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

Signals of Selection in Host Genome (Nectin4 as Proxy)
GDW2019 Group Exercise with Nectin 4: Hypothesis, Genomic Workflow, Findings, and Future Directions

GDW
Genomics of Disease in Wildlife

Nectin4 Workflow –Step 1

Hypothesis

Research shows CDV co-receptor Nectin4 is linked with the neurological form of CDV in carnivores, particularly felids and canids.

- HO: Nectin4 exhibits neutral evolution among mammalian lineages, tracking speciation.
- H1: Nectin4 has variants linked with neurologic CDV susceptibility

GDW
Genomics of Disease in Wildlife

Nectin4: Step 2

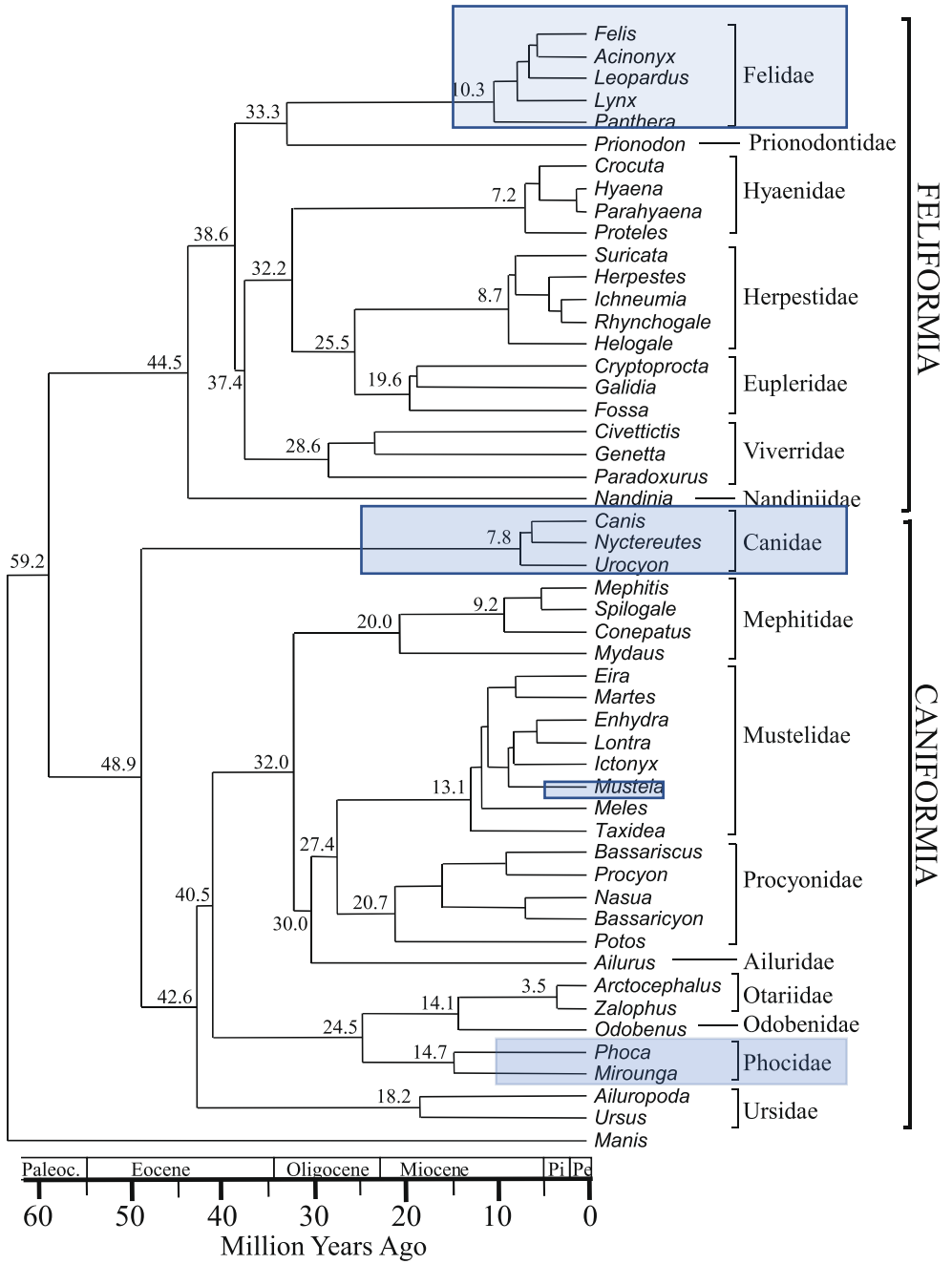
Assembling Data for a Pilot Study

1) Understand the Host: Mammalian Evolution and Phylogeny

2)Genome mining of existing sequences from representative lineages (NCBI, ENSEMBL) refGene

3)Download and translate into coding sequences (e.g. Geneious)

GDW
Genomics of Disease in Wildlife



Carnivore Order

Genomics of Disease in Wildlife

Nectin4: Step 3

Alignment of Nectin4 Sequences from 32 Taxa

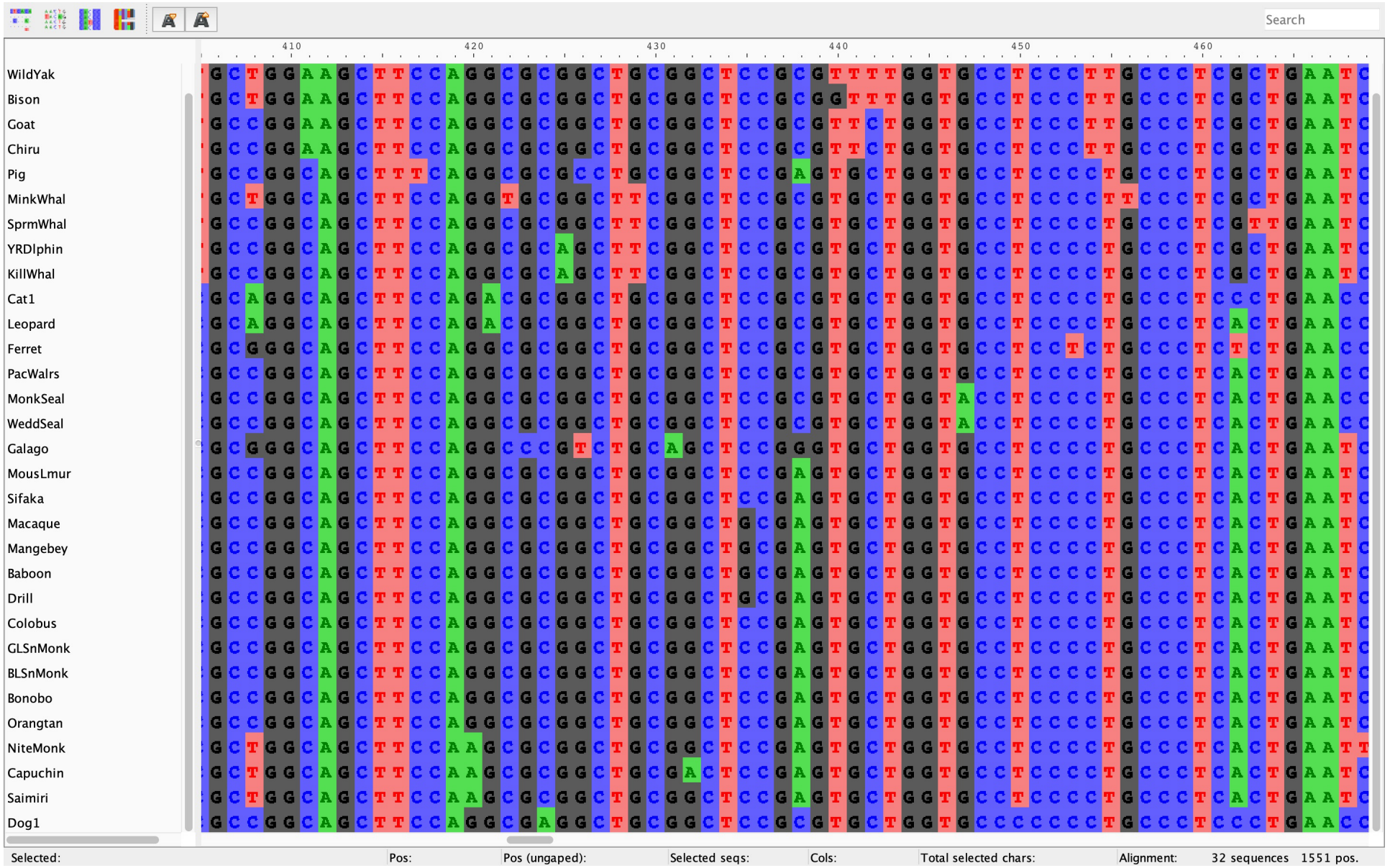
Programs: MUSCLE, Clustal Omega, Mafft

Which works best with coding regions?

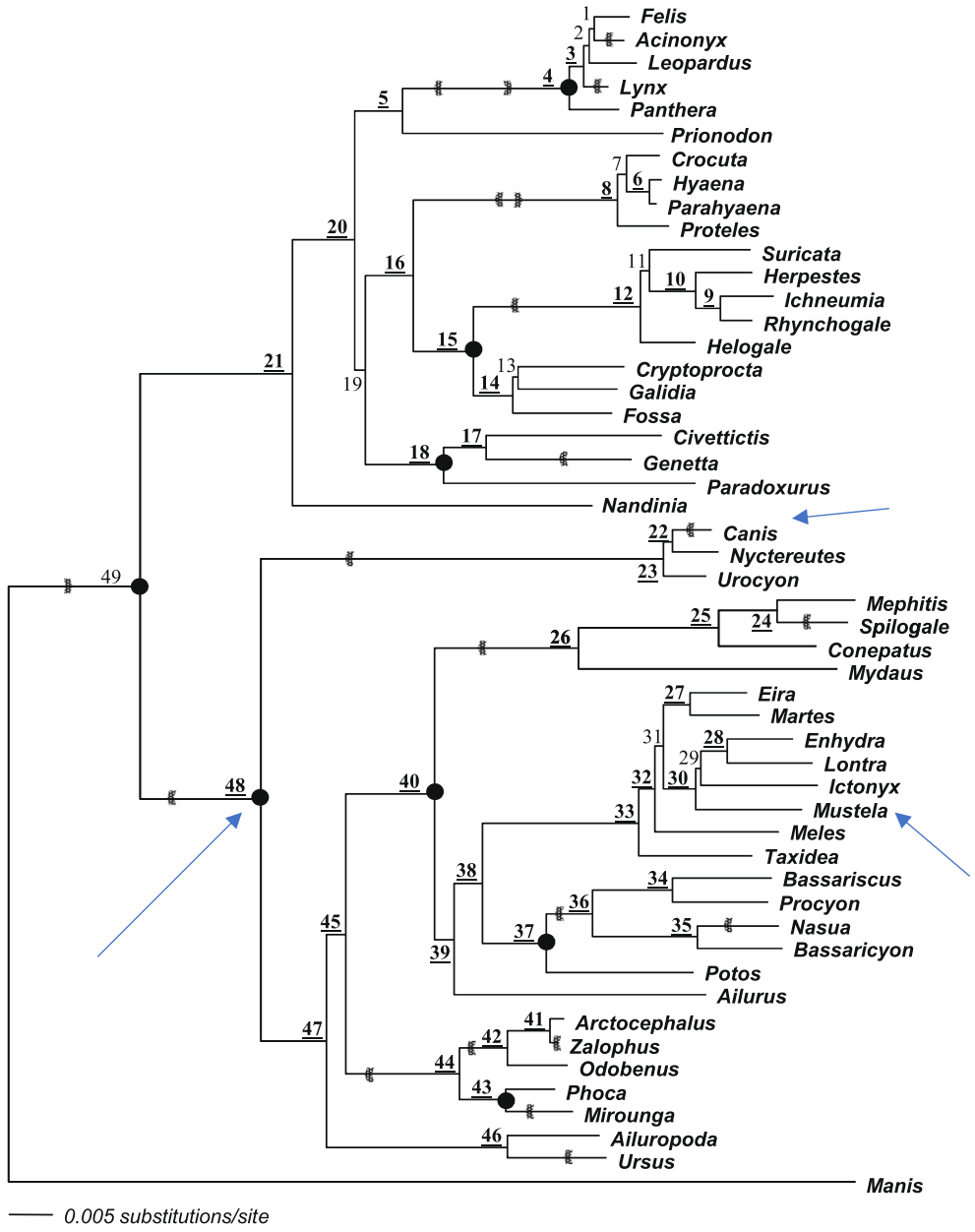
What is the key feature of multiple sequence files (MSA) that alignment programs must resolve ?

Indels

Genomics of Disease in Wildlife



Genomics of Disease in Wildlife



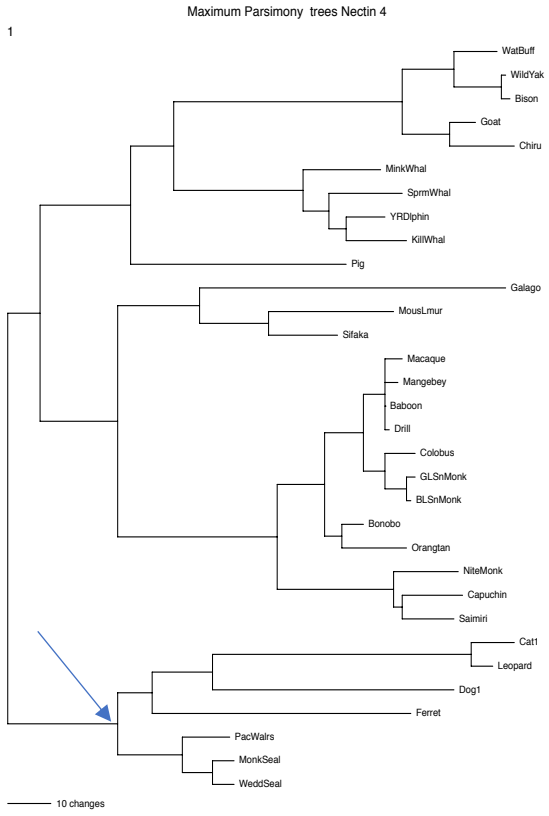
Feliformia

Caniformia

Genomics of Disease in Wildlife

Nectin4: Step 4

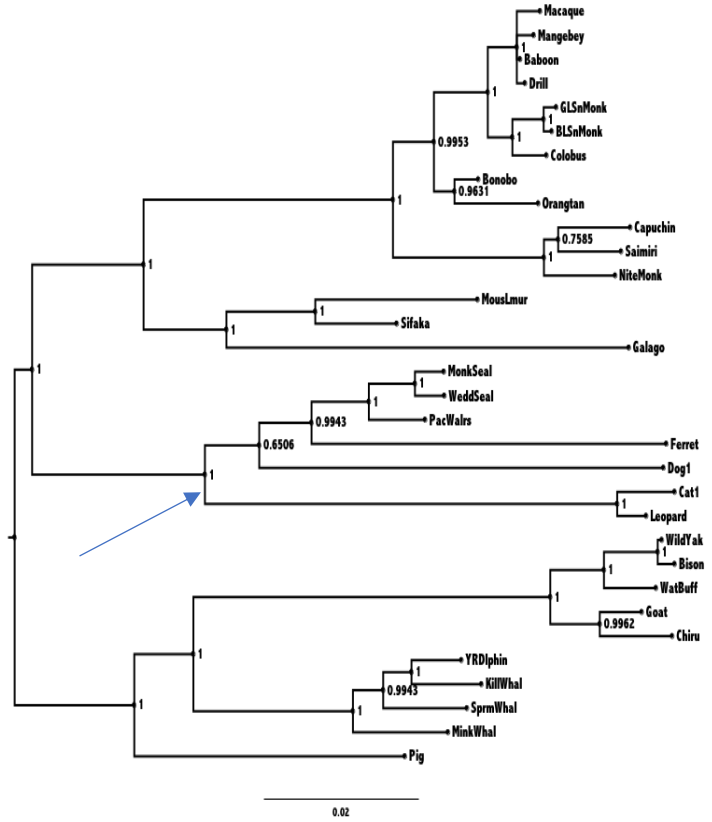
MP



ME

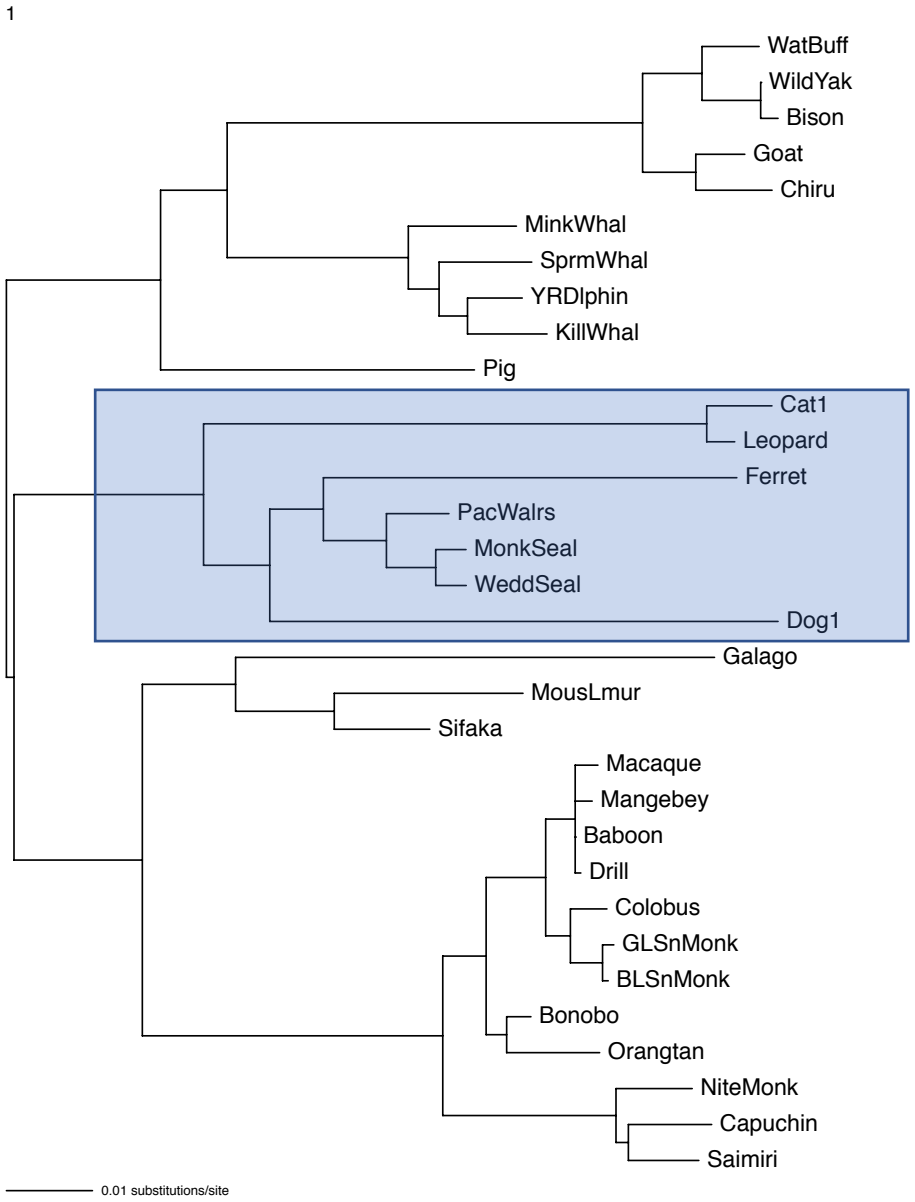
ML

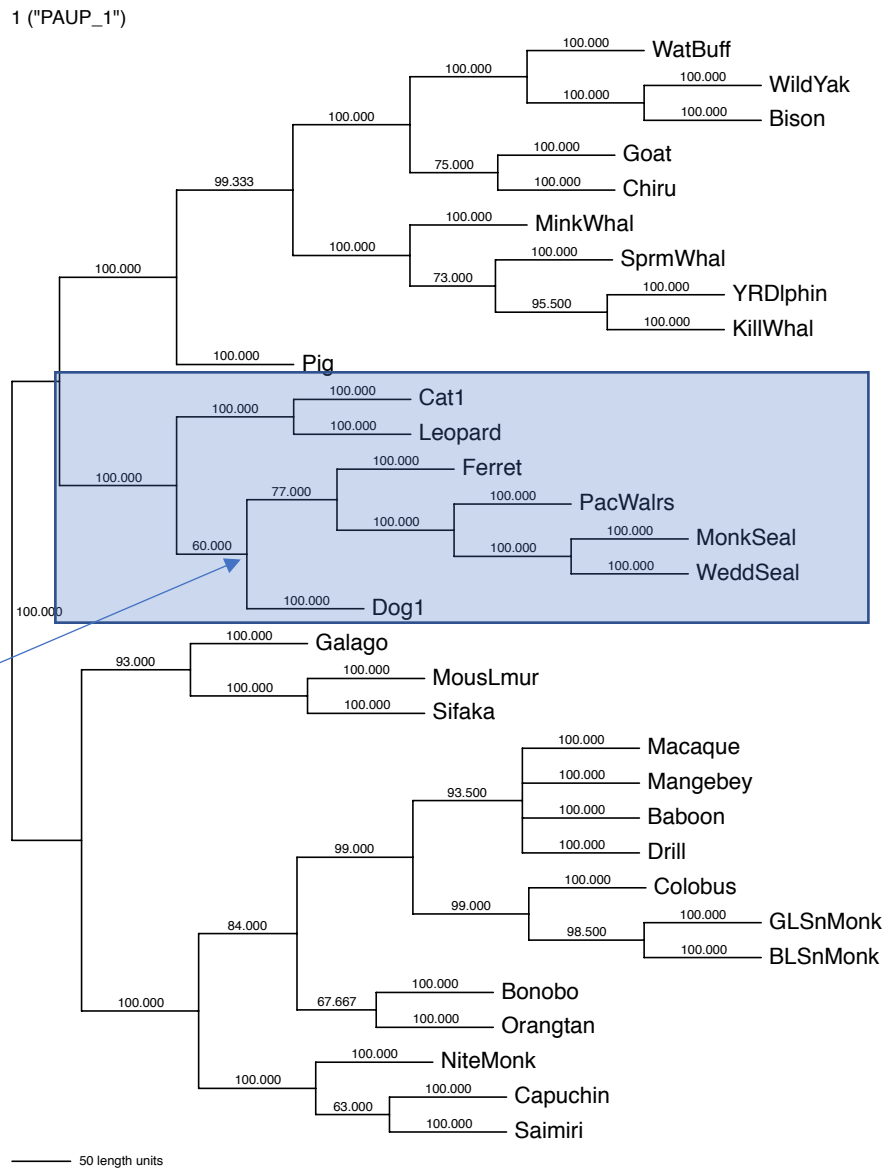
Bayes



Genomics of Disease in Wildlife

ML Phylogeny
100 BS reps





Genomics of Disease in Wildlife

Testing Coding Regions for Signals of Adaptive Evolution

- Is there evidence of selection operating on a gene?
- Where does selection occur within the gene? What regions, motifs, amino acids are under selection?
- Mapping the selection event on the phylogenetic tree. Is there a specific species or lineage that is experiencing selection?
- Assessing the form of selection (i.e. negative or positive) and identifying the codon (s), and *a priori* testing for statistical rigor.
- Are other genes exhibiting compensatory changes coincident with selection?

GDW
Genomics of Disease in Wildlife

Testing for Selection in Aligned Sequences (MSA)

- Aligned codon sequences must be in frame with no stop codons.

- Remove recombination motifs and/or analyze partitions of MSA that are confirmed identical by descent.
  - Permits unbiased estimates of parameters

- A resolved phylogenetic tree of the multiple sequence file.
  - Pre-existing species tree established from other analyses
  - A poorly resolved tree will be unable to adequately test for selection and result in spurious results.

GDW
Genomics of Disease in Wildlife

Molecular Selection In Coding Sequences

dN – nonsynonymous (missense) substitution

dS – synonymous substitutions

$\omega = dN/dS$

$\omega = 1$ (Neutral), no functional effect

$\omega < 1$ (Purifying selection), selected to maintain function

$\omega > 1$ (Diversifying selection), selected for adaptation to change in function

GDW
Genomics of Disease in Wildlife

Categories of Codon Selection Models

- Among sites:
  - Ho: Variable selection pressure possible among sites within YGOI (your-gene-of-interest) but no sites exhibit positive selection
- Among branches:
  - Ho: Average dN/dS is the same among all branches in the phylogeny for YGOI.
- Among clades:
  - Ho: Average dN/dS for YGOI is the same for each lineage within the phylogeny
- Branch-site:
  - Ho: Variable selection pressure is possible among sites with YGOI and no sites exhibit positive selection in any particular lineage relative to the rest of the phylogeny.

GDW
Genomics of Disease in Wildlife

HyPHy: HYpothesis testing using PHYlogenies

Step 1: GARD (**G**enetic **A**lgorithm for **R**ecombination **D**etection)
Pre-Analyses for Selection
Identify Recombinant Regions of Alignment

Step 2: Among Sites Model across Phylogeny (pervasive selection)- Programs listed progressive larger datasets
FEL (**F**ixed **E**ffects **L**ikelihood) ,
SLAC (**S**ingle-**L**ikelihood **A**ncestor **C**ounting) ,
FUBAR (**F**ast, **U**nconstrained **B**ayesian **A**pp**R**oximation)

Step 3: Among sites Models within a subset of branches within a phylogeny
MEME (**M**ixed **E**ffects **M**odel of **E**volution)

Step 4: Branch-site Models
aBSREL (**a**daptive **B**ranch-**S**ite **R**andom **E**ffects **L**ikelihood)

GDW
Genomics of Disease in Wildlife

Basic Steps to CodeML

Create 3 files: data, tree and control

Load files into GUI interface

Select Parameters for appropriate test

Depending on numbers of sequences, genetic information, pattern of mutation, length of sequence…..

CodeML can take minutes or hours to run

Record Ln likelihood value

Compare with ln likelihood of null model.

Determine significance by the log-likelihood ratio model (LRT).

$\Delta\lambda = 2\ (l_1 - l_0)$ chi-square 2 d.f.

For Branch-Site Models

$\Delta\lambda = 2\ (l_1 - l_0)$ chi-square 2 d.f. P-value/2

GDW
Genomics of Disease in Wildlife

Site Models

Model	NSsites	np	Free parameters
M0 (one ratio)	NSsites = 0	1	ω
M1a (NearlyNeutral): p_0 ($p_1 = 1 - p_0$) $\omega_0 < 1$, $\omega_1 = 1$	NSsites = 1	2	p_0, $\omega_0 < 1$
M2a (PositiveSelection): p_0, p_1 ($p_2 = 1 - p_0 - p_1$) $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$	NSsites = 2	4	p_0, p_1, $\omega_0 < 1$, $\omega_2 > 1$
M3 (discrete): p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2	NSsites = 3	5	p_0, p_1, ω_0, ω_1, ω_2
M7 (beta): p, q	NSsites = 7	2	p, q
M8 (beta&ω): p_0 ($p_1 = 1 - p_0$) p, q, $\omega_s > 1$	NSsites = 8	4	$p_0, p, q, \omega_s > 1$

LRT

Model M1a and Model M2a, 2 df

Model M7 and Model M8, 2 df

GDW
Genomics of Disease in Wildlife

Branch Site Models

- Branch-site
  - Model A recommended
    - Model = 2
    - NSsites=2
  - Compare LRT with Null Model A
    - Model=2
    - NSsites=2
    - But, ω=1, fixed.

- Results include MLC files and Rst files which will include any BEB analyses that will identify putative sites under selection.

Branch site model A: Old and New

Site class	Proportion	Old model A (np = 3)		New model A (np = 4)	
		Background	Foreground	Background	Foreground
0	p_0	$\omega_0 = 0$	$\omega_0 = 0$	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)\, p_0/(p_0 + p_1)$	$\omega_0 = 0$	$\omega_2 > 1$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)\, p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$	$\omega_1 = 1$	$\omega_2 > 1$

Genomics of Disease in Wildlife

Nectin4: Step 5

Tests for Selection with PAML

We use a **constraint tree** corrected according to known mammalian evolution.

Likelihood values:

Among Sites Models	Branch Site Models
Model 0	Carnivores
Model 1	Dog-like
Model 2	Cat-like
Model 7	
Model 8	

Genomics of Disease in Wildlife

Nectin4: Among Sites Models

Table 1. Tests for selection among codons of Nectin4 in Mammals using models implemented in PAML 4.8 using the likelihood method

Criteria	Model	Parameter estimates	Ln likelihood	LRT 2 df	Selected Codon, BEB Posterior Probability
Among Sites	M0 (one ratio)	k = 3.26438 w_0= 0.12726	-7210.145636	NA	NA
	M1a (nearly neutral)	k = 3.37983 w_0= 0.04628 (p_0=0.88353) w_1= 1.00000 (p_1=0.11647)	-7131.607388	NA	NA
	M2a (selection)	k = 3.37982 w_0=0.04628 (p_0=0.88353) w_1= 1.00000 (p_1=0.116470) w_2= 24.47759 (p_2=0.00000)	-7131.607388	M2a vs M1a NS	409 S 0.646
	M7 (b distribution, neutral)	k =3.29736 b distribution p= 0.13854 q=0.81699	-7126.396981	NA	NA
	M8 (b distribution, selection)	k = 3.29588 w_s= 0.02930 (p_1 = 1.22390) b distribution p_0= 0.97070 p=0.16546 q=1.29682	-7124.109459	M8 vs M7 4.6, P=0.1, 2 d.f. NS	14 A 0.617 17 W 0.845 32 L 0.863 74 A 0.614 78 G 0.907 193 T 0.638 309 P 0.804 314 T 0.769 341 A 0.581 342 P 0.711 409 S 0.967* 481 R 0.688

Genomics of Disease in Wildlife

Nectin4: Branch-Site Models

Table 2. Test of selection of Nectin4 among lineages using revised branch-site Model A and test 2 as implemented in PAML 4.8

Model A	Parameter Estimates					Ln likelihood LRT test 2,1df	Selected Codon, Posterior P Positive sites for foreground lineages Prob (w>1):	
		Site Class	p	w0 background	w1 foreground			
Felid lineage	k=3.37927	0 1 2a 2b	0.87897 0.11127 0.00867 0.00110	0.04491 1.00000 0.04491 1.00000	0.04491 1.00000 5.55434 5.55434	LnL0 = -7129.218766 LnL1 = -7128.155700, P=0.07	25 A 365 C	0.932 0.936
Caniform Lineage	k=3.37985	0 1 2a 2b	0.88353 0.11647 0.00000 0.00000	0.04628 1.00000 0.04628 1.00000	0.04628 1.00000 1.00000 1.00000	LnL0=-7130.813821 LnL1=-7131.607388 P=0.10 NS	Both back ground and foreground share same values.	
Carnivore Lineage	k=3.37305	0 1 2a 2b	0.88088 0.10867 0.00930 0.00115	0.04536 1.00000 0.04536 1.00000	0.04536 1.00000 1.65644 1.65644	Lnl0= -7130.813821 Lnl1= -7130.579671 NS	25 A 33 A 86 S 87 K 277 Q 365 C	0.675 0.679 0.817 0.709 0.586 0.689

GDW
Genomics of Disease in Wildlife

Future Directions?