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

Nectin4: Step 2

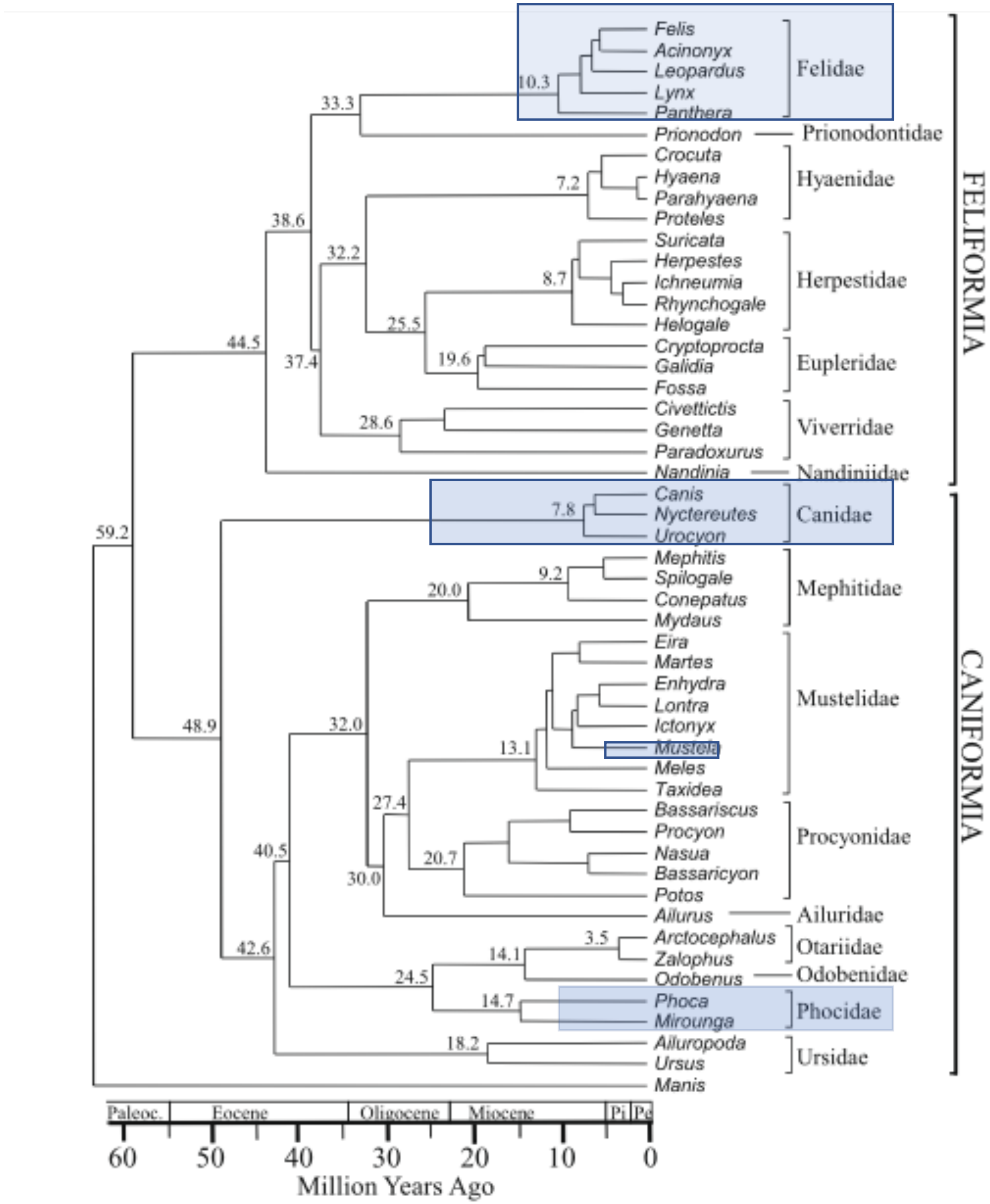
Assembling Data for a Pilot Study

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)

E. Eizirik et al. / Molecular Phylogenetics and Evolution 56 (2010) 49–63



Carnivore Order

Genomics of Disease in Wildlife

Nectin4: Step 3

Alignment of Nectin4 Sequences from 32 Taxa

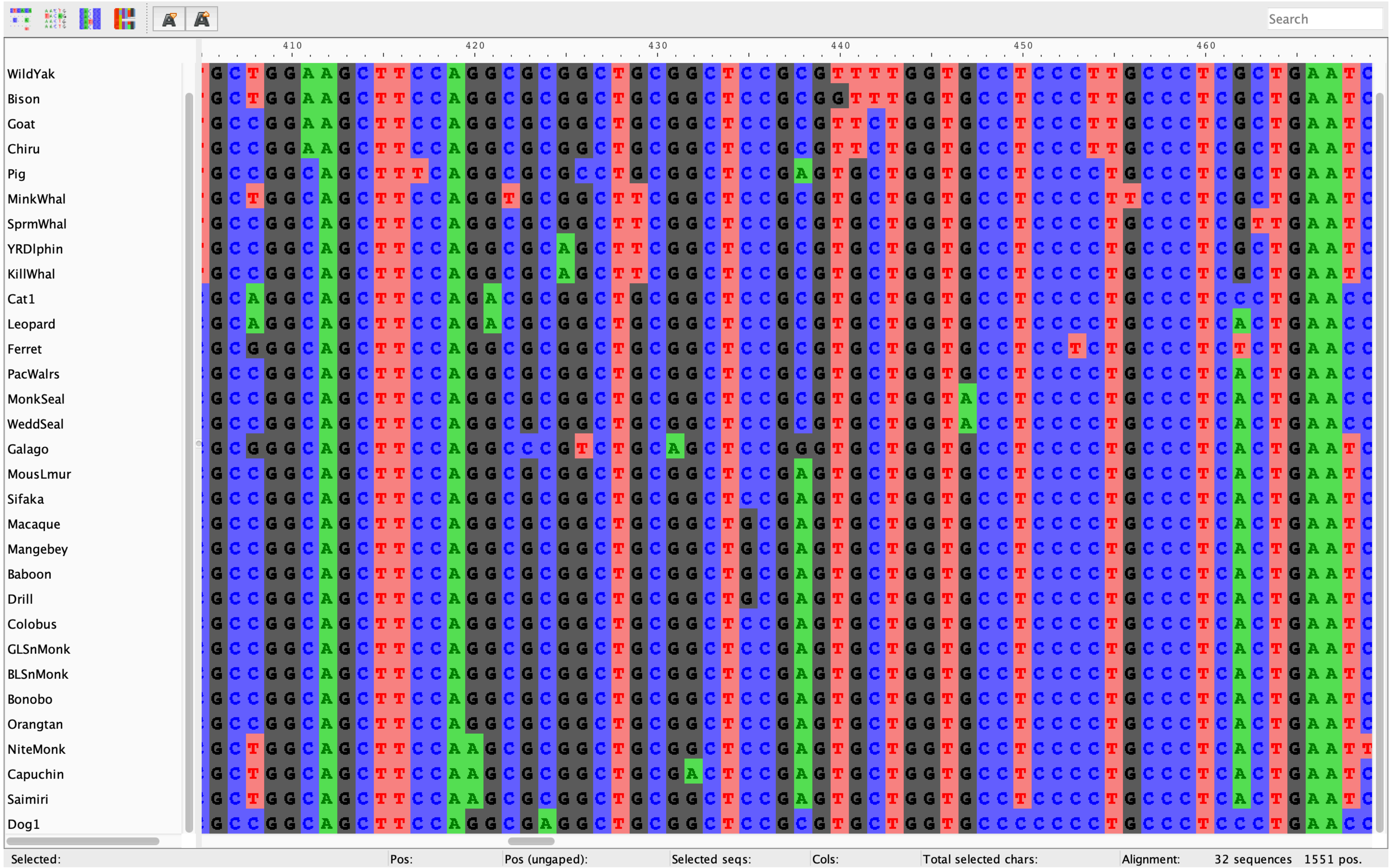
Programs: Aliview (Muscle), ClustalX, Clustal Omega, Mafft, PRANK

Which works best with coding regions?

MUSCLE, PRANK or adapting programs to identify codon (aligned as translated codons)

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

Indels

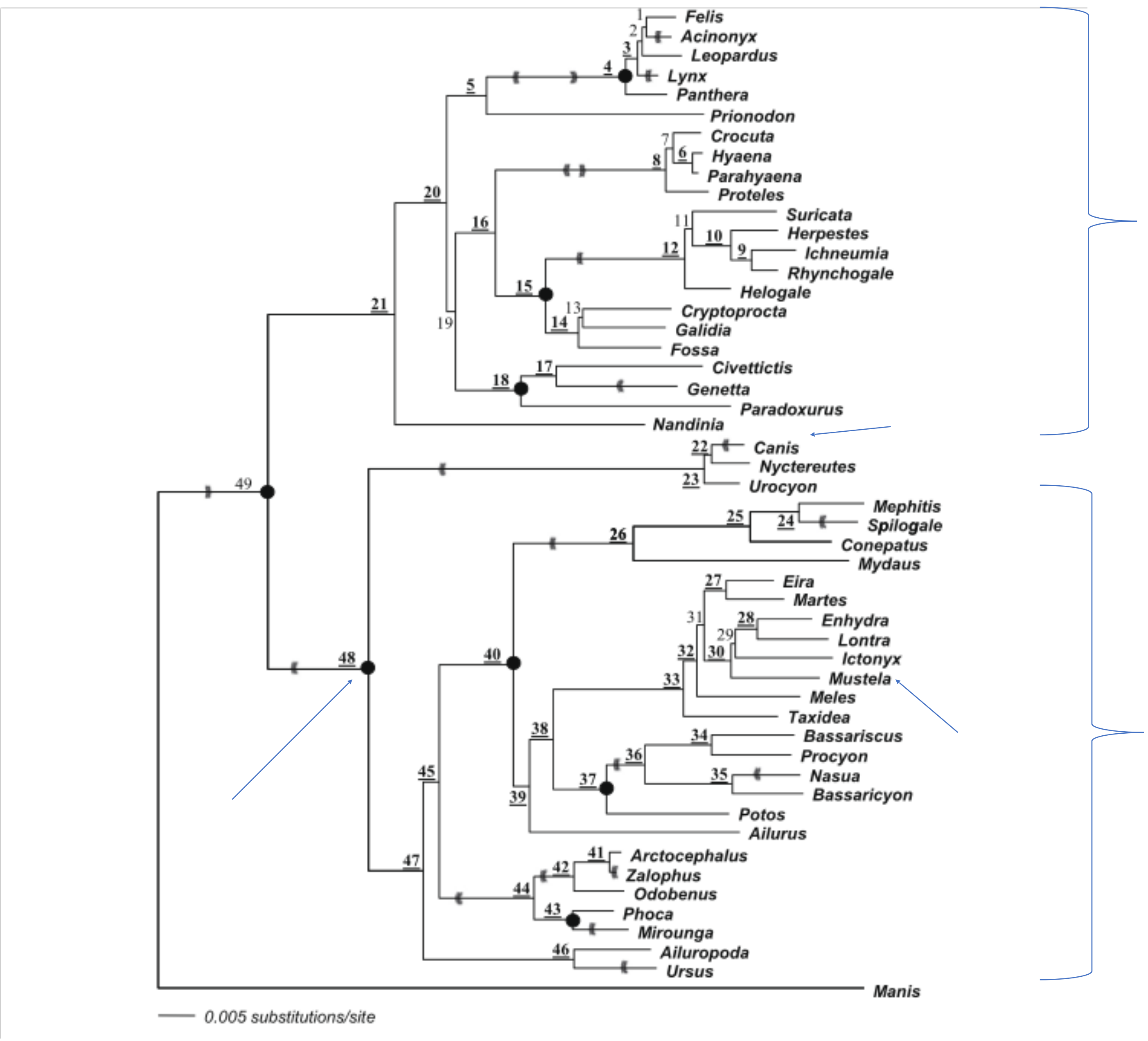


```
                     410       420       430       440       450       460
WildYak    GCTGGAAGCTTCCAGGCGCGGCTGCGGCTCCGCGTTTTGGTGCCTCCCTTGCCCTCGCTGAATC
Bison      GCTGGAAGCTTCCAGGCGCGGCTGCGGCTCCGCGGTTTGGTGCCTCCCTTGCCCTCGCTGAATC
Goat       GCCGGAAGCTTCCAGGCGCGGCTGCGGCTCCGCGTTCTGGTGCCTCCCTTGCCCTCGCTGAATC
Chiru      GCCGGAAGCTTCCAGGCGCGGCTGCGGCTCCGCGTTCTGGTGCCTCCCTTGCCCTCGCTGAATC
Pig        GCCGGCAGCTTTCAGGCGCGCCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCGCTGAATC
MinkWhal   GCTGGCAGCTTCCAGGTGCGGCTTCGGCTCCGCGTGCTGGTGCCTCCCCTTCCCTCGCTGAATC
SprmWhal   GCCGGCAGCTTCCAGGCGCGGCTTCGGCTCCGCGTGCTGGTGCCTCCCCTGCCCTCGTTGAATC
YRDlphin   GCCGGCAGCTTCCAGGCGCAGCTTCGGCTCCGCGTGCTGGTGCCTCCCCTGCCCTCGCTGAATC
KillWhal   GCCGGCAGCTTCCAGGCGCAGCTTCGGCTCCGCGTGCTGGTGCCTCCCCTGCCCTCGCTGAATC
Cat1       GCAGGCAGCTTCCAGACGCGGCTGCGGCTCCGCGTGCTGGTGCCTCCCCTGCCCTCCCTGAACC
Leopard    GCAGGCAGCTTCCAGACGCGGCTGCGGCTCCGCGTGCTGGTGCCTCCCCTGCCCTCACTGAACC
Ferret     GCGGGCAGCTTCCAGGCGCGGCTGCGGCTCCGCGTGCTGGTGCCTCCTCTGCCCTCTCTGAACC
PacWalrs   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGCGTGCTGGTGCCTCCCCTGCCCTCACTGAACC
MonkSeal   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGCGTGCTGGTACCTCCCCTGCCCTCACTGAACC
WeddSeal   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGCGTGCTGGTACCTCCCCTGCCCTCACTGAACC
Galago     GCGGGCAGCTTCCAGGCCCGTCTGCAGCTCCGGGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
MousLmur   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Sifaka     GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Macaque    GCCGGCAGCTTCCAGGCGCGGCTGCGGCTGCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Mangebey   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTGCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Baboon     GCCGGCAGCTTCCAGGCGCGGCTGCGGCTGCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Drill      GCCGGCAGCTTCCAGGCGCGGCTGCGGCTGCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Colobus    GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
GLSnMonk   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
BLSnMonk   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Bonobo     GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Orangtan   GCCGGCAGCTTCCAGGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
NiteMonk   GCTGGCAGCTTCCAAGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATT
Capuchin   GCTGGCAGCTTCCAAGCGCGGCTGCGACTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Saimiri    GCTGGCAGCTTCCAAGCGCGGCTGCGGCTCCGAGTGCTGGTGCCTCCCCTGCCCTCACTGAATC
Dog1       GCCGGCAGCTTCCAGGCGAGGCTGCGGCTCCGCGTGCTGGTGCCCCCCCTGCCCTCCCTGAACC
```

Selected: | Pos: | Pos (ungaped): | Selected seqs: | Cols: | Total selected chars: | Alignment: 32 sequences 1551 pos.

GDW
Genomics of Disease in Wildlife

E. Eizirik et al. / Molecular Phylogenetics and Evolution 56 (2010) 49–63

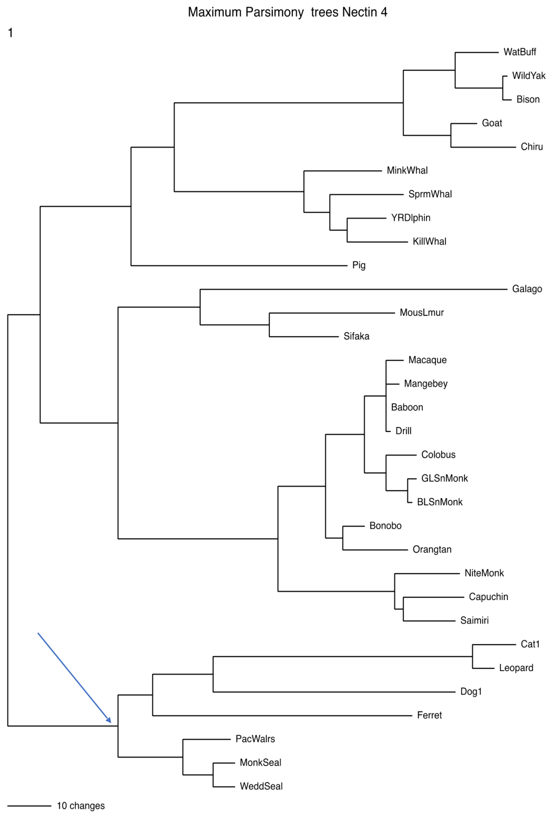


Feliformia

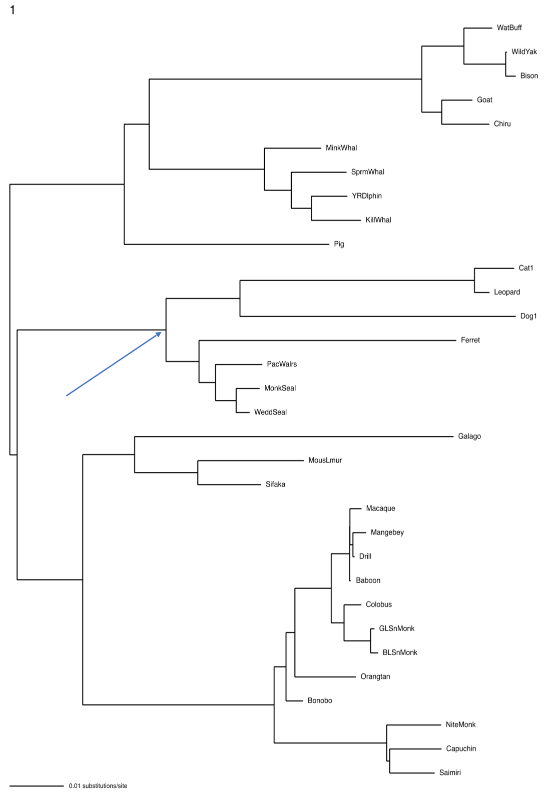
Caniformia

Nectin4: Step 4

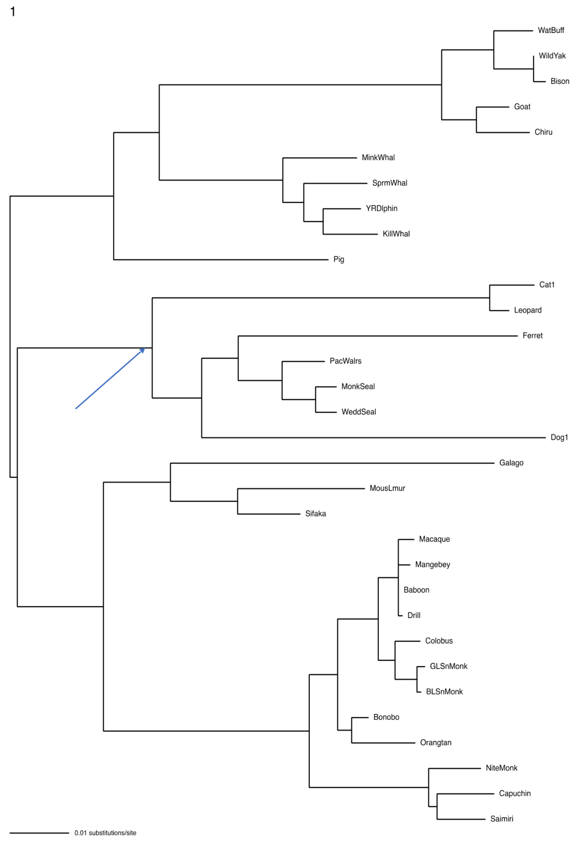
MP



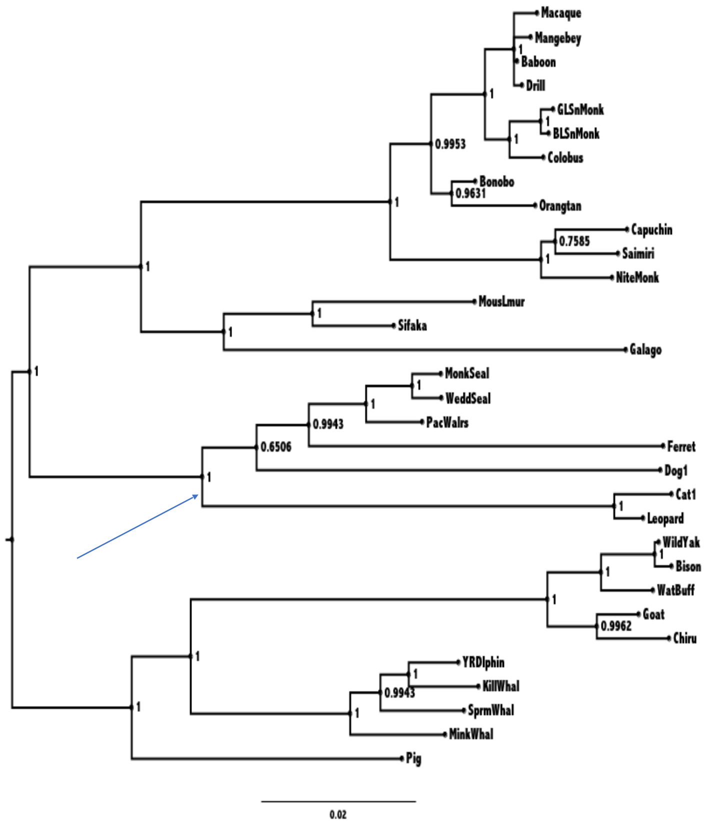
ME



ML

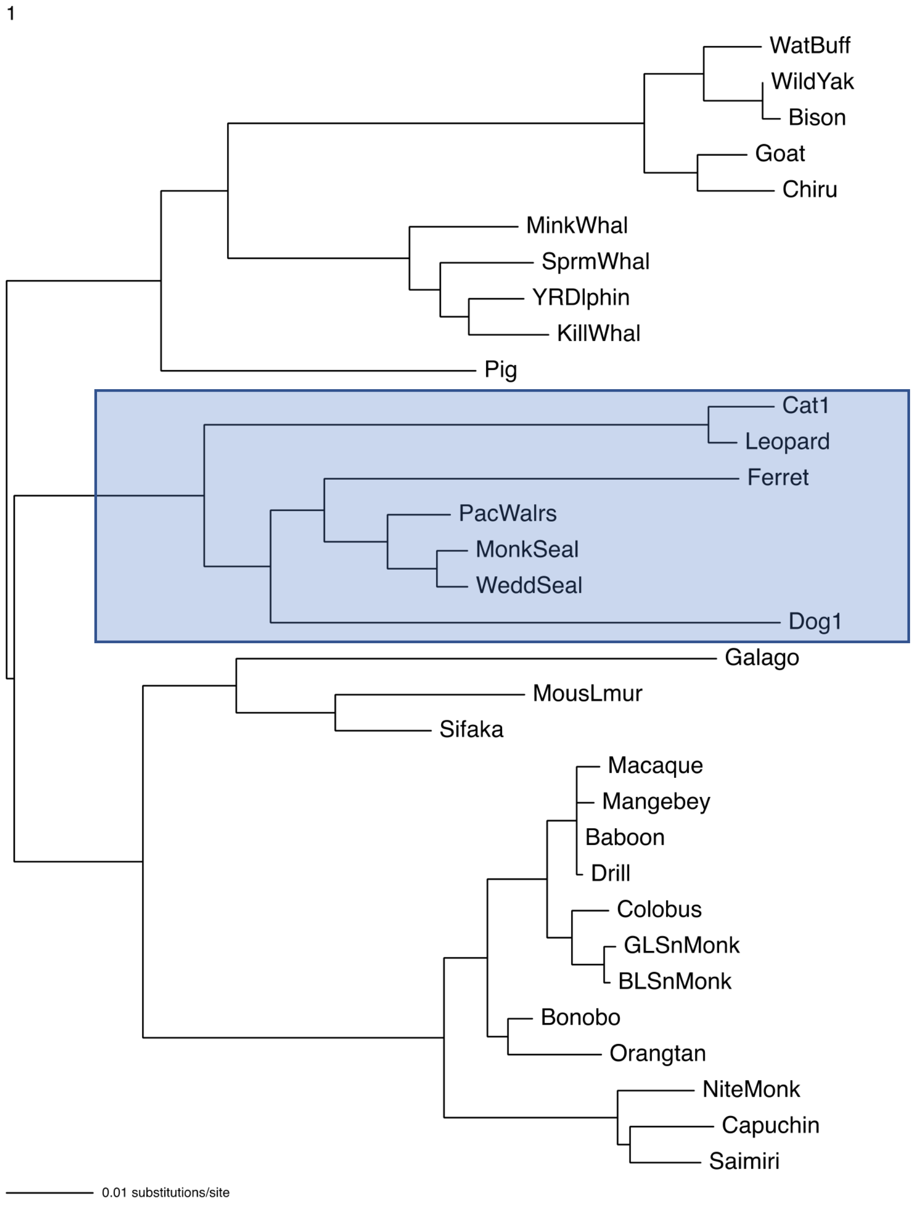


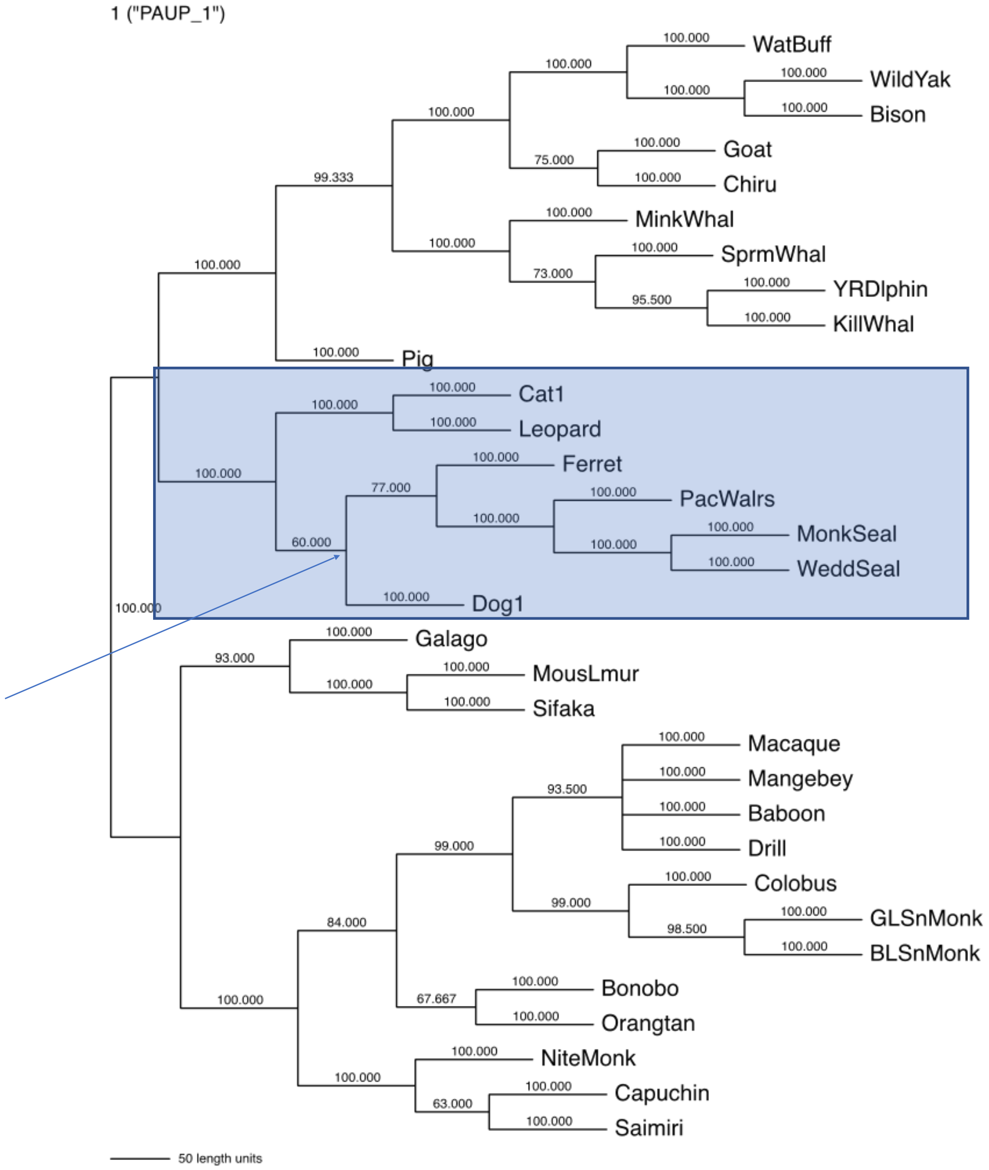
Bayes



Genomics of Disease in Wildlife

ML Phylogeny
100 BS reps





Genomics of Disease in Wildlife

Will Human Biomedical Best Practices for Characterizing DNA Variants Work for Wildlife Disease?

- Multiple unrelated families for rare monogenic conditions
  - **(No)**
- 1000’s to 10,000’s samples of ‘patients’ and ‘controls’
  - **(Maybe-depends on Scale… taxa/pathogen)**
- Consortiums to gather, curate samples
  - **(Why not!)**
- Pool cohorts guided by demography, population genetics, and phenotypes
  - **(Yes)**
- For presumed monogenic diseases, multiple families (pedigrees), but if a de novo condition, analyze the parents, using hypothesis driven testing of genes.
  - **(Yes, captive species)**
- Test hypotheses in genes known to be linked with phenotype (e.g. GO terms) or Systems Biology approaches
  - **(Yes, comparative genomics, evolutionary genomics)**
- For cases and control studies based on whole exome: Bonferroni adjusted of 0.05/30,000 genes (or whatever the number of genes in organisms genome)
  - **(Yes)**

GDW
Genomics of Disease in Wildlife

Human to Wildlife: How to identify causative variants within a gene linked with disease in Host Species

1) Association: the variant is significantly enriched in cases compared to controls.

2) Segregation: the variant is co-inherited with disease status within affected families and additional co-segregating pathogenic variants are unlikely or have been excluded.

3) Population frequency: the variant is found at a low frequency, consistent with the proposed inheritance model and disease prevalence, in large population cohorts with similar ancestry to patients.

4) Conservation: the site of the variant displays evolutionary conservation consistent with deleterious effects of sequence changes at that location.

GDW
Genomics of Disease in Wildlife

Human to Wildlife: How to identify causative variants within a gene linked with disease in Host Species

5) Predicted effect on function: variant is found at the location within the protein predicted to cause functional disruption (for example, enzyme active site, protein-binding region).

6) Gene disruption: the variant significantly alters levels, splicing or normal biochemical function of the product of the affected gene. (in vitro verification)

7) Phenotype recapitulation: introduction of the variant, or an engineered gene product carrying the variant, into a cell line or animal model results in a phenotype that is consistent with the disease and that is unlikely to arise from disruption of genes selected at random.

8) Rescue: the cellular phenotype in patient-derived cells, model organisms, or engineered equivalents can be rescued by addition of wild-type gene product or specific knockdown of the variant allele.

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?

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.

Molecular Selection In Coding Sequences

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

dN – nonsynonymous (missense) substitution

dS – synonymous substitutions

ω = dN/dS

ω = 1 (Neutral), no functional effect

ω < 1 (Purifying selection), selected to maintain function

ω > 1 (Diversifying selection), selected for adaptation to change in function

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.

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.

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

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.

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).

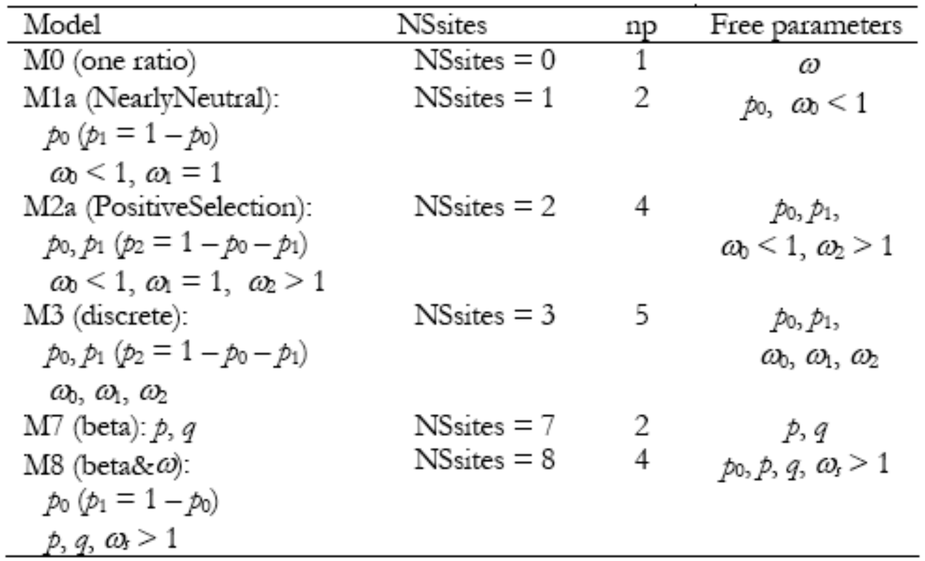
- $\Delta\lambda = 2\ (l_1 - l_0)$ chi-square 2 d.f.

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

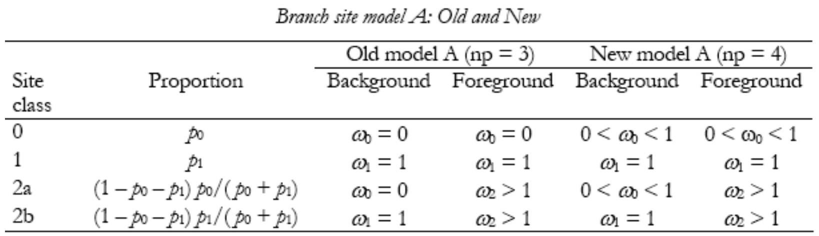
Model M1a and Model M2a, 2 df

Model M7 and Model M8, 2 df

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

		Old model A (np = 3)		New model A (np = 4)	
Site class	Proportion	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	$(1 - p_0 - p_1)\, p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$	$\omega_1 = 1$	$\omega_2 > 1$

Nectin4: Step 5

Tests for Selection with PAML

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

Likelihood values:

Among Sites Models

- Model 0
- Model 1
- Model 2
- Model 7
- Model 8

Branch Site Models

- Carnivores
- Dog-like
- Cat-like

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)

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

Future Directions?