PoSeiDon

A Web Server for the Detection of Evolutionary Recombination Events and Positive Selection

Martin Hölzer

February 15, 2017

Friedrich Schiller University Jena
RNA Bioinformatics and High Throughput Analysis

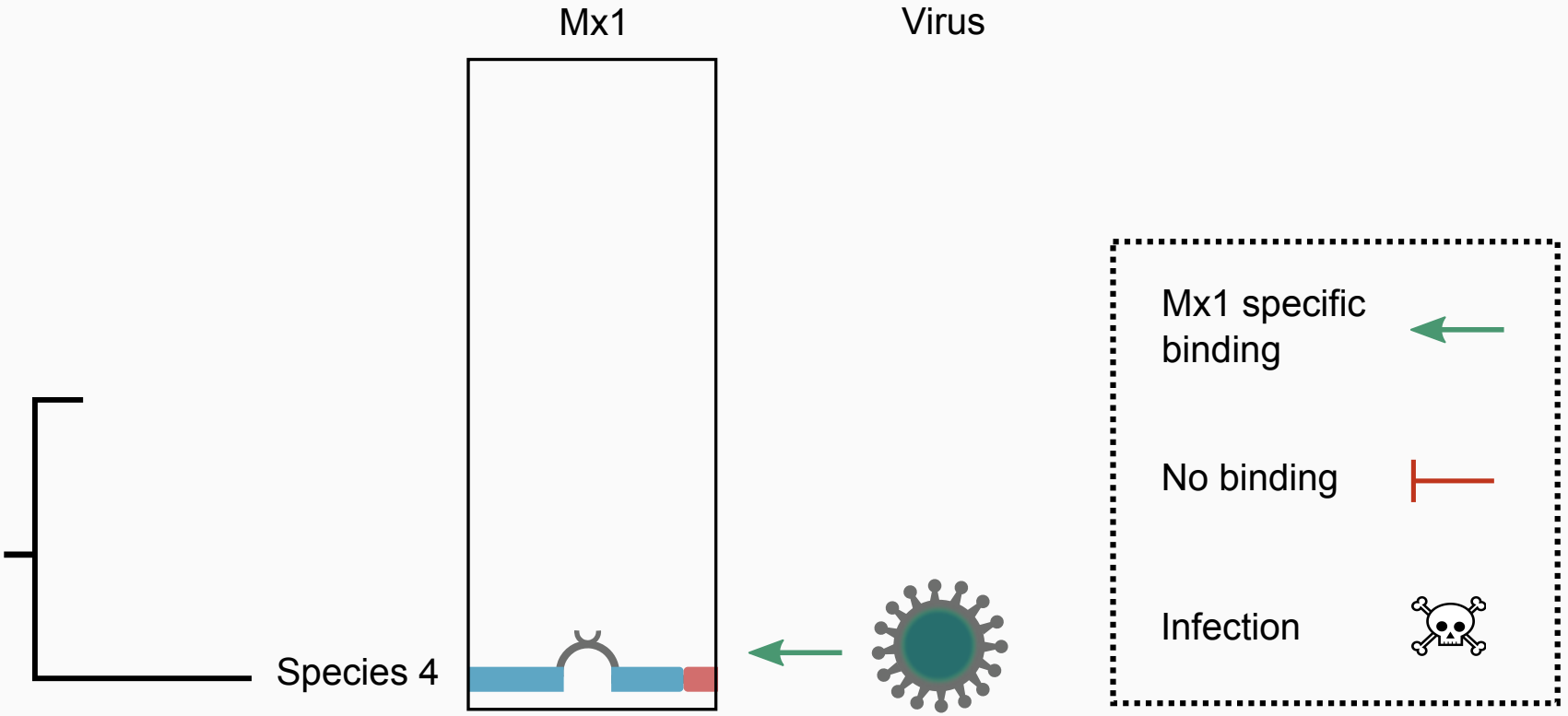
32nd TBI Winterseminar in Bled

Evolutionary Biology

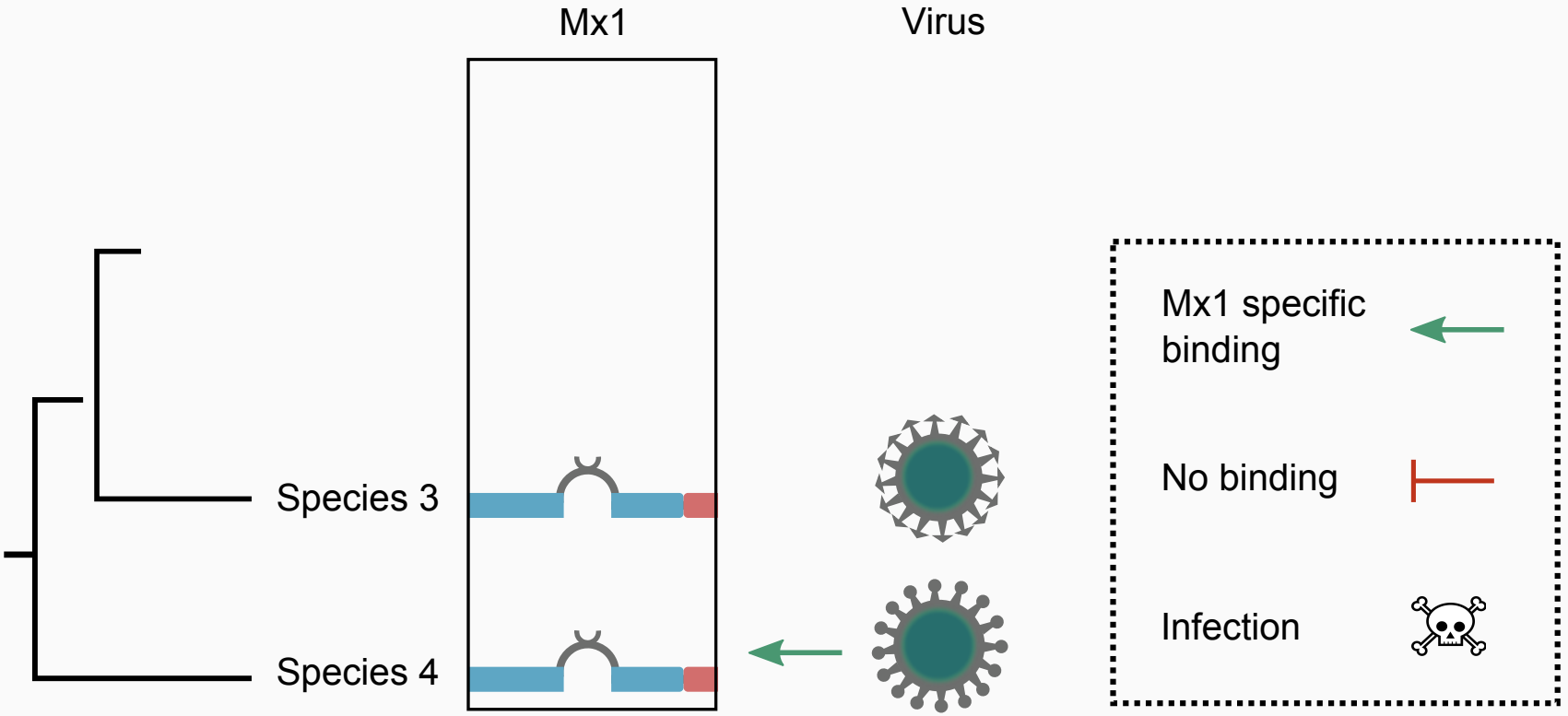
How do protein-coding genes evolve through time and in the various organisms that exist today?

Positive Selection

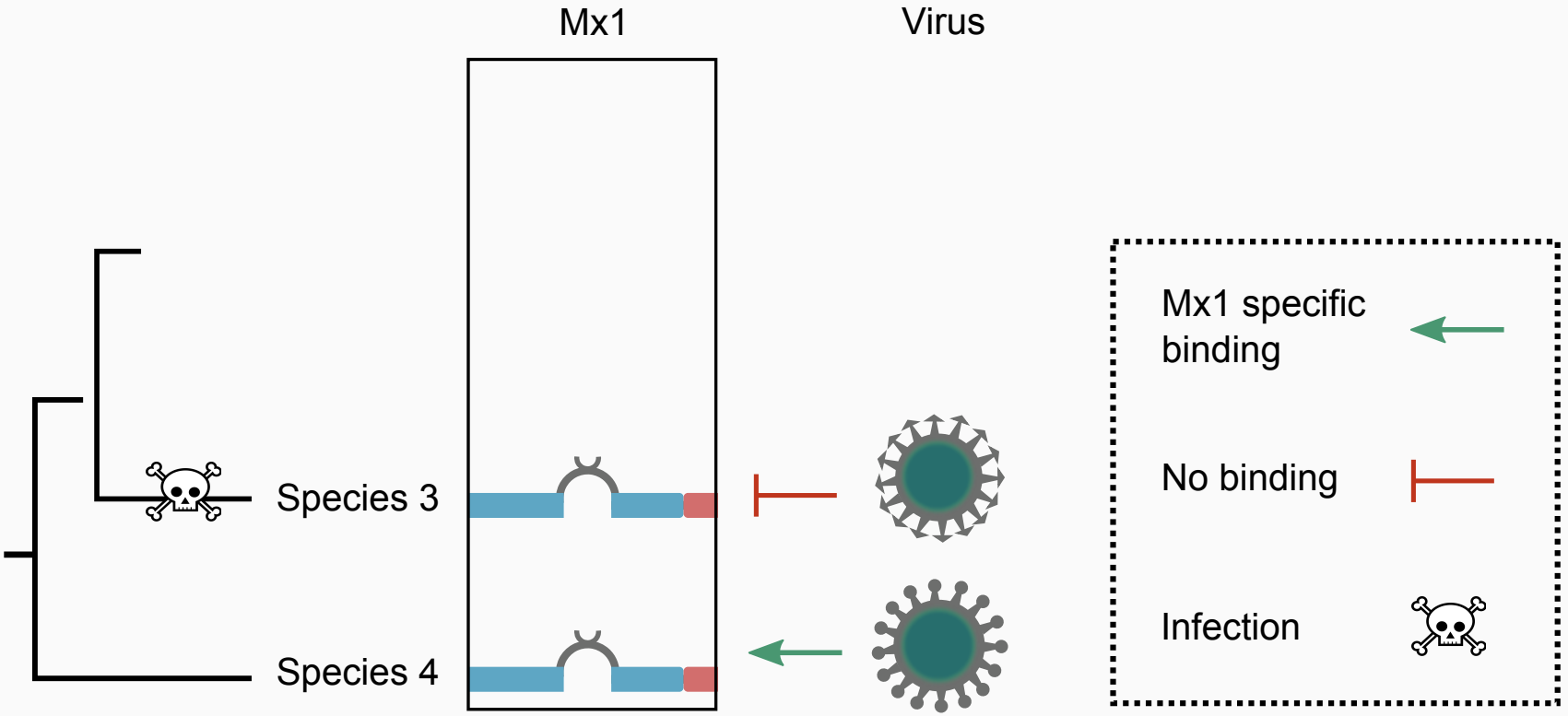
'Arms Race'


Mx1
Virus
Species 4
Mx1 specific binding
No binding
Infection

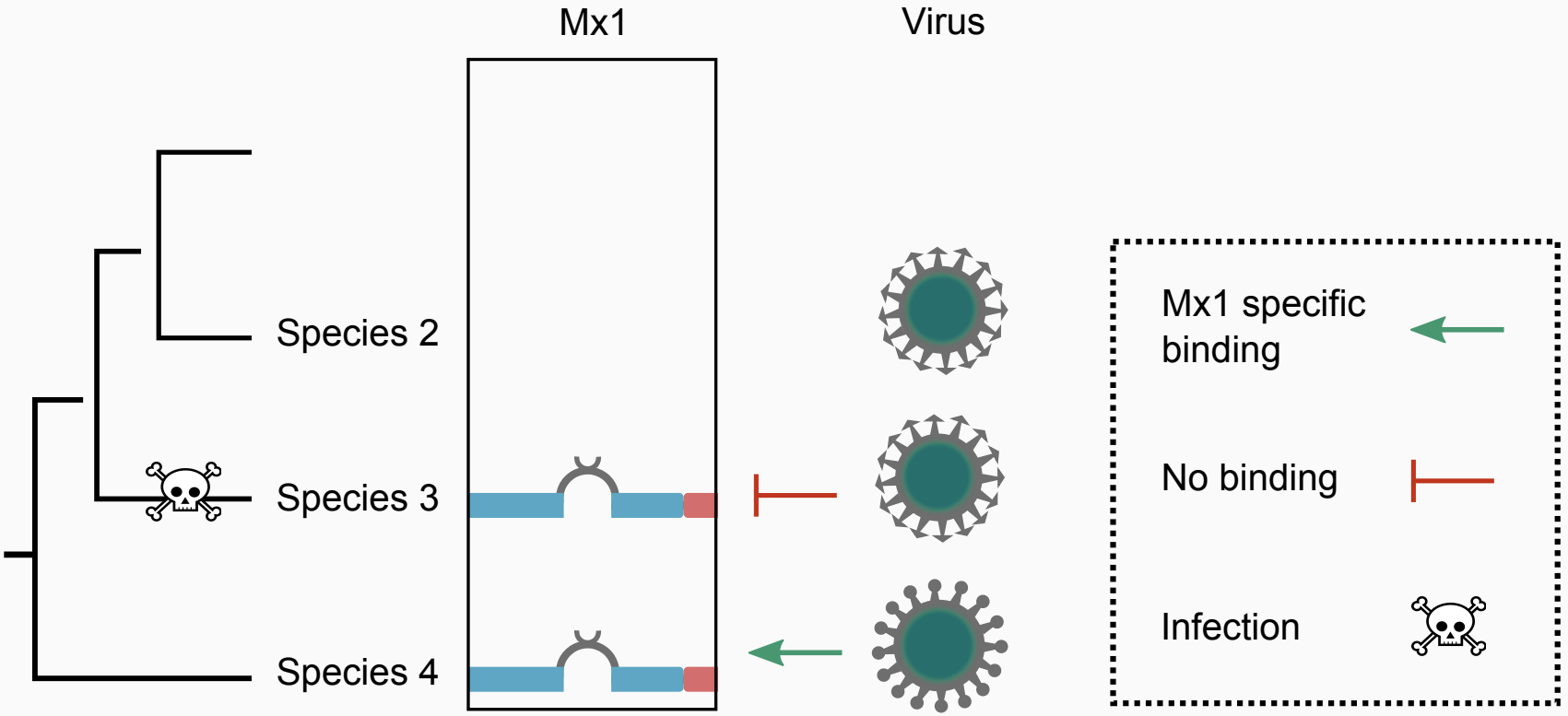
'Arms Race'


Mx1
Virus
Species 3
Species 4
Mx1 specific binding
No binding
Infection

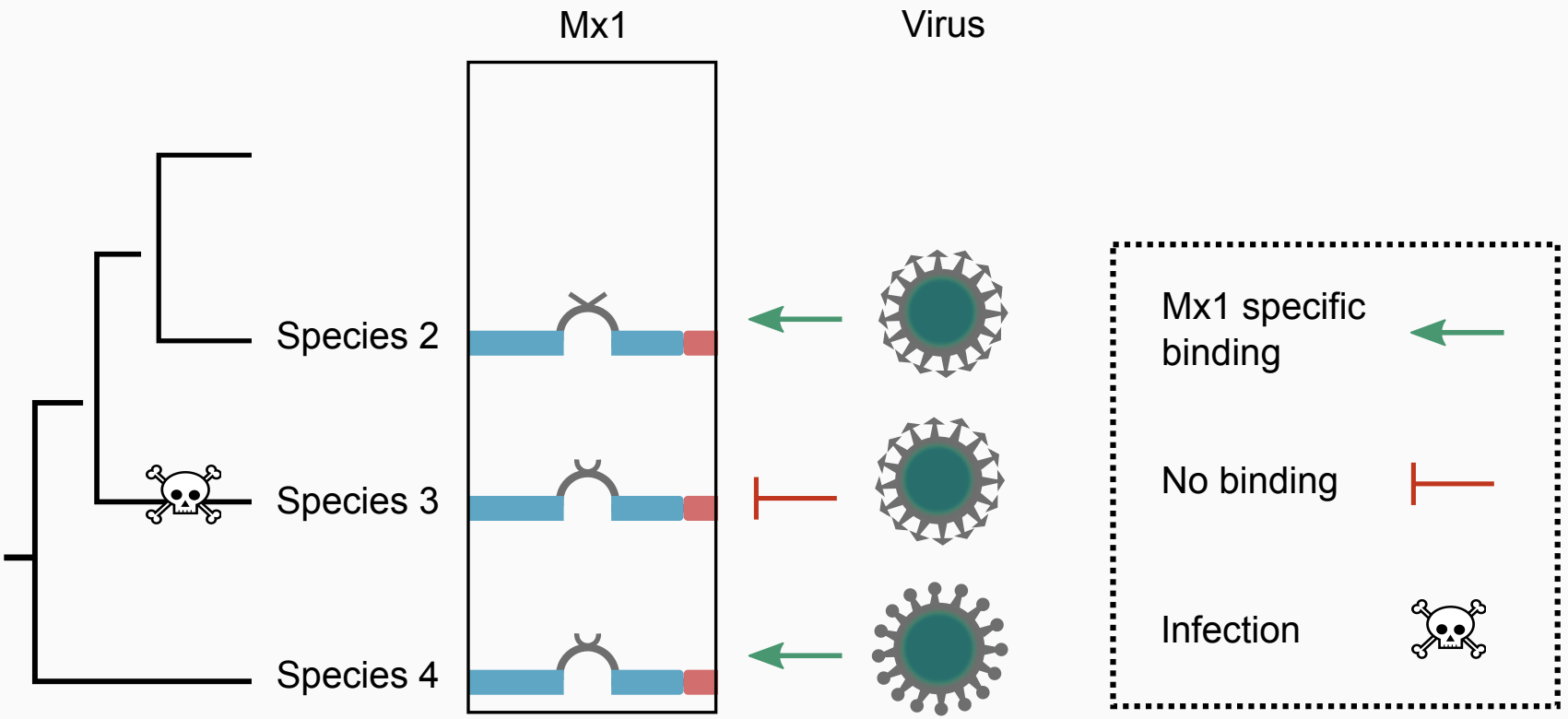
'Arms Race'


Mx1
Virus
Species 3
Species 4
Mx1 specific binding
No binding
Infection

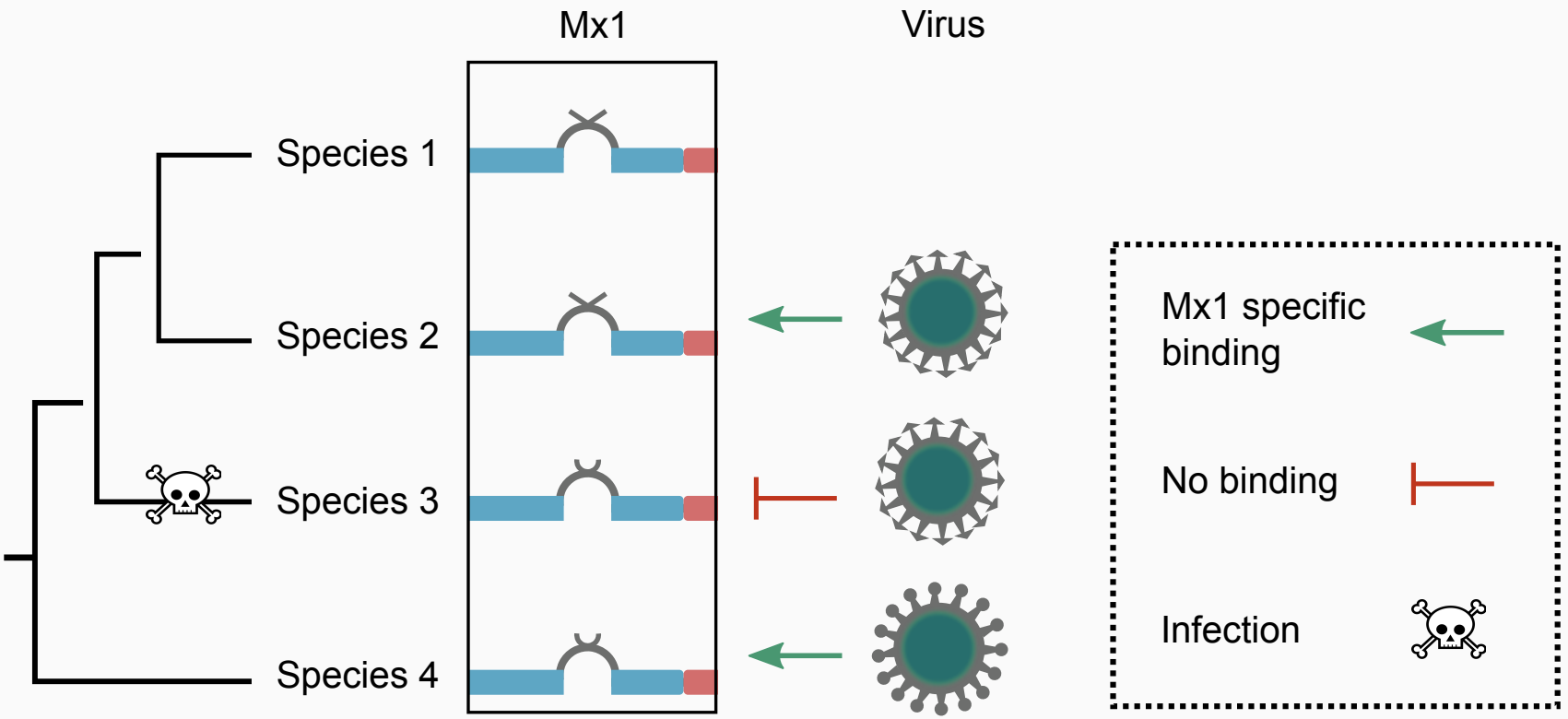
'Arms Race'


Mx1
Virus
Species 2
Species 3
Species 4
Mx1 specific binding
No binding
Infection

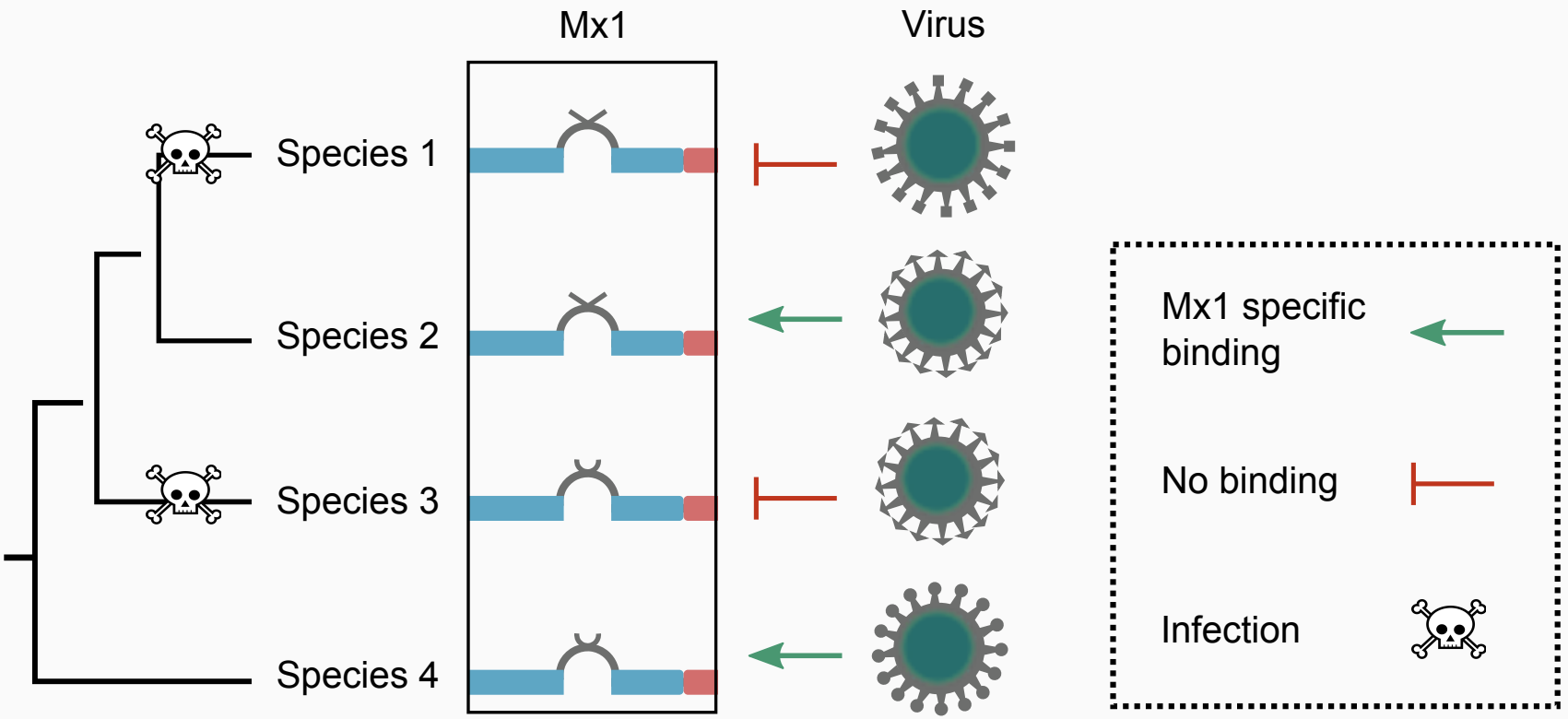
'Arms Race'


Mx1
Virus
Species 2
Species 3
Species 4
Mx1 specific binding
No binding
Infection

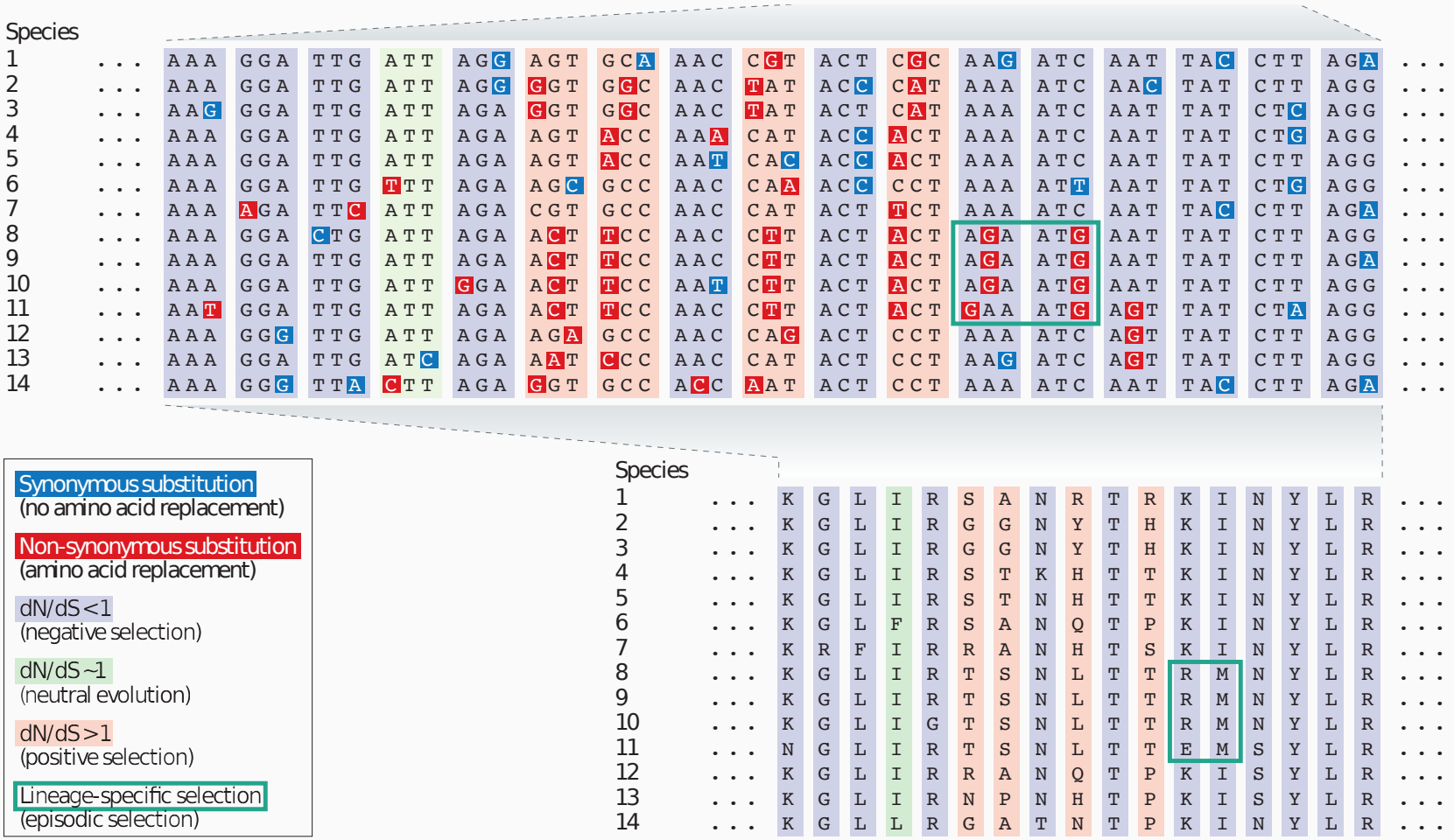
'Arms Race'


Mx1
Virus
Species 1
Species 2
Species 3
Species 4
Mx1 specific binding
No binding
Infection

'Arms Race'

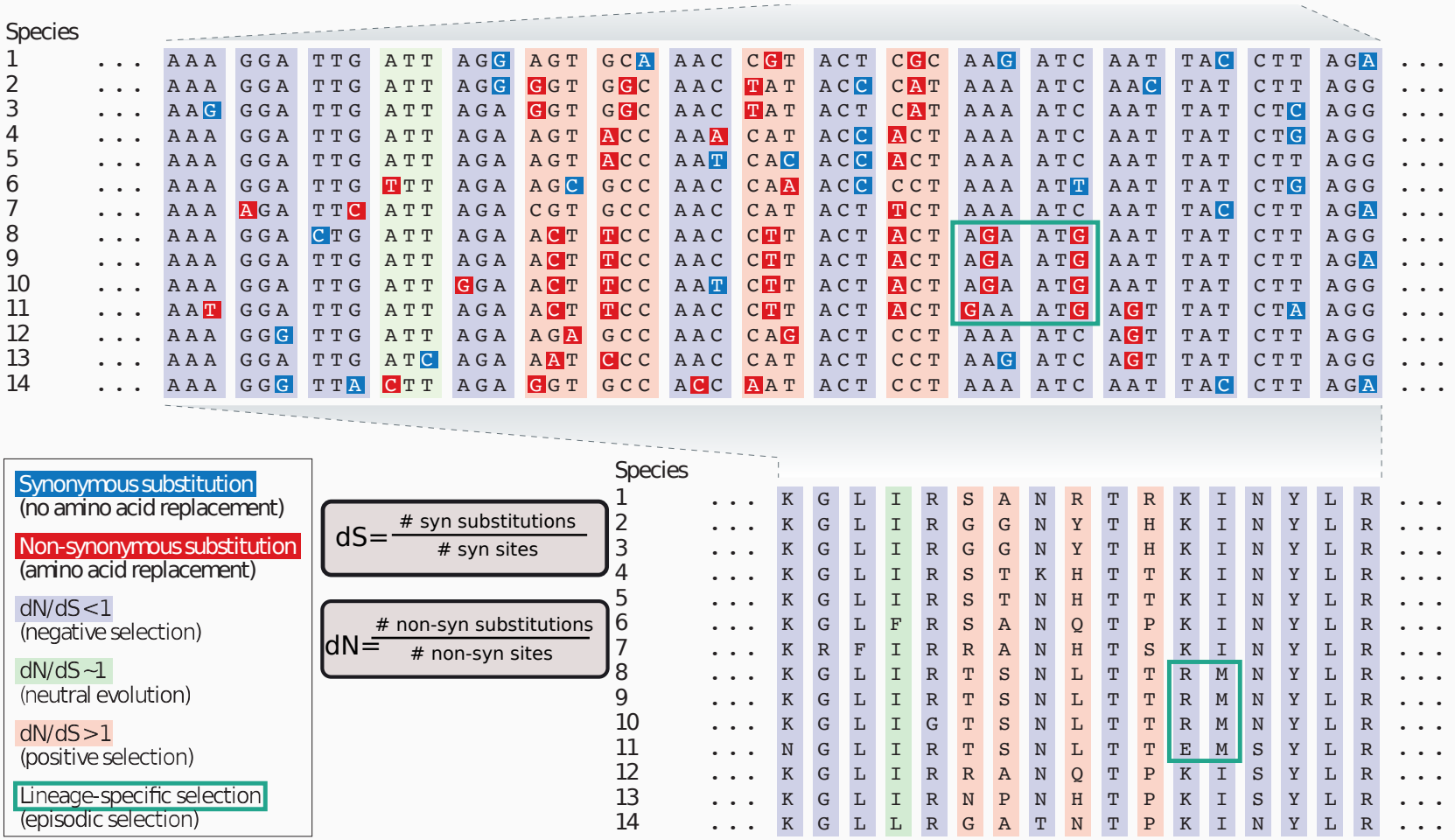

Mx1
Virus
Species 1
Species 2
Species 3
Species 4
Mx1 specific binding
No binding
Infection

Co-Evolution and detecting natural selection



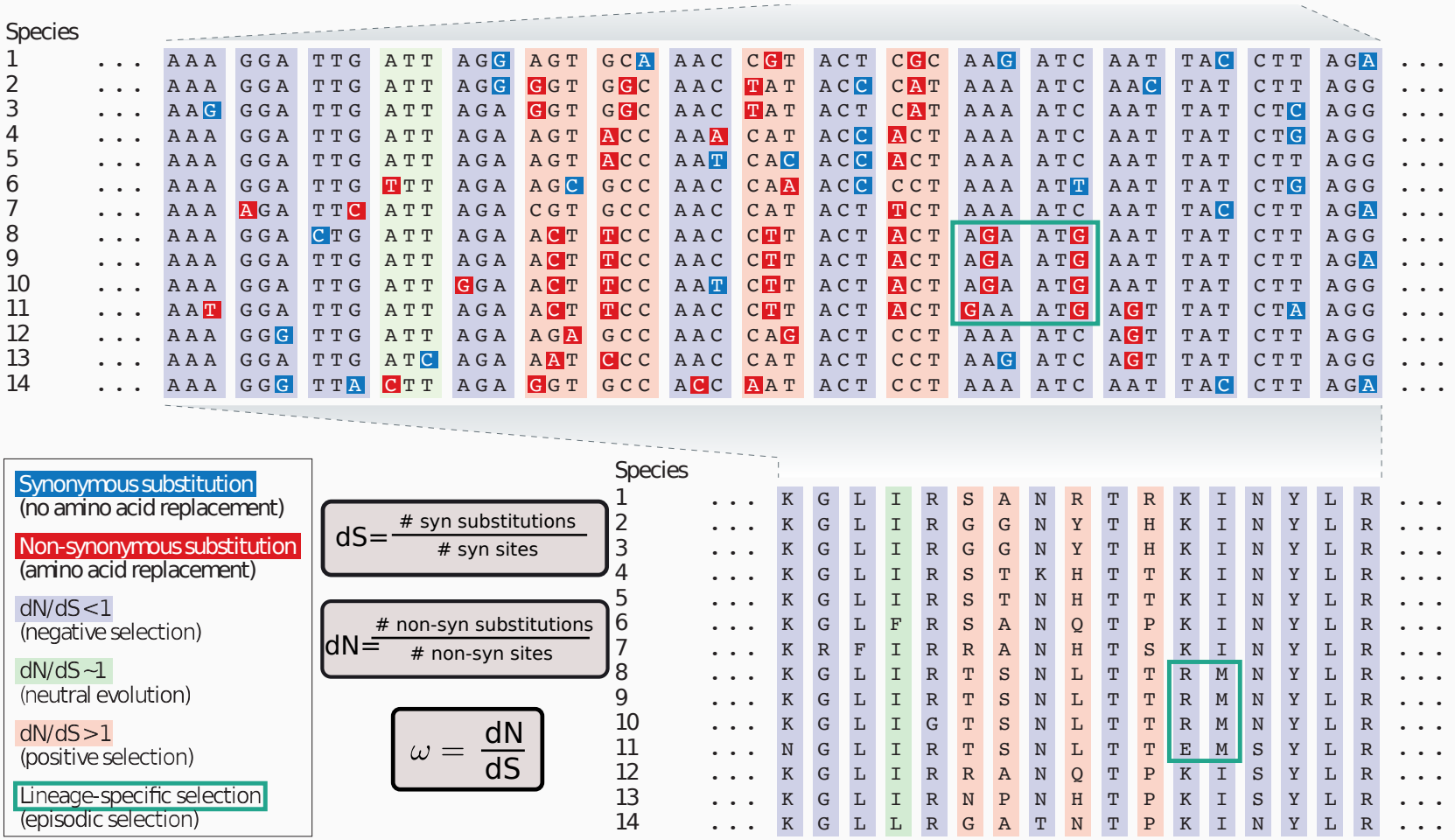
Sironi, Manuela, *et al.* "Evolutionary insights into host-pathogen interactions from mammalian sequence data." *Nature Reviews Genetics* 16.4 (2015): 224-236.

Co-Evolution and detecting natural selection



Sironi, Manuela, *et al.* "Evolutionary insights into host-pathogen interactions from mammalian sequence data." *Nature Reviews Genetics* 16.4 (2015): 224-236.

Co-Evolution and detecting natural selection

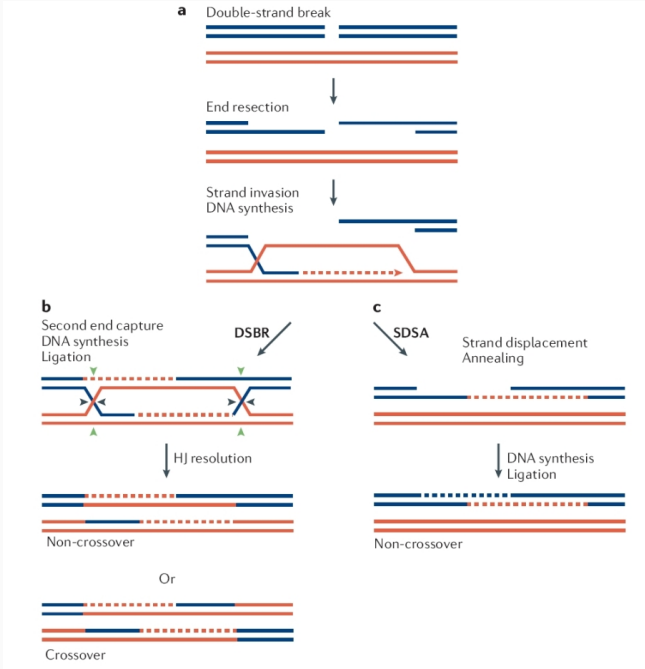


Sironi, Manuela, *et al.* "Evolutionary insights into host-pathogen interactions from mammalian sequence data." *Nature Reviews Genetics* 16.4 (2015): 224-236.

Recombination

Genetic Recombination

- Rearrangement of genetic information within and among DNA molecules
- If recombination, then possibly no unique tree topology can describe the evolutionary history of the whole sequence

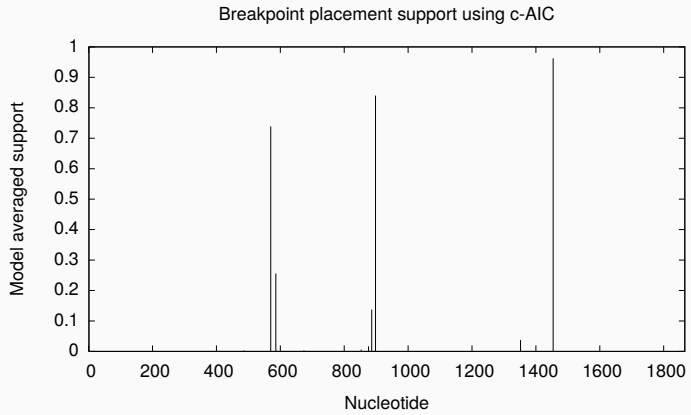


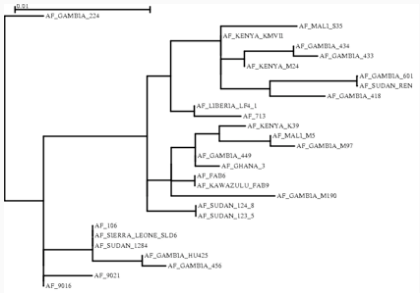
Clancy, S. (2008) "Genetic recombination." *Nature Education* 1(1):40

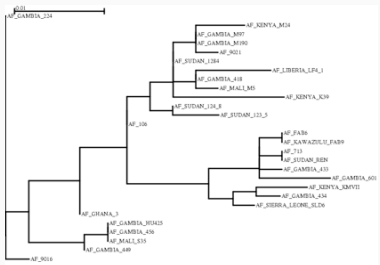
Genetic Recombination

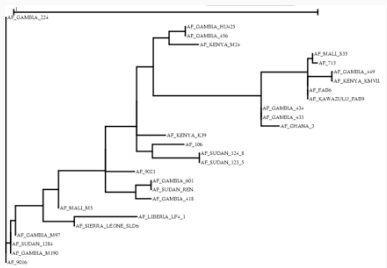
GARD found evidence of 3 breakpoints

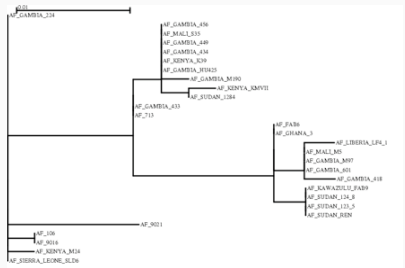
BPs	c-AIC	Δ c-AIC	Segments
0	7902.02		1-1866
1	7642.2	259.812	898
2	7518.5	123.708	724 1454
3	7424.64	93.8553	570 898 1454
4	7424.64	0	570 898 1454







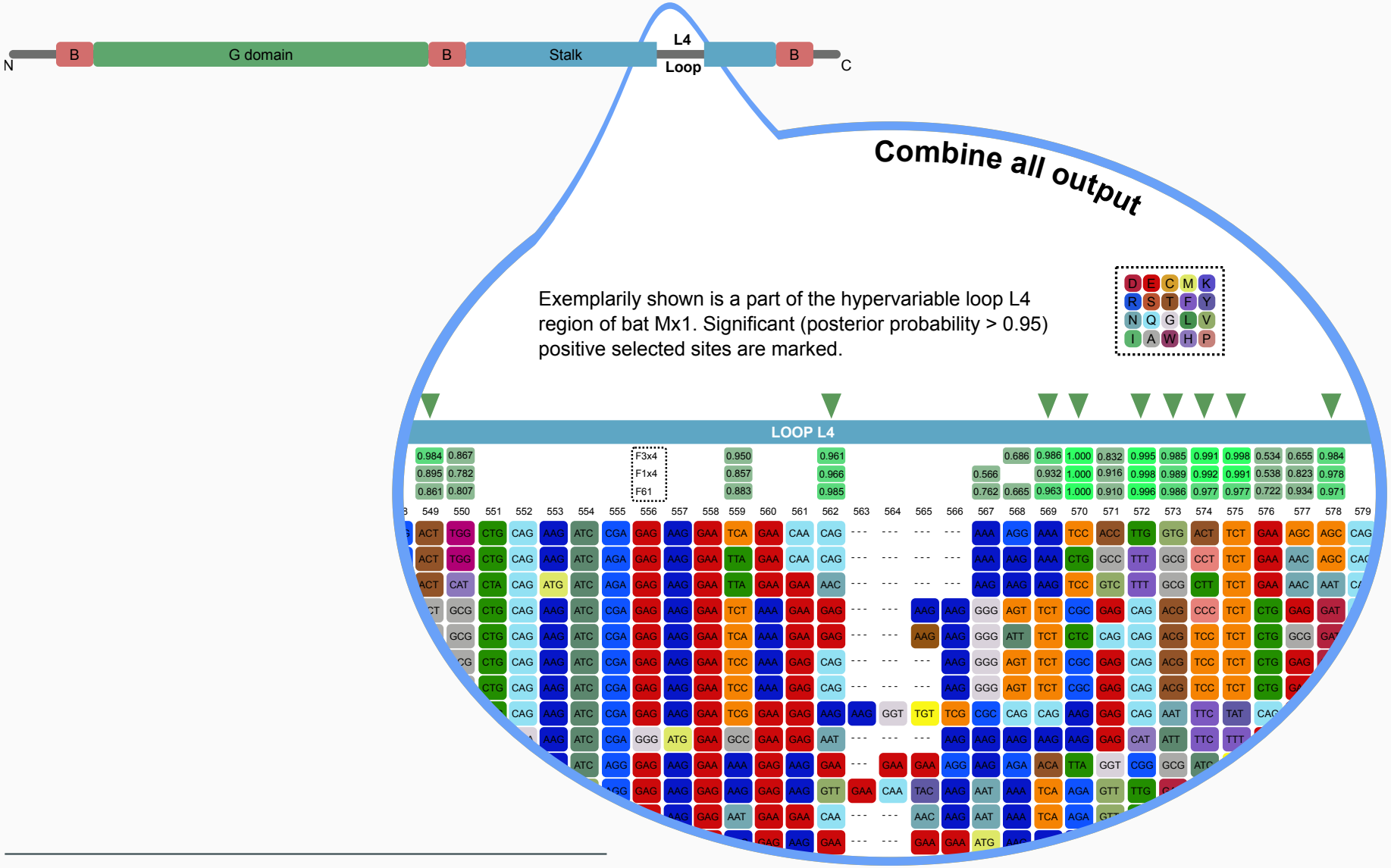




Pond, Sergei L. Kosakovsky, *et al.* "GARD: a genetic algorithm for recombination detection." *Bioinformatics* 22.24 (2006): 3096-3098.

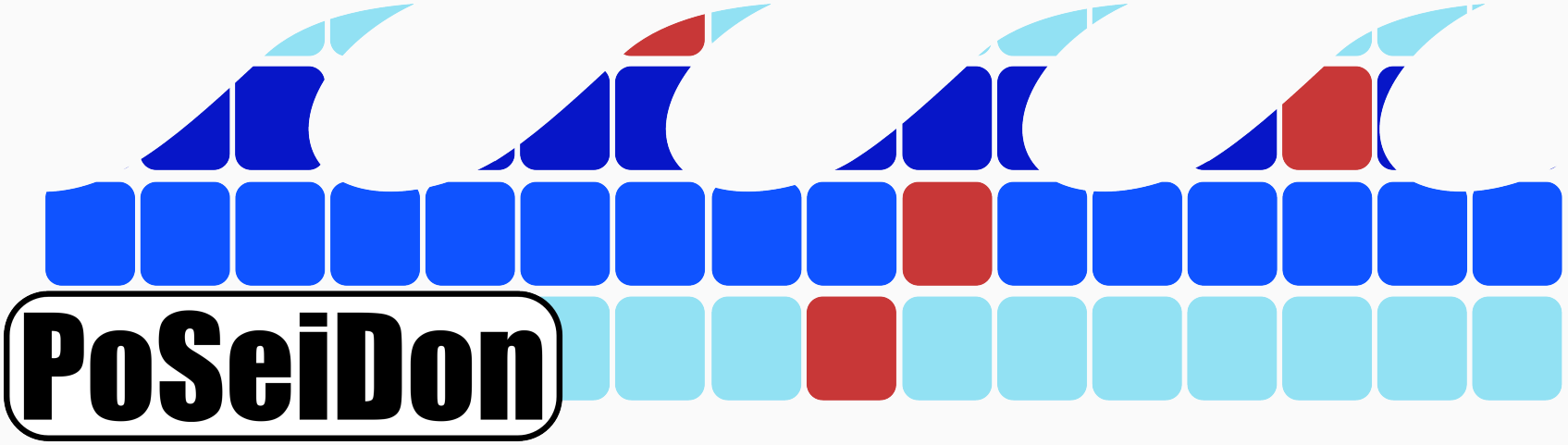
Previous Work

Positive Selection in bat MX1



Fuchs, Jonas *et al.*, "Evolution and antiviral specificity of interferon-induced Mx proteins of bats." Submitted.

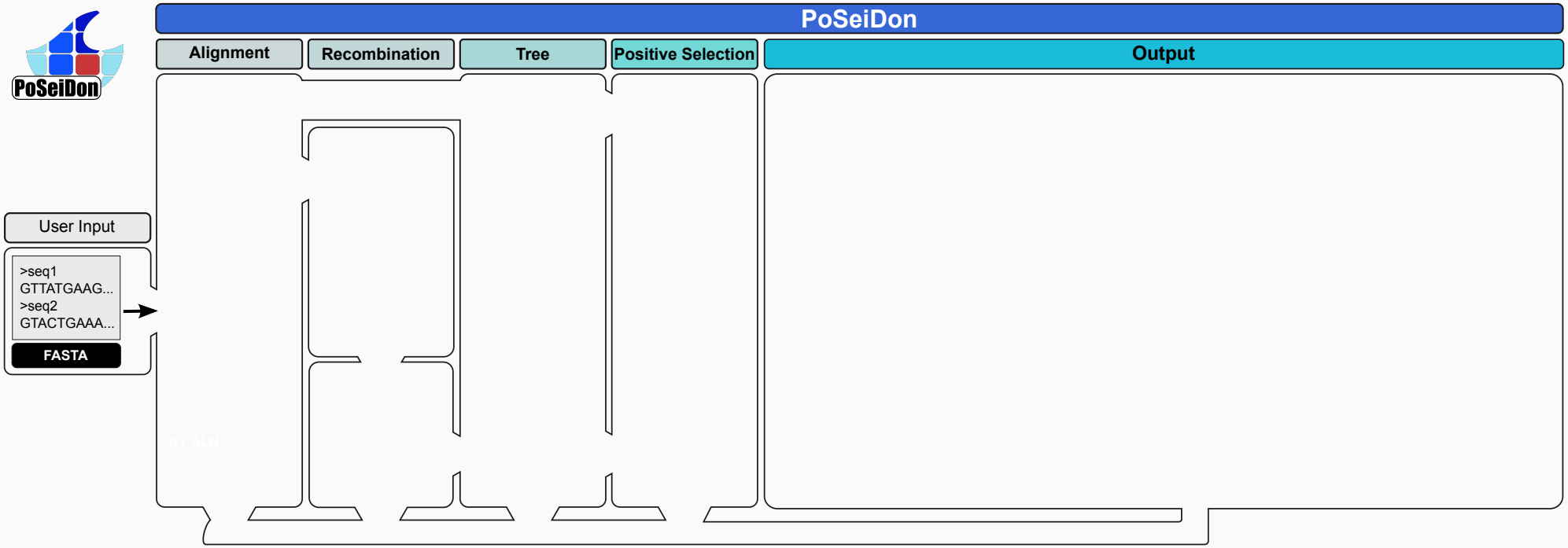
Positive Selection Detection and Recombination Analysis



Hölzer, Martin and Marz, Manja "PoSeiDon: A Web Server for the Detection of Evolutionary Recombination Events and Positive Selection." Submitted as Applications Note.

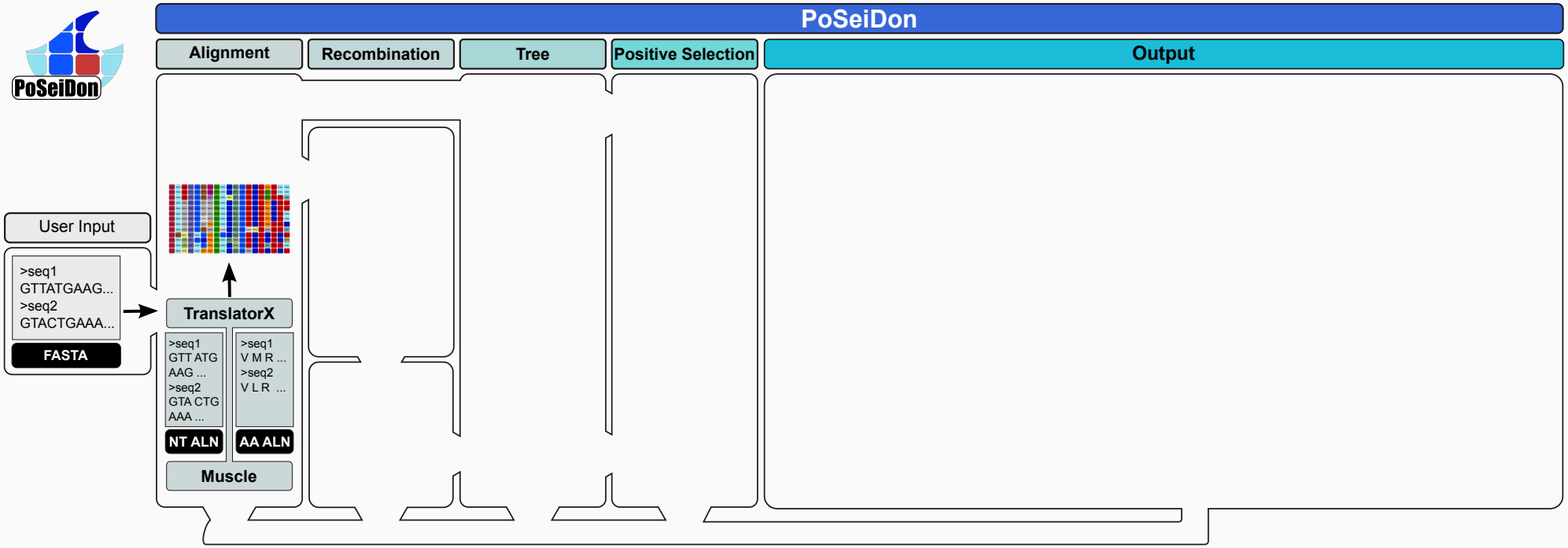
The Pipeline

The Pipeline



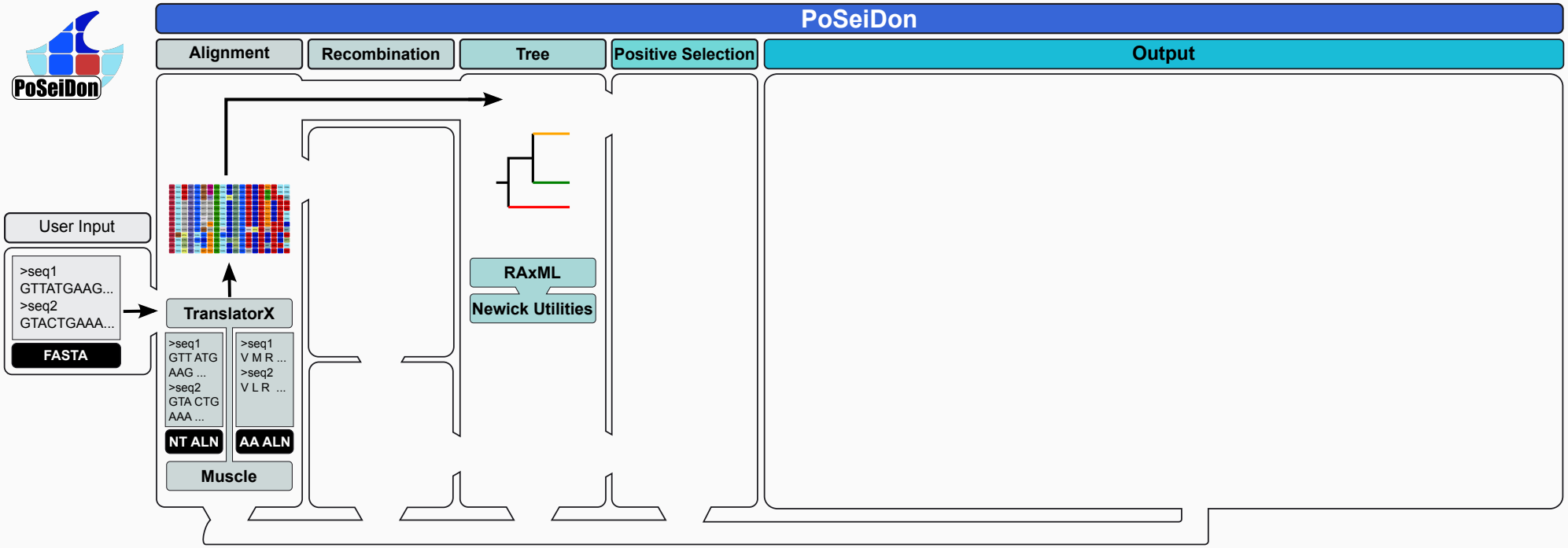
Hölzer, Martin and Marz, Manja "PoSeiDon: A Web Server for the Detection of Evolutionary Recombination Events and Positive Selection." Submitted as Applications Note.

The Pipeline



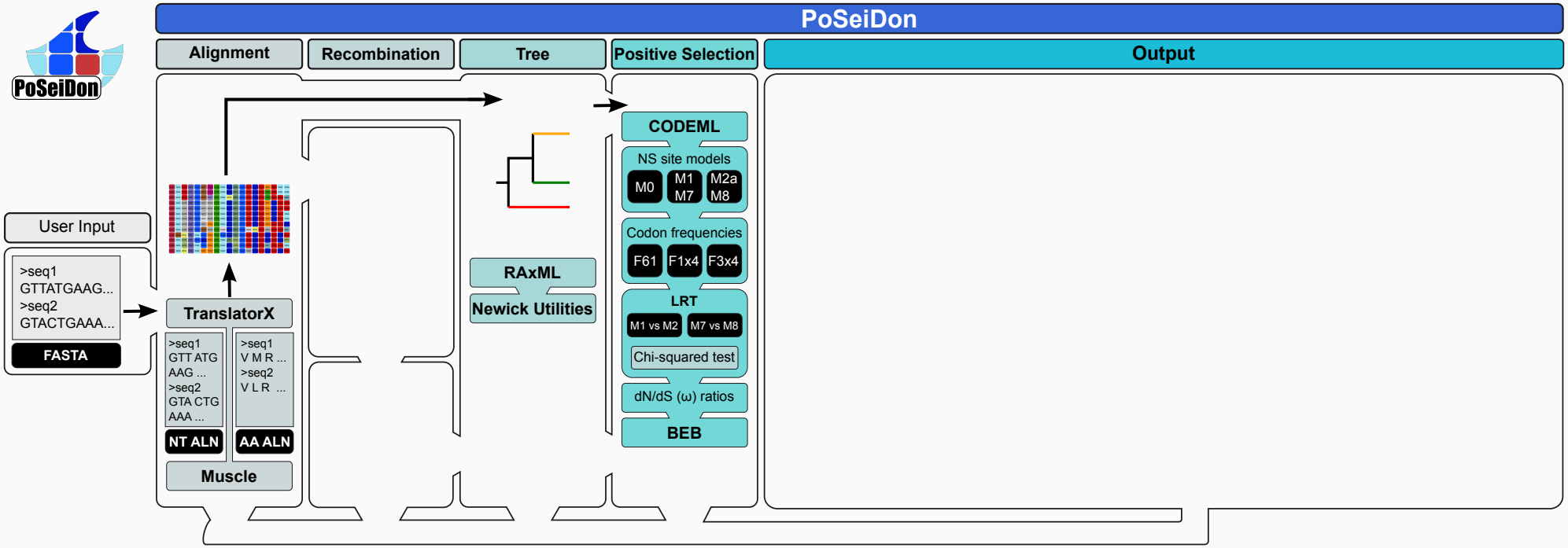
Abascal, Federico, Rafael Zardoya, and Maximilian J. Telford. "TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations." *Nucleic acids research* (2010): gkq291.

The Pipeline



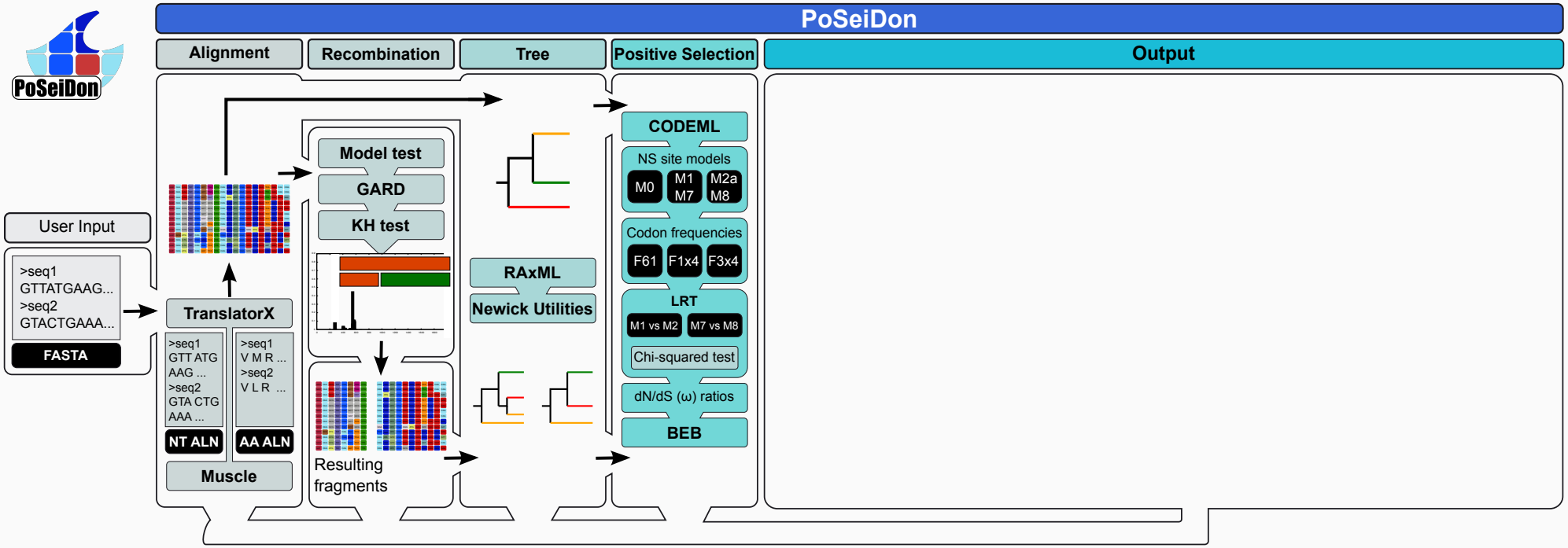
Stamatakis, Alexandros. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30.9 (2014): 1312-1313.

The Pipeline



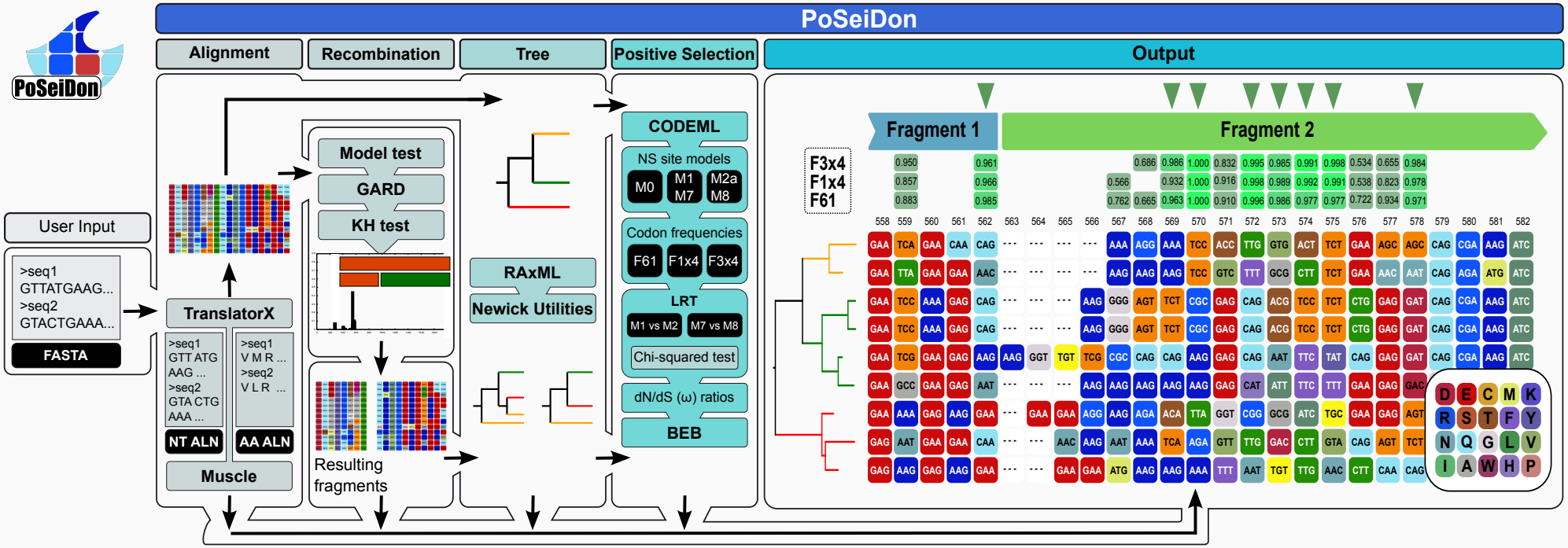
Yang, Ziheng. "PAML 4: phylogenetic analysis by maximum likelihood." *Molecular biology and evolution* 24.8 (2007): 1586-1591.

The Pipeline



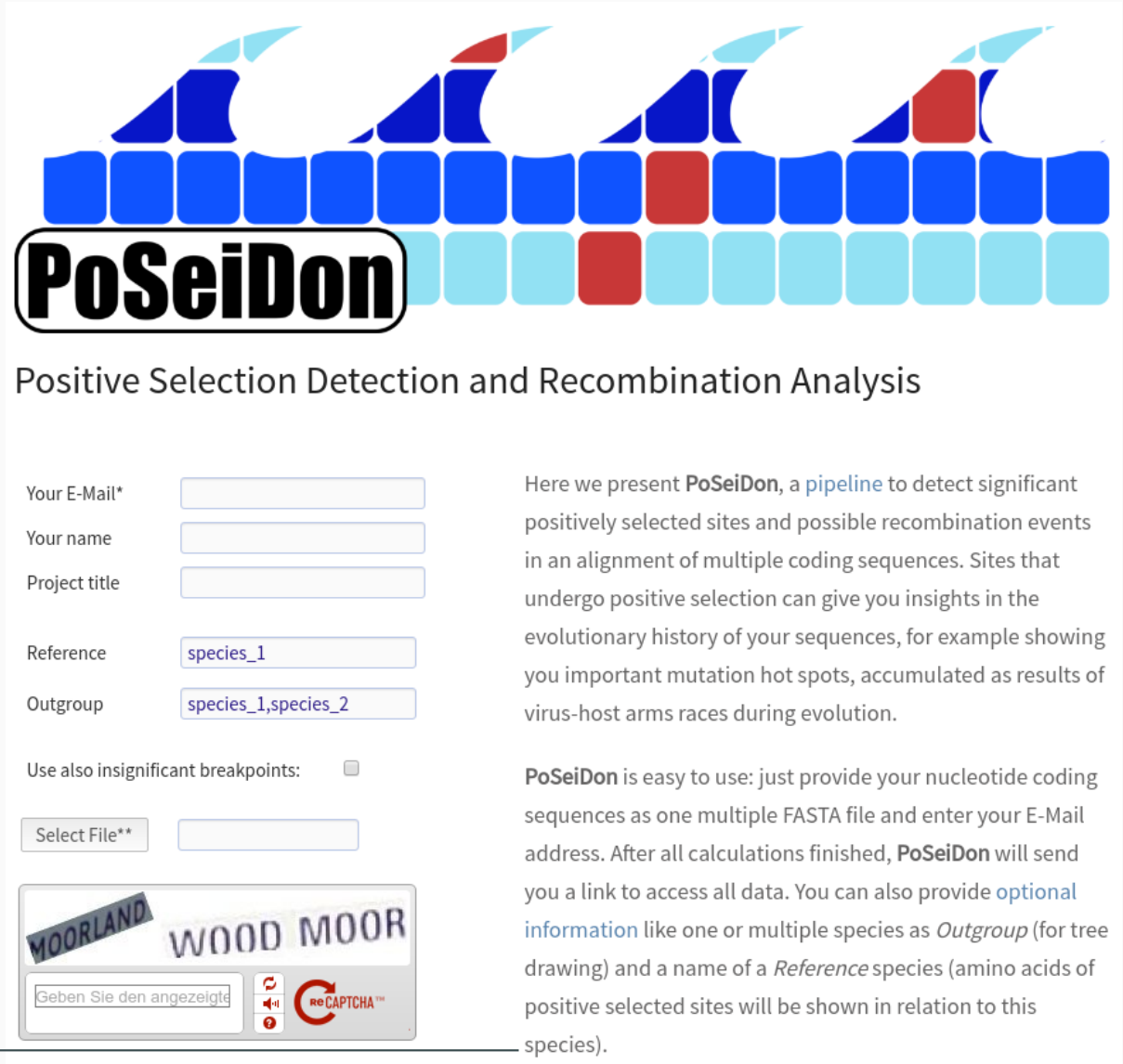
Pond, Sergei L. Kosakovsky, et al. "GARD: a genetic algorithm for recombination detection." *Bioinformatics* 22.24 (2006): 3096-3098.

The Pipeline



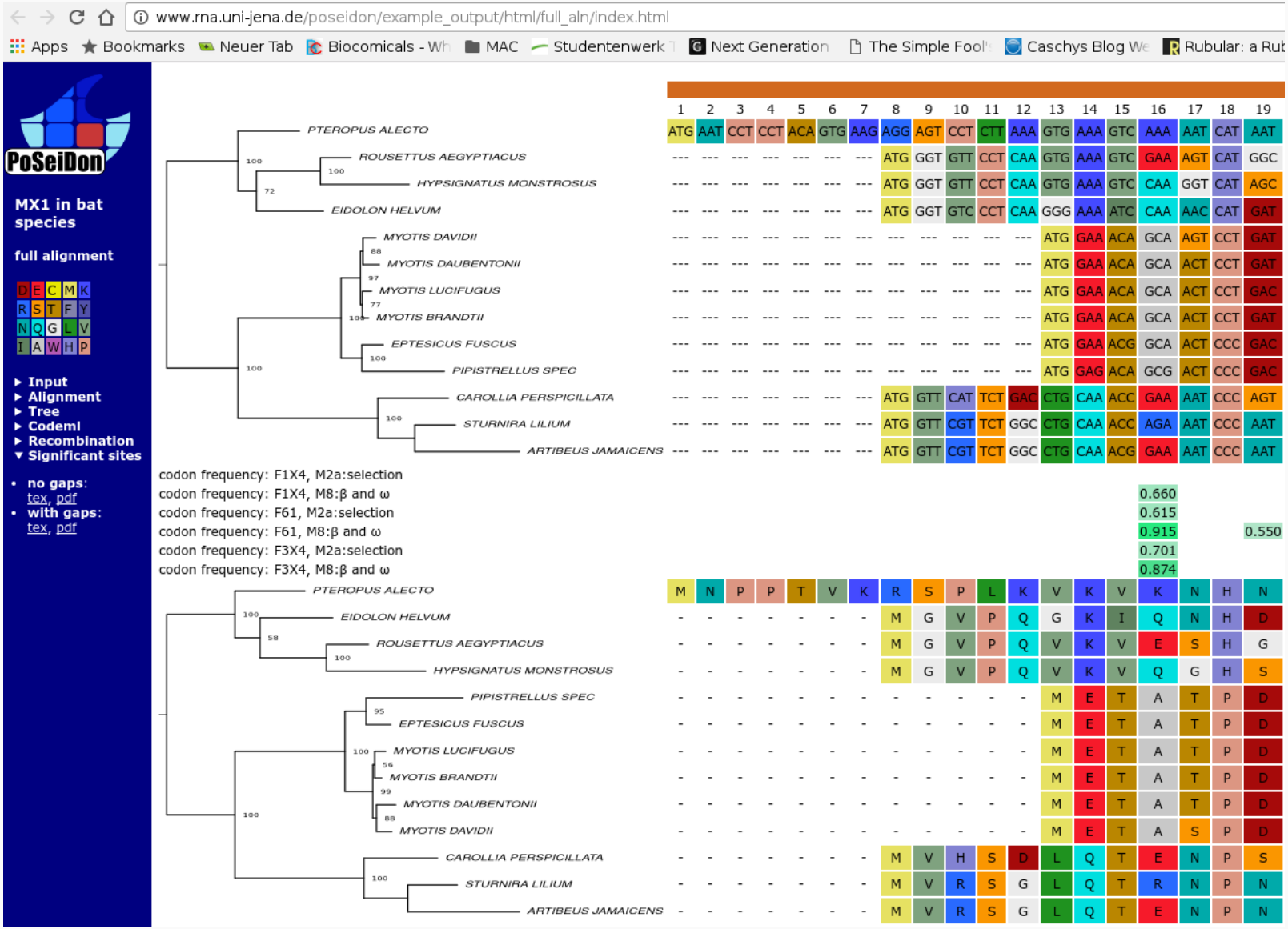
Hölzer, Martin and Marz, Manja "PoSeiDon: A Web Server for the Detection of Evolutionary Recombination Events and Positive Selection." Submitted as Applications Note.

The Interface



`www.rna.uni-jena.de/poseidon`

The Output


www.rna.uni-jena.de/poseidon/example_output/html/full_aln/index.html
Apps
Bookmarks
Neuer Tab
Biocomicals - Wh
MAC
Studentenwerk T
Next Generation
The Simple Fool
Caschys Blog We
Rubular: a Rub
PoSeiDon
MX1 in bat species
full alignment
Input
Alignment
Tree
Codeml
Recombination
Significant sites
no gaps: tex, pdf
with gaps: tex, pdf
PTEROPUS ALECTO
ROUSETTUS AEGYPTIACUS
HYPSIGNATUS MONSTROSUS
EIDOLON HELVUM
MYOTIS DAVIDII
MYOTIS DAUBENTONII
MYOTIS LUCIFUGUS
MYOTIS BRANDTII
EPTESICUS FUSCUS
PIPISTRELLUS SPEC
CAROLLIA PERSPICILLATA
STURNIRA LILIUM
ARTIBEUS JAMAICENS
codon frequency: F1X4, M2a:selection
codon frequency: F1X4, M8:β and ω
codon frequency: F61, M2a:selection
codon frequency: F61, M8:β and ω
codon frequency: F3X4, M2a:selection
codon frequency: F3X4, M8:β and ω
0.660
0.615
0.915
0.550
0.701
0.874

Outlook

Outlook

- improve interface
- support `branch-site` models
- upload your own tree/alignment
- distribute source code to run on local machines

Outlook

- improve interface
- support `branch-site` models
- upload your own tree/alignment
- distribute source code to run on local machines

- `www.rna.uni-jena.de/poseidon`

Outlook

- improve interface
- support `branch-site` models
- upload your own tree/alignment
- distribute source code to run on local machines
- check and include suggestions of the reviewers 

- `www.rna.uni-jena.de/poseidon`

Thanks!

Questions?

Backup slides

"Now, here, you see, it takes all the running you can do, to keep in the same place."

– Red Queen to Alice in Lewis Carroll's *"Through the Looking-Glass"*

Run for your Lives!

"Now, here, you see, it takes all the running you can do, to keep in the same place."

Iron Maiden *et al.* "The Number of the Beast (album)." EMI, *Battery Studios*, 39:11, London, England (1982)

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site
  - K_s: number of synonymous substitutions per synonymous site

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site
  - K_s: number of synonymous substitutions per synonymous site
- homologous genes with a K_a/K_s ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site
  - K_s: number of synonymous substitutions per synonymous site
- homologous genes with a K_a/K_s ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site
  - K_s: number of synonymous substitutions per synonymous site
- homologous genes with a K_a/K_s ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1
- however, if some of the mutations are advantageous and some disadvantageous, the ratio could be less than 1

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site
  - K_s: number of synonymous substitutions per synonymous site
- homologous genes with a K_a/K_s ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1
- however, if some of the mutations are advantageous and some disadvantageous, the ratio could be less than 1
- K_a/K_s – dN/dS – ω

Calculating selective pressure, K_a/K_s

- K_a/K_s ratio is an indicator of selective pressure acting on protein-coding genes
  - K_a: number of non-synonymous substitutions per non-synonymous site
  - K_s: number of synonymous substitutions per synonymous site
- homologous genes with a K_a/K_s ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1
- however, if some of the mutations are advantageous and some disadvantageous, the ratio could be less than 1
- K_a/K_s – dN/dS – ω κ – Ts/Tv

Example: calculating K_a/K_s

Val			Met			Arg			Thr			
G	T	T	A	T	G	A	A	G	A	C	C	Total

Example: calculating K_a/K_s

		Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	

Example: calculating K_a/K_s

	Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$

Example: calculating K_a/K_s

	Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

Example: calculating K_a/K_s

	Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

- nucleotide sites give $9\frac{2}{3}$ non-synonymous sites and $2\frac{1}{3}$ synonymous sites in this peptide

Example: calculating K_a/K_s

	Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

- nucleotide sites give $9\frac{2}{3}$ non-synonymous sites and $2\frac{1}{3}$ synonymous sites in this peptide

Val			**Leu**			Arg			Thr			
G	T	**A**	**C**	T	G	A	A	**A**	A	C	C	Total

Example: calculating K_a/K_s

	Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

- nucleotide sites give $9\frac{2}{3}$ non-synonymous sites and $2\frac{1}{3}$ synonymous sites in this peptide

	Val			**Leu**			Arg			Thr			
	G	T	**A**	**C**	T	G	A	A	**A**	A	C	C	Total
# substitutions													
# non-syn	0	0	0	1	0	0	0	0	0	0	0	0	1

Example: calculating K_a/K_s

	Val			Met			Arg			Thr			
	G	T	T	A	T	G	A	A	G	A	C	C	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

- nucleotide sites give $9\frac{2}{3}$ non-synonymous sites and $2\frac{1}{3}$ synonymous sites in this peptide

	Val			**Leu**			Arg			Thr			
	G	T	**A**	**C**	T	G	A	A	**A**	A	C	C	Total
# substitutions													
# non-syn	0	0	0	1	0	0	0	0	0	0	0	0	1
# syn	0	0	1	0	0	0	0	0	1	0	0	0	2

Example: calculating K_a/K_s

Now,

$$K_a = \frac{\#nonsyn\_substitutions}{\#nonsyn\_sites} = \frac{1}{9\frac{2}{3}} = 0.103$$

and

$$K_s = \frac{\#syn\_substitutions}{\#syn\_sites} = \frac{2}{2\frac{1}{3}} = 0.857$$

Thus,

$$\frac{K_a}{K_s} = \frac{dN}{dS} = \omega = \frac{0.103}{0.857} = 0.12$$

CodeML: **Output**

- *CodonFreq=F3x4* (estimation of codon frequency distribution)
- *NSsites=M2a* (positive selection)

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

NOTE.—The site models are implemented using the control variable NSsites in CODEML, and p is the number of free parameters in the ω distribution.

$\Longrightarrow$ [codeml.variable.mlc]