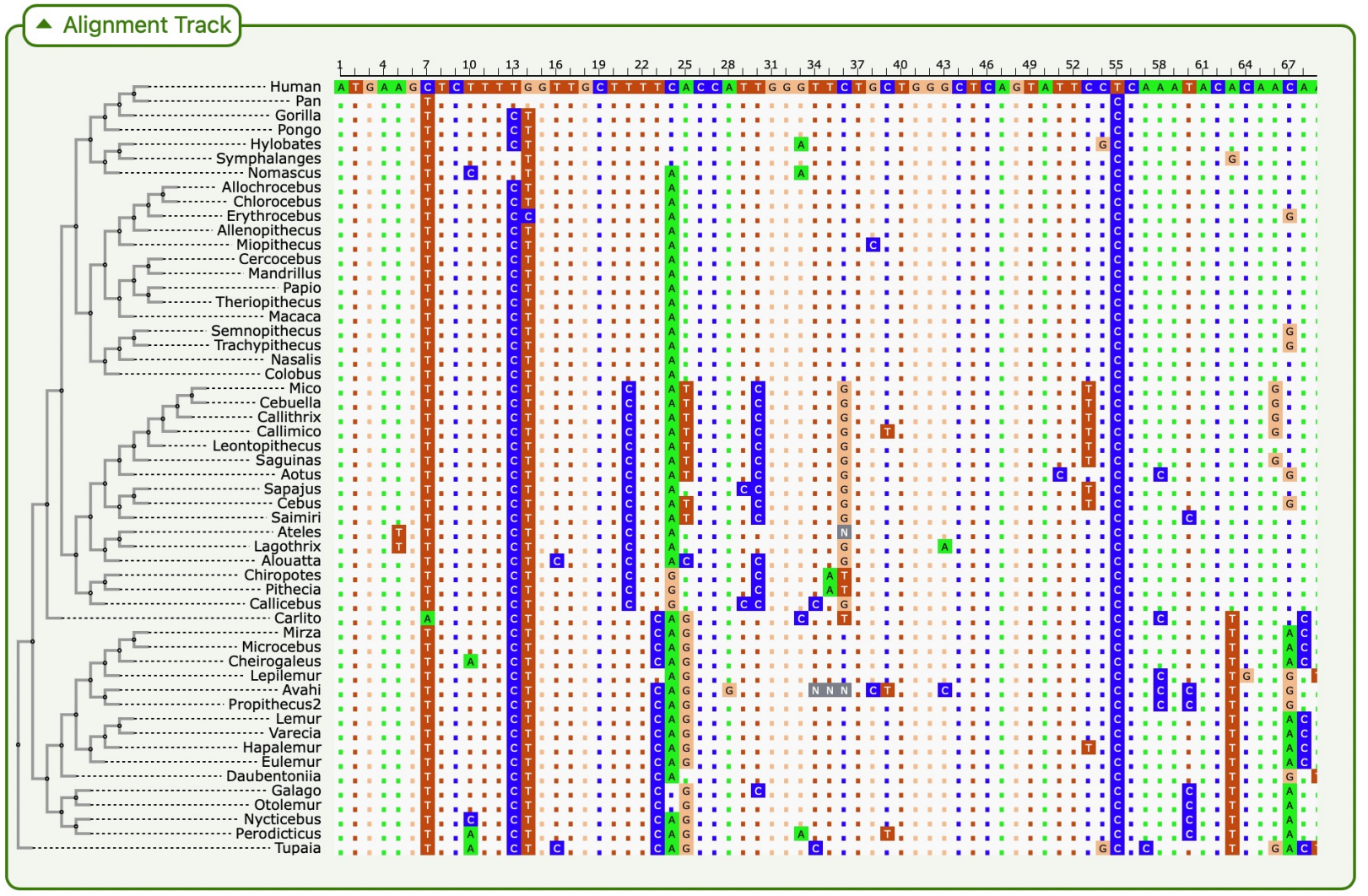
Genomics of Disease in Wildlife

Amylase Gene: Aligned in Primates (NT)



Genomics of Disease in Wildlife

Amylase Gene: Aligned in Primates (NT Matched)



Genomics of Disease in Wildlife

Alignments of Gene Families

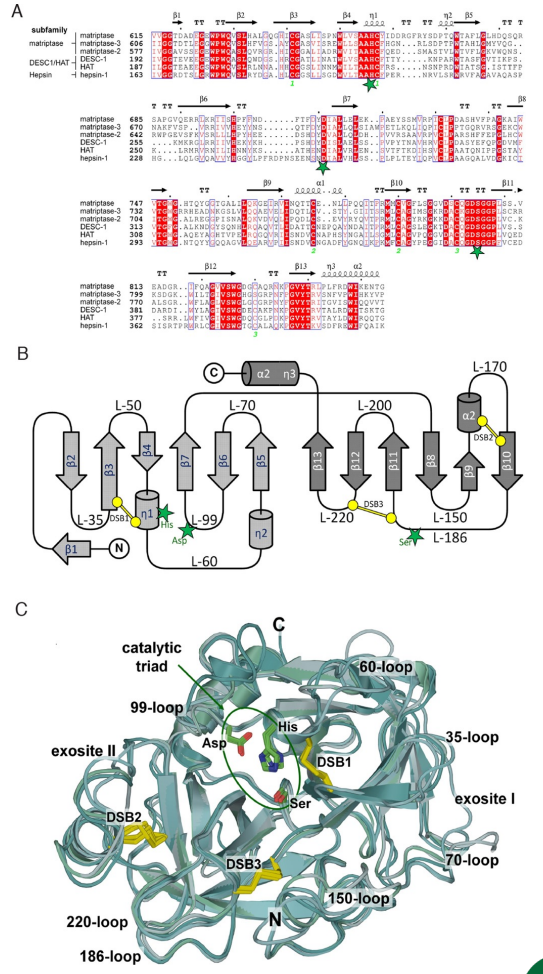
Within and Between Genomes

Many Genes Are Part of a Gene Family

Do they have different functions in biology?

How did they duplicate over time?

How do they differ in structure?



Genomics of Disease in Wildlife

True simultaneous alignment would be multi-dimensional and prohibitively computer intensive

General components to alignment process of most programs:

Series of pairwise alignments combined into multiple alignments.

Guide trees constructed for intermediate levels of addition and tested.

Different alignment programs use different ways of creating the final multiple sequence file using both global and local methods

Traditional approach uses a progressive alignment based on evolutionary relationships of sequences

Iterative alignment methods (MCMC) that also include biological features to help align (e.g. circular genomes)

GDW
Genomics of Disease in Wildlife

Commonly Used MSA Programs

- MUSCLE MUltiple Sequence Comparison by Log-Expectation
- ClustalX
- Clustal Omega*
- MAFFT Multiple Alignment Using Fast Fourier Transform*
- PRANK: Probabilistic Alignment Kit*

Each offers a different criteria or algorithm for the intermediate stages in refinement (guide trees and partial alignments) between input and final form.

Because of this, some may work better than others for your data set.

Speed versus accuracy

*Scalable for large datasets such as whole genomes

GDW
Genomics of Disease in Wildlife

Translation Alignments with Codons

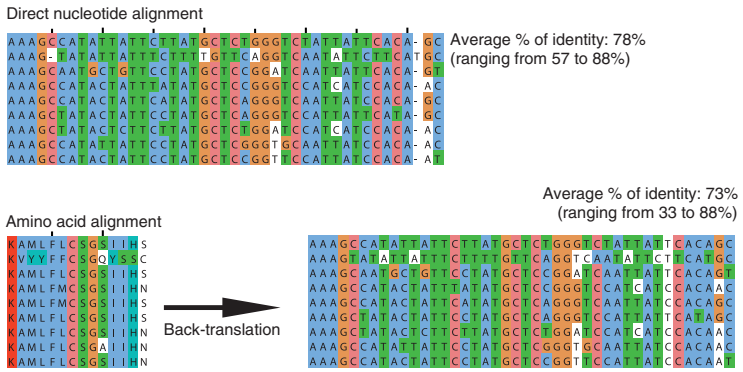


Figure 1. Example illustrating the different performance of the direct and back-translated nucleotide alignments (multiple alignments were built with Muscle with default parameters).

Codon Triplet: 1st, 2nd, 3rd Positions

Non-degenerate (2nd position): all encode nonsynonymous

Two-fold degenerate (1 position): both nonsynonymous and synonymous substitutions

Four-Fold (3 position): All Synonymous

Applications:

Difficult alignment from genomes characterized by:

- Ancient divergence events
- Rapidly mutating genomes

Or

Identification of conserved motifs for function

Common Problems:

Presence of Stop Codons

Presence of Indels

Gene Conversion or Recombination Events

Genomics of Disease in Wildlife

Translation Aligner: MACZE

MACSE

MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons.

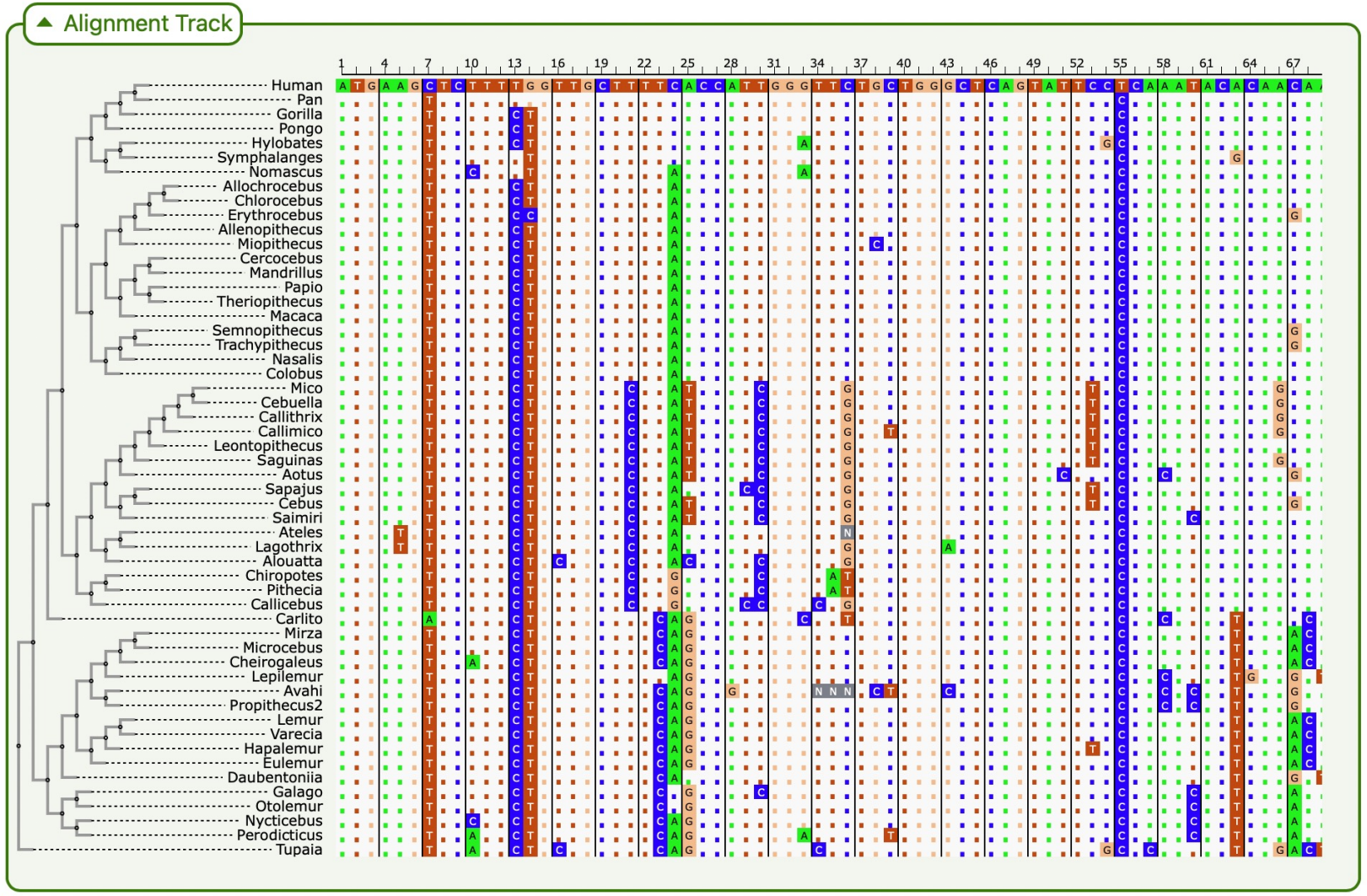
A wide range of molecular analyses relies on multiple sequence alignments (MSA). Until now the most efficient solution to align nucleotide (NT) sequences containing open reading frames was to use indirect procedures that align amino acid (AA) translation before reporting the inferred gap positions at the codon level. There are two important pitfalls with this approach. Firstly, any premature stop codon impedes using such a strategy. Secondly, each sequence is translated with the same reading frame from beginning to end, so that the presence of a single additional nucleotide leads to both aberrant translation and alignment.

MACSE aligns coding NT sequences with respect to their AA translation while allowing NT sequences to contain multiple frameshifts and/or stop codons. MACSE is hence the first automatic solution to align protein-coding gene datasets containing non-functional sequences (pseudogenes) without disrupting the underlying codon structure. It has also proved useful in detecting undocumented frameshifts in public database sequences and in aligning next-generation sequencing reads/contigs against a reference coding sequence.

For further details about the underlying algorithm see the original publication:
MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons.
Vincent Ranwez, Sébastien Harispe, Frédéric Delsuc, Emmanuel JP Douzery
PLoS One 2011, 6(9): e22594.

Genomics of Disease in Wildlife

Amylase Gene Codon Structure



Genomics of Disease in Wildlife

Alignment of Amino Acids

Underlying Assumption:

Construct alignment that minimize cost to preserve physical and chemical properties of protein sequence

Types of Probability Matrices:

Empirical Data Models:

PAM-DayHoff, JTT, WAG , Blosum

Organelle specific matrices:

MtREV

cpREV10, cpREV64

Pathogens:

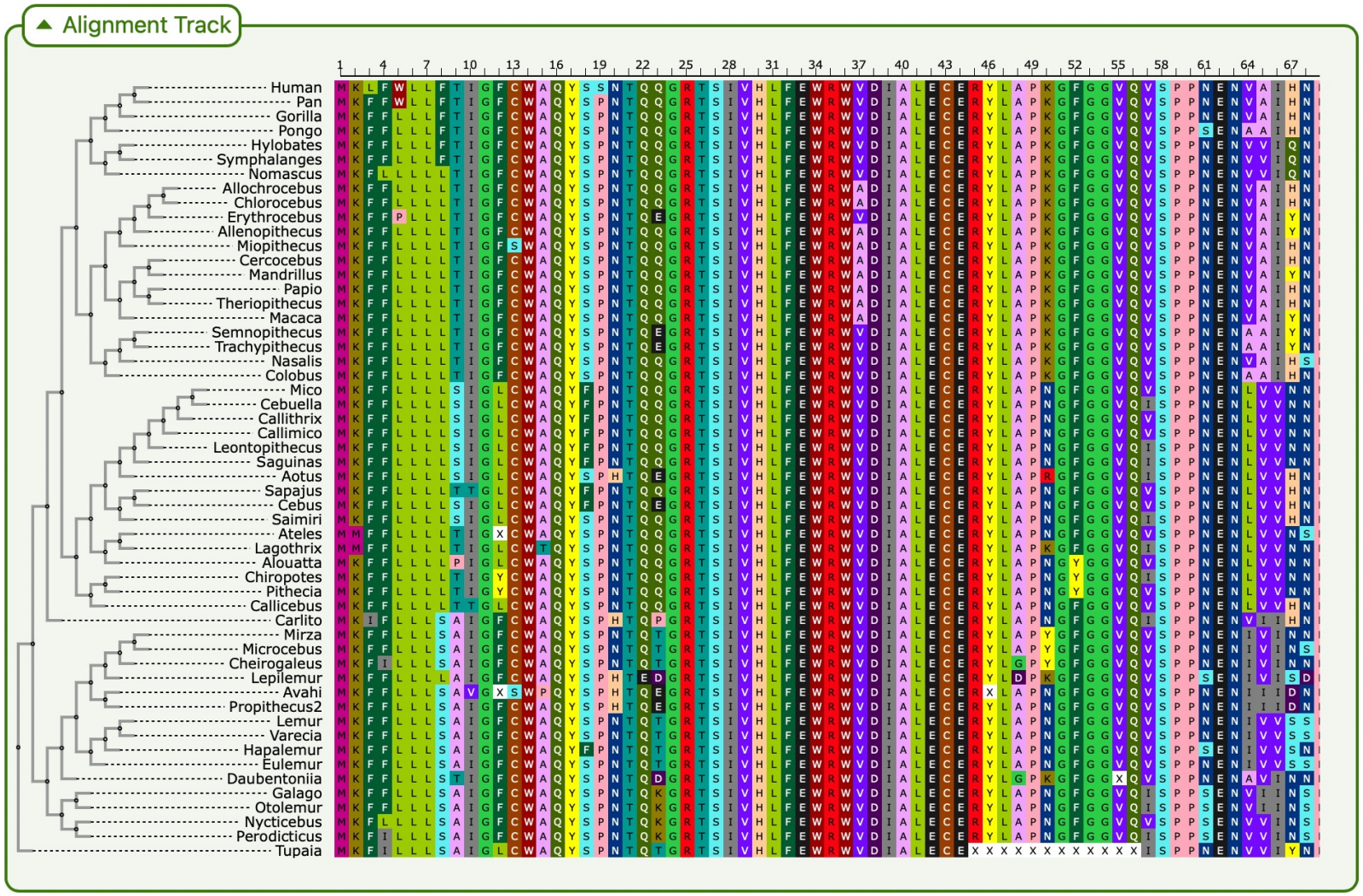
FLU Influenza

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp	D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln	Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met	M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	33	6
Trp	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val	V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

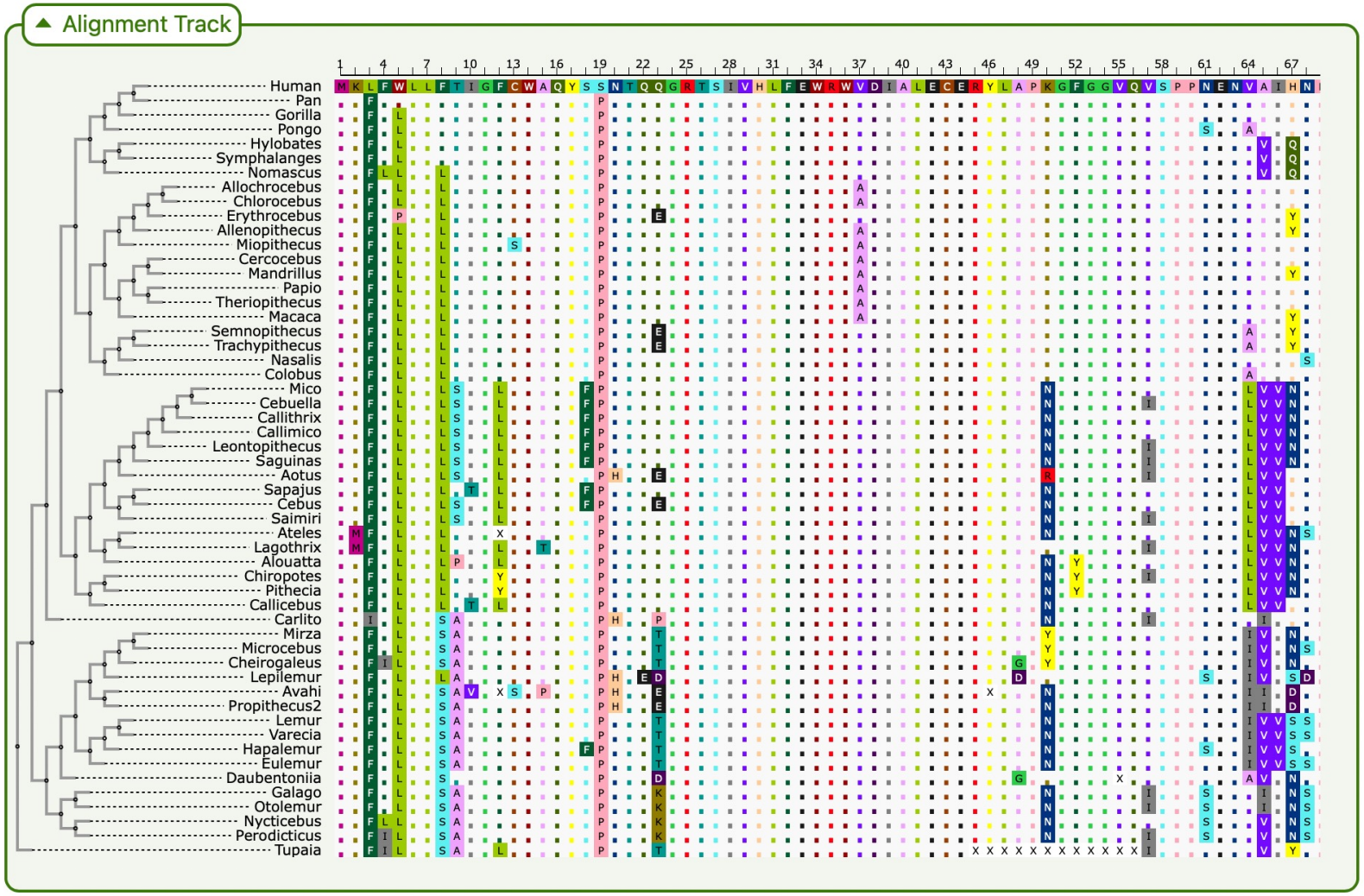
A PAM250 matrix. Each column has been adjusted so that the columns sum to 100.

GDW
Genomics of Disease in Wildlife

Amylase Gene: Aligned in Primates (AA)



Amylase Gene: Aligned in Primates (AA Matched)



Genomics of Disease in Wildlife

How Do We Know It Worked?

?

Why Would it Fail?