

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.

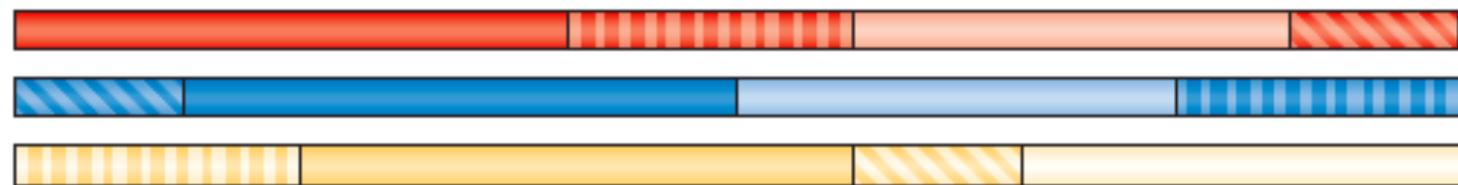
Importance:

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



Alignment Criteria: Optimal collinearity of homologous regions

a Sequenced genomes

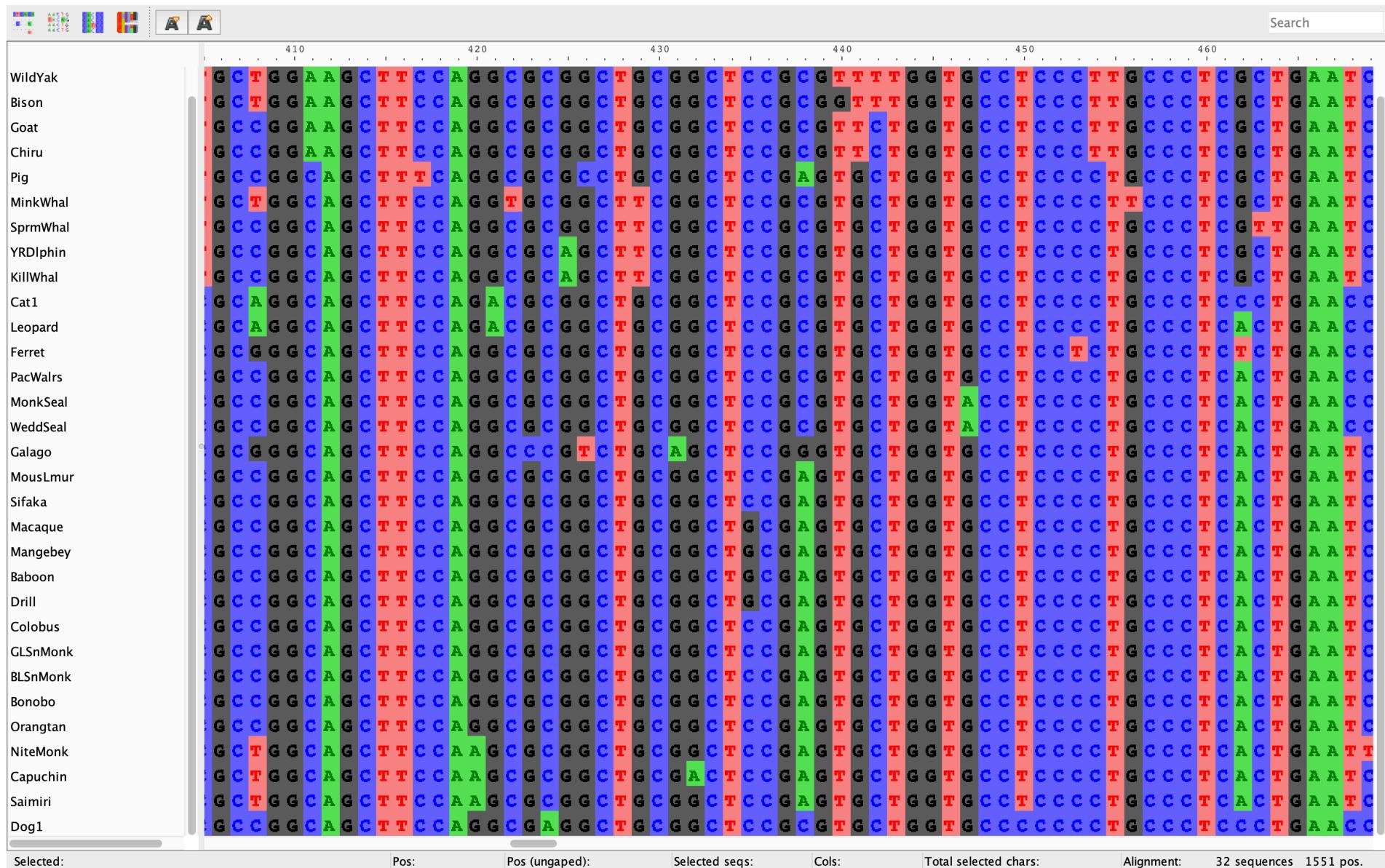


b Reconstruction of homologous collinearity relationships



MARGULIES AND BIRNEY NAT REV 2008



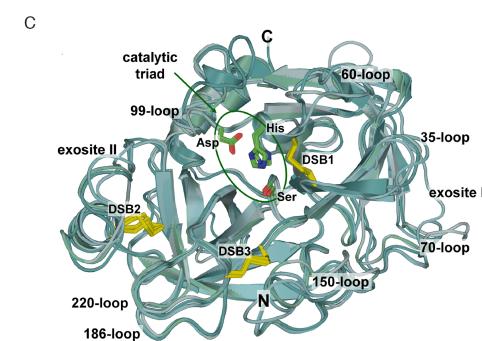
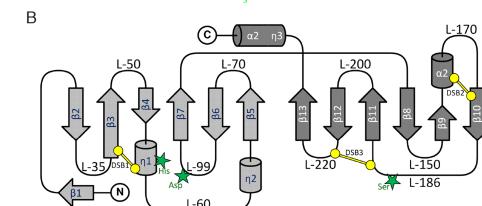
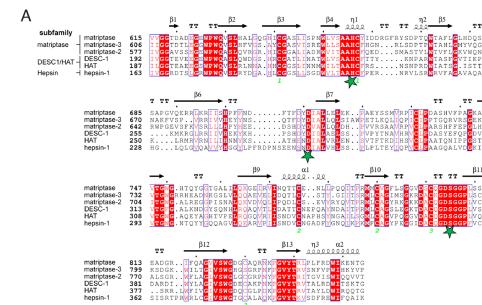


Crucial Input for Comparative Genomic Applications

Comparing within Genomes

Gene families

Structure, function, evolution



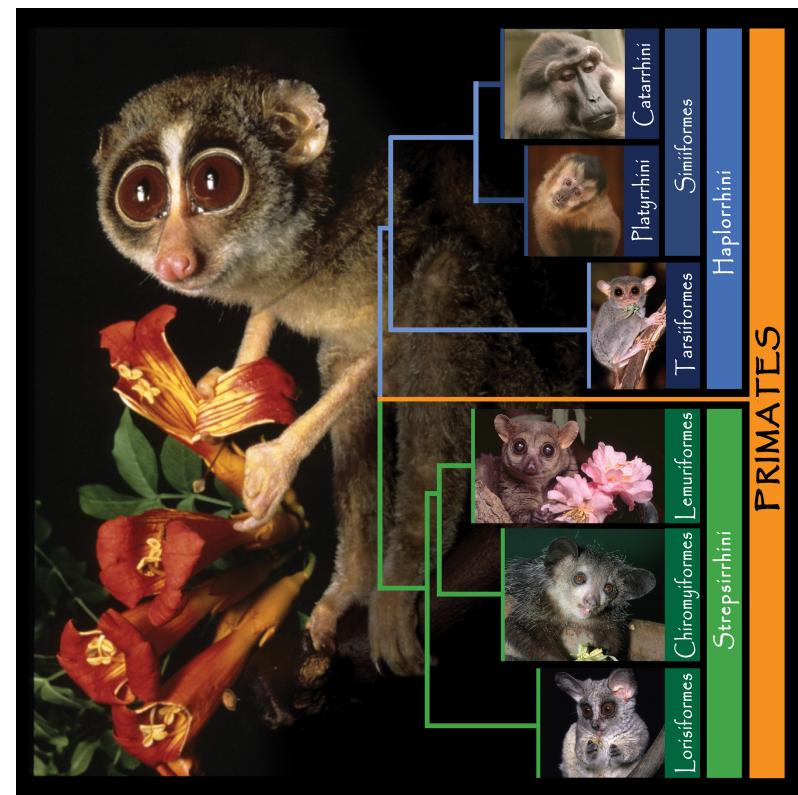
Crucial Input for Comparative Genomic Applications

Host:

Individuals
Populations
Species

Pathogen:

Within host
Between host



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)



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



Translation Alignments with Codons

- Codon Triplet: 1st, 2nd, 3rd Positions
Different Probabilities for Synonymous (dS) and Nonsynonymous (dN) Substitutions

Non-degenerate (2nd position): all encode nonsynonymous (amino acid altering) substitutions
Two-fold degenerate (1st position): both nonsynonymous and synonymous substitutions
Four-fold (3rd position): all synonymous

- Translation into amino acid residues to aid alignment

Applications:

Difficult alignment from genomes characterized by:

Ancient divergence events

Rapidly mutating genomes

Or

Identification of conserved motifs for coding regions

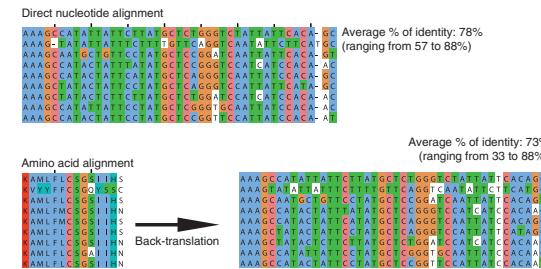


Figure 1. Example illustrating the different performance of the direct and back-translated nucleotide alignments (multiple alignments were built with Muscle with default parameters).

Evolution of Amino Acids

Underlying assumption:

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

Types of probability matrices:

1) Empirical Data Models

PAM-DayHoff, JTT, WAG

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	
Ala	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
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	1	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	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	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.

2) Organelle specific matrices:

MtREV

cpREV10, cpREV64

3) Pathogens

FLU Influenza

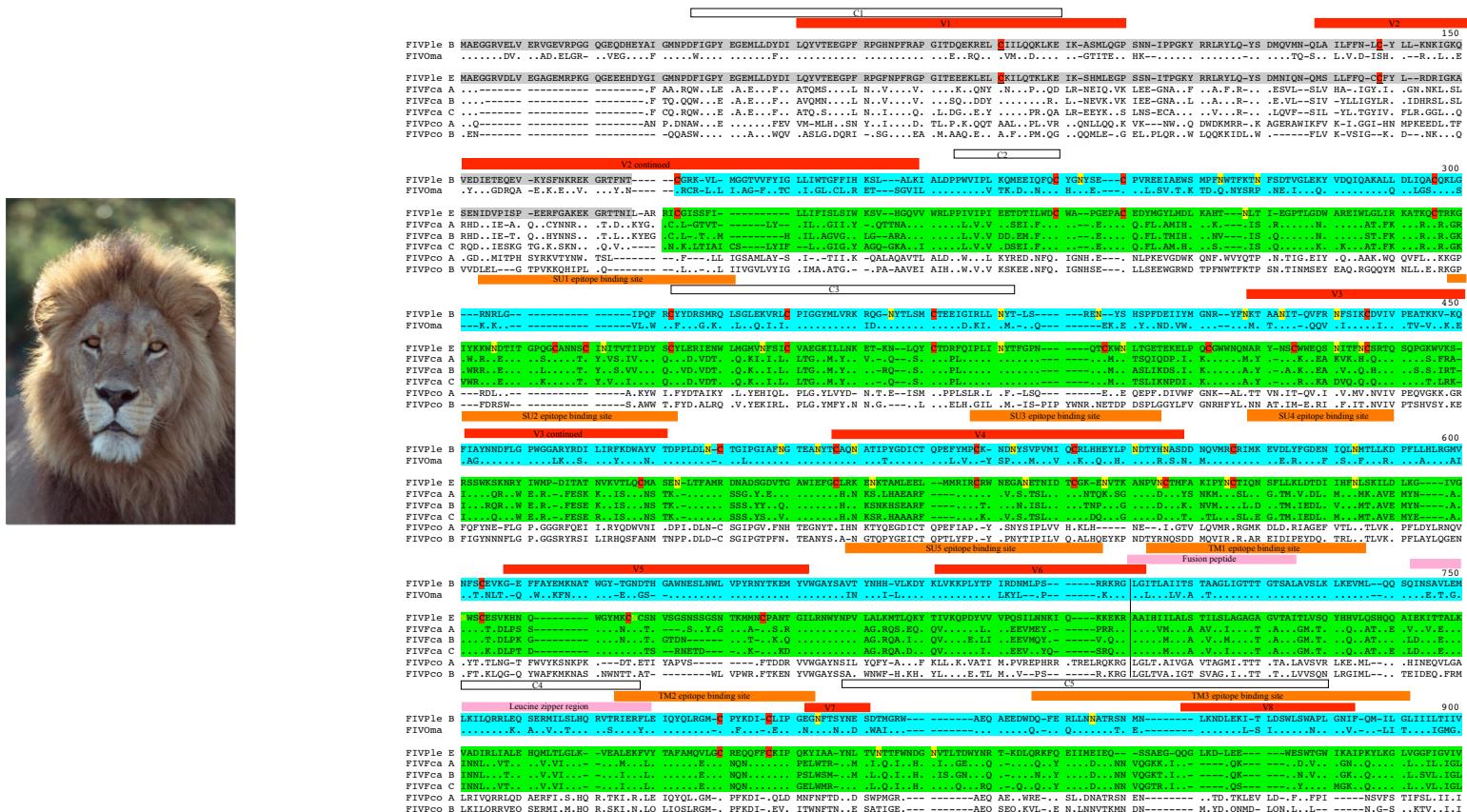
How do we know it worked?

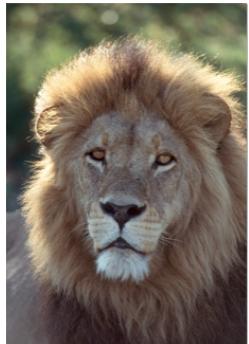
?

Why would it fail?



Recombination in Alignments





Subtype B Serengeti Population: Wild type



Subtype E Okavango Delta Population: Recombinant



*leader and
ORF rev exon 1*

*env surface and
transmembrane*

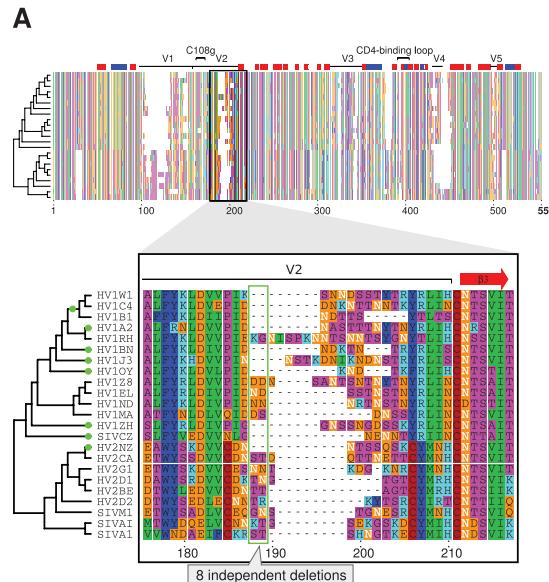
putative RRE



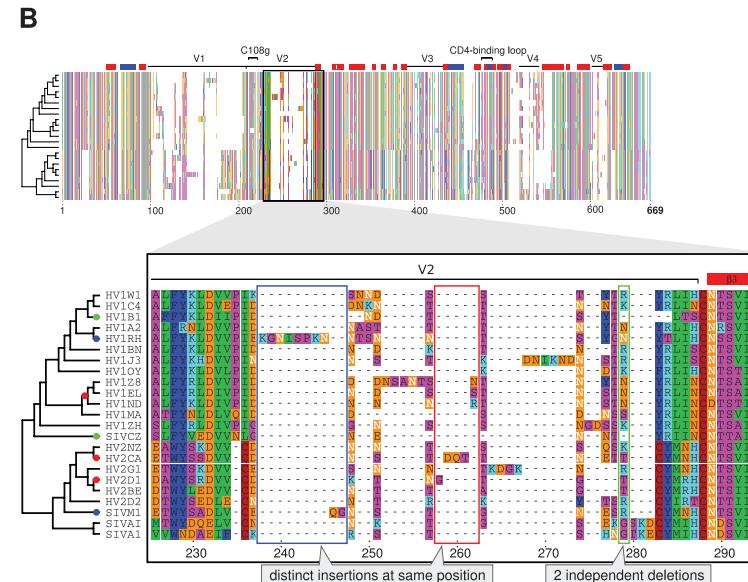
Insertion/Deletions-Can we align and interpret correctly?

SIV/HIV alignment gp120: Hypervariable regions with excessive rates of insertion/deletion

Clustal Omega



PRANK+F



Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis Ari Löytynoja, et al. Science 320, 1632 (2008)

Larger-Sized Indels: (>150-200 bp) Can Confound Alignment Process

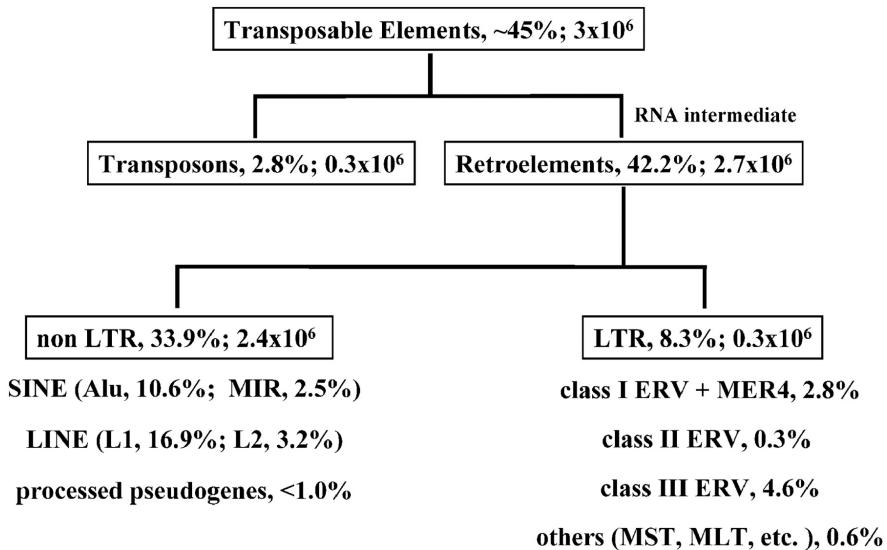
RepeatMasker database if needed
Currently over 56% of human genomic sequence is identified and masked by the program.
Based on curated libraries of repeats RepBase, dfam.

<https://www.girinst.org>

This may not work for new genome sequences from novel organisms

May wish to understand the mechanisms or functional outcome of TE insertion events

Transposable Elements Human Genome



How it should look if alignment method recognized large indels

Walters-Conte et al. BMC Evolutionary Biology 2014, 14:137
<http://www.biomedcentral.com/1471-2148/14/137>

Page 8 of 15

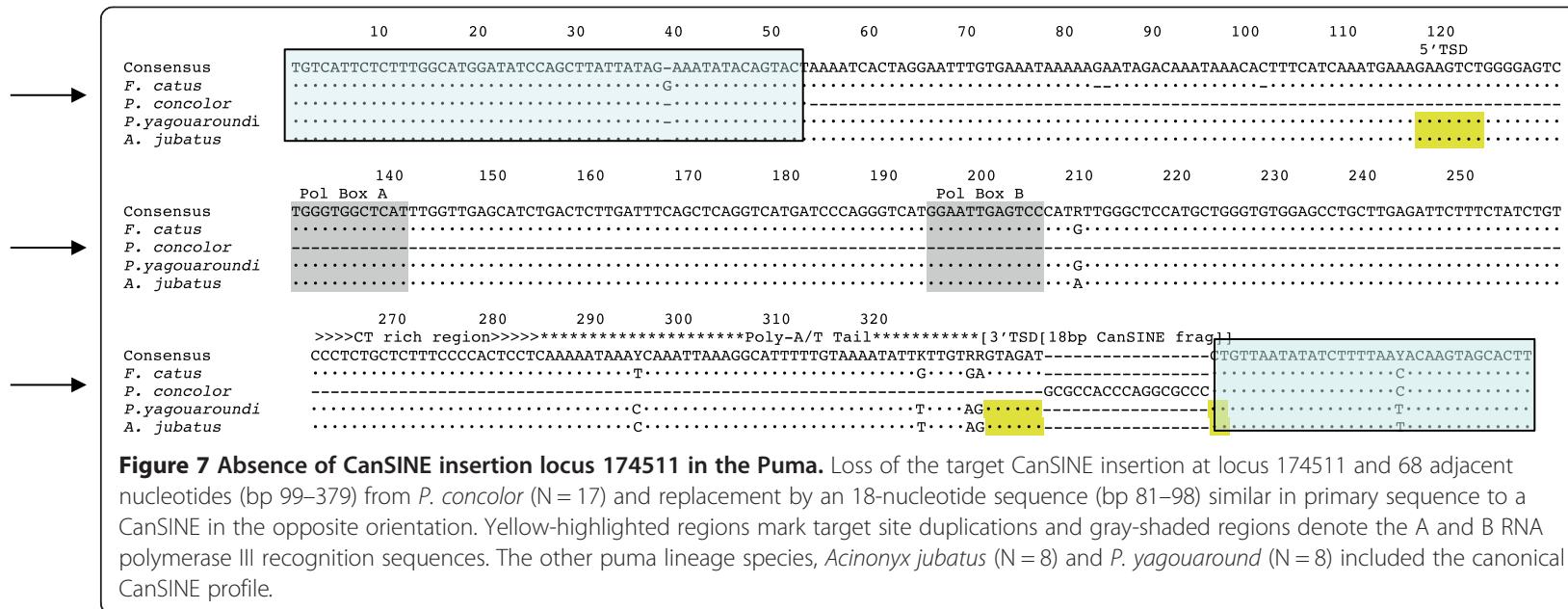
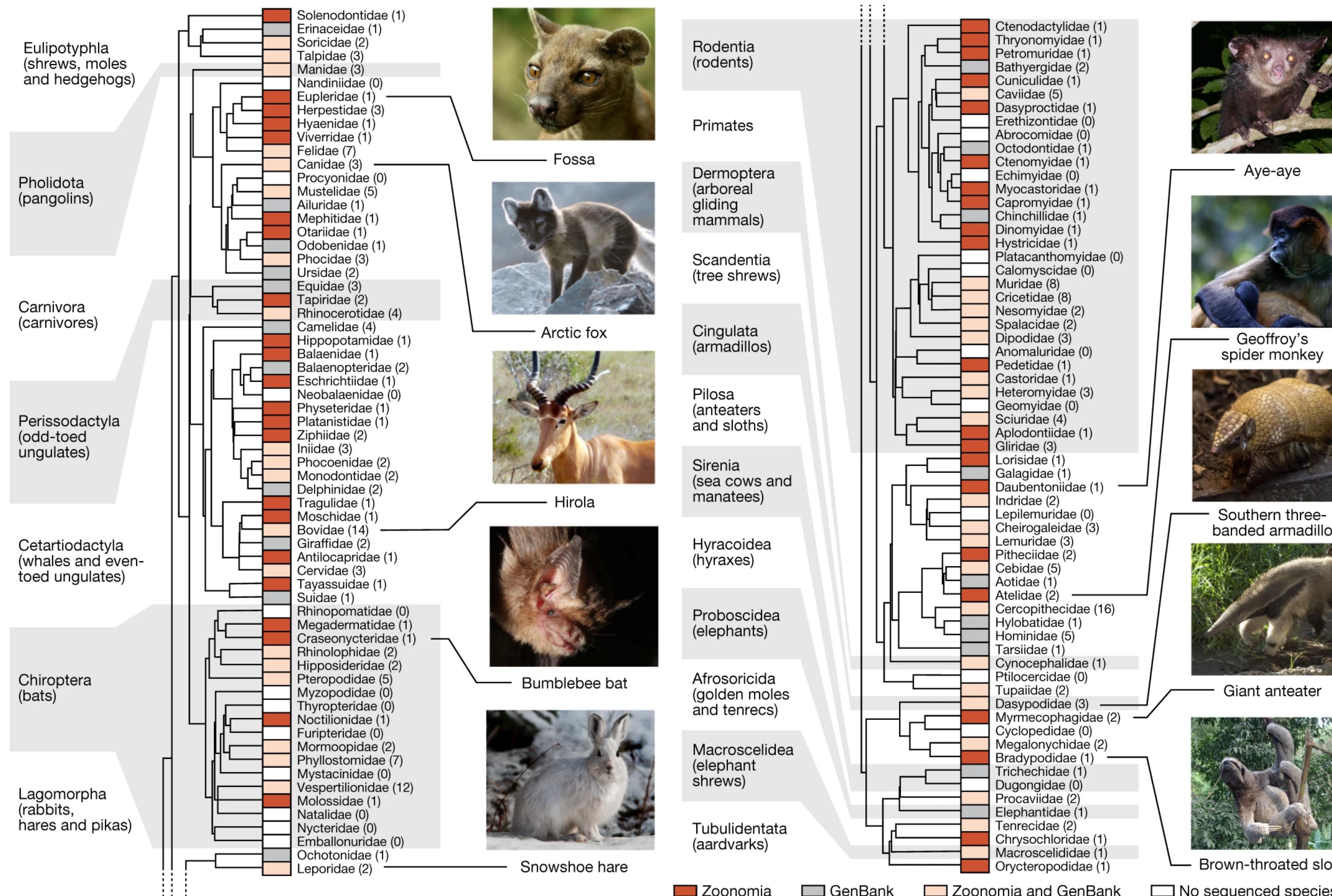


Fig. 1: The Zoonomia Project brings the fraction of eutherian families that are represented by at least one assembly to 83%.

From: A comparative genomics multitool for scientific discovery and conservation



zoonomia

[THE PROJECT](#)[MAMMALIAN TREE](#)[THE DATA](#)[WHITE PAPER](#)[CASE STUDIES](#)[OUR TEAM](#)

The Mammal Tree

[TREE VIEW](#)[LIST VIEW](#)

Conservation Status

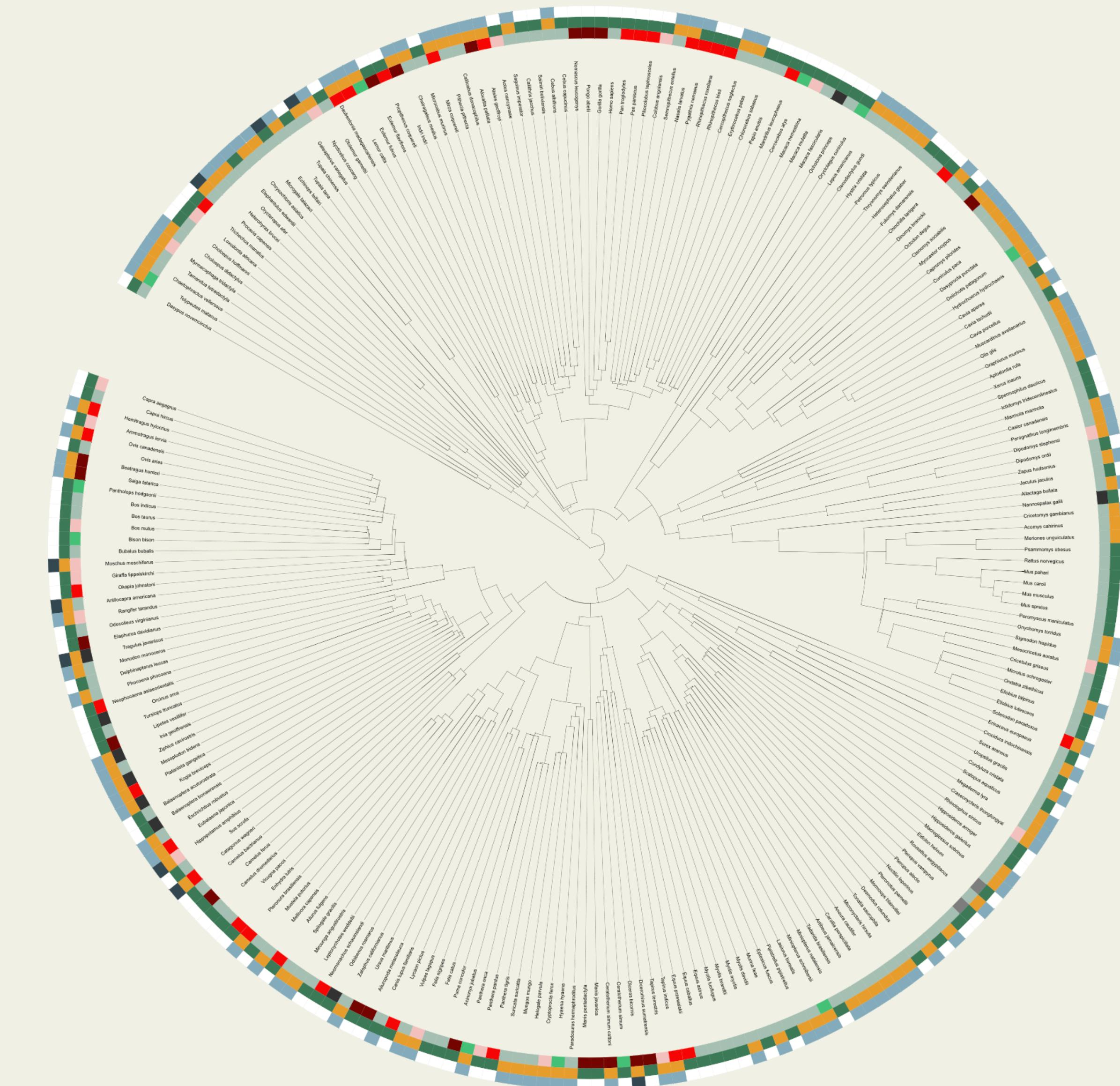
- Data-deficient
- Least Concern
- Near Threatened
- Vulnerable
- Endangered
- Critically Endangered

Assembly Source

- Zoonomia
- Existing Assembly

Assembly Method

- DISCOVAR
- Upgraded DISCOVAR
- Other



Article | [Open Access](#) | Published: 11 November 2020

Progressive Cactus is a multiple-genome aligner for the thousand-genome era

Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefina Stiller, Diane Genereux, Jeremy Johnson, Voichita Dana Marinescu, Jessica Alföldi, Robert S. Harris, Kerstin Lindblad-Toh, David Haussler, Elinor Karlsson, Erich D. Jarvis, Guojie Zhang  & Benedict Paten 

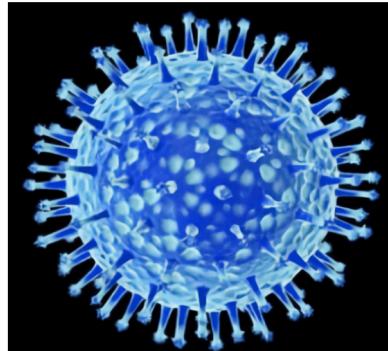
Nature **587**, 246–251 (2020) | [Cite this article](#)

18k Accesses | **11** Citations | **304** Altmetric | [Metrics](#)

Tutorial Exercise: The Nectin4 Gene in Mammals

Morbilliviruses and host receptors: A case study with Nectin4 (poliovirus receptor-related protein 4 PVRL4) and Canine Distemper Virus

Canine Distemper Virus



Deadly Outbreaks In Lions



A lion with myoclonus, involuntary muscle spasms, possibly associated with previous infection with canine distemper virus during a 1994 outbreak.
Image Credit: Serengeti Carnivore Disease Project

Our Question: Is Host receptor Nectin4 correlated with neurologic form of CDV infection in felids and canids?

Background:

There are seven known Morbilliviruses that infect Carnivores, Cetaceans, Phocids, Primates, ungulates, small ruminants:

- Feline morbilliviruses
- Canine distemper virus
- Phocine morbillivirus
- Cetacean morbillivirus
- Rinderpest virus
- Small ruminant morbillivirus
- Measles morbillivirus

CDV is known to access cell entry into the host through host genes SLAM and NECTIN4 (PVRL4). It is hypothesized that NECTIN4 is linked with neurological forms of canine distemper. We will conduct a comparative genomic analysis of NECTIN4 in mammals to determine structure, function and evolution. Our study will focus only on the coding regions (CDS) of the gene.

Pilot Study Using Genome Mining:

We have searched RefSeq and NCBI for CDS (exons) of Nectin4 from affected mammalian taxa and downloaded representative sequences that are full-length CDS of NECTIN4.

