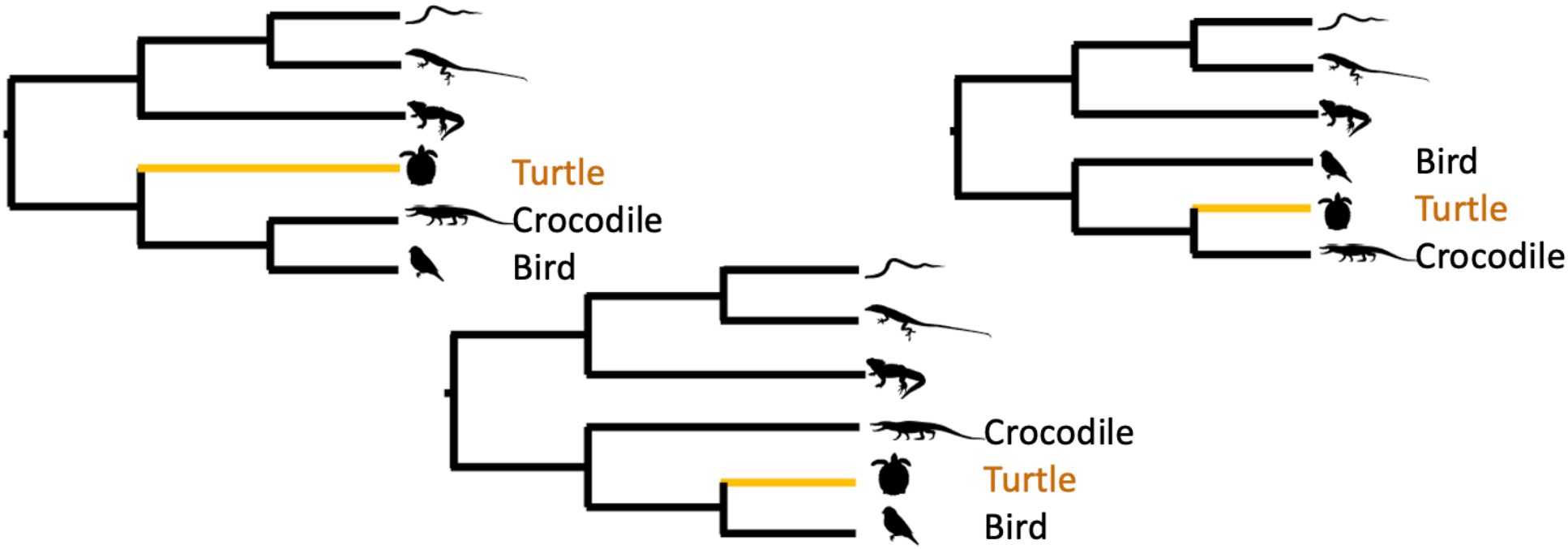
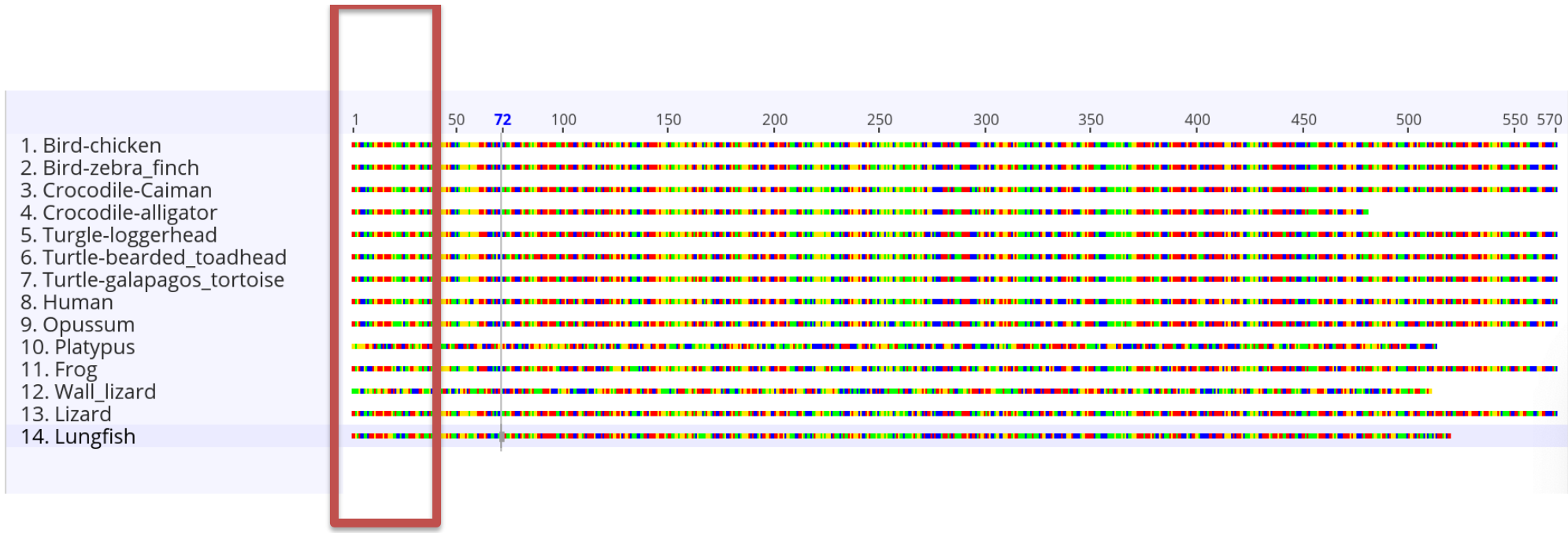
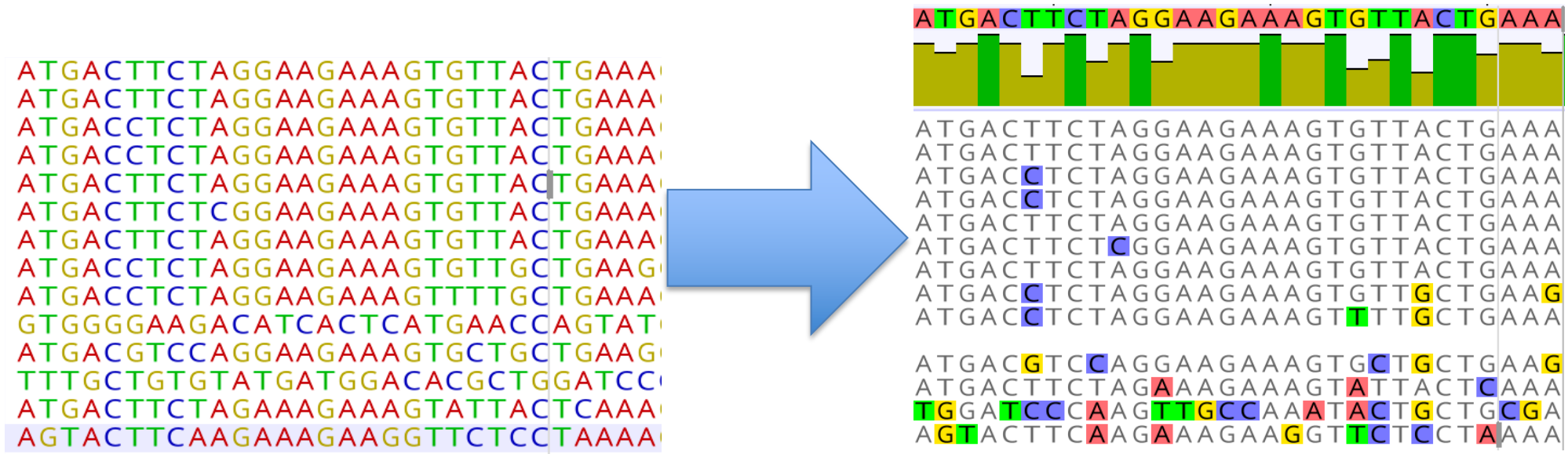
Alignment

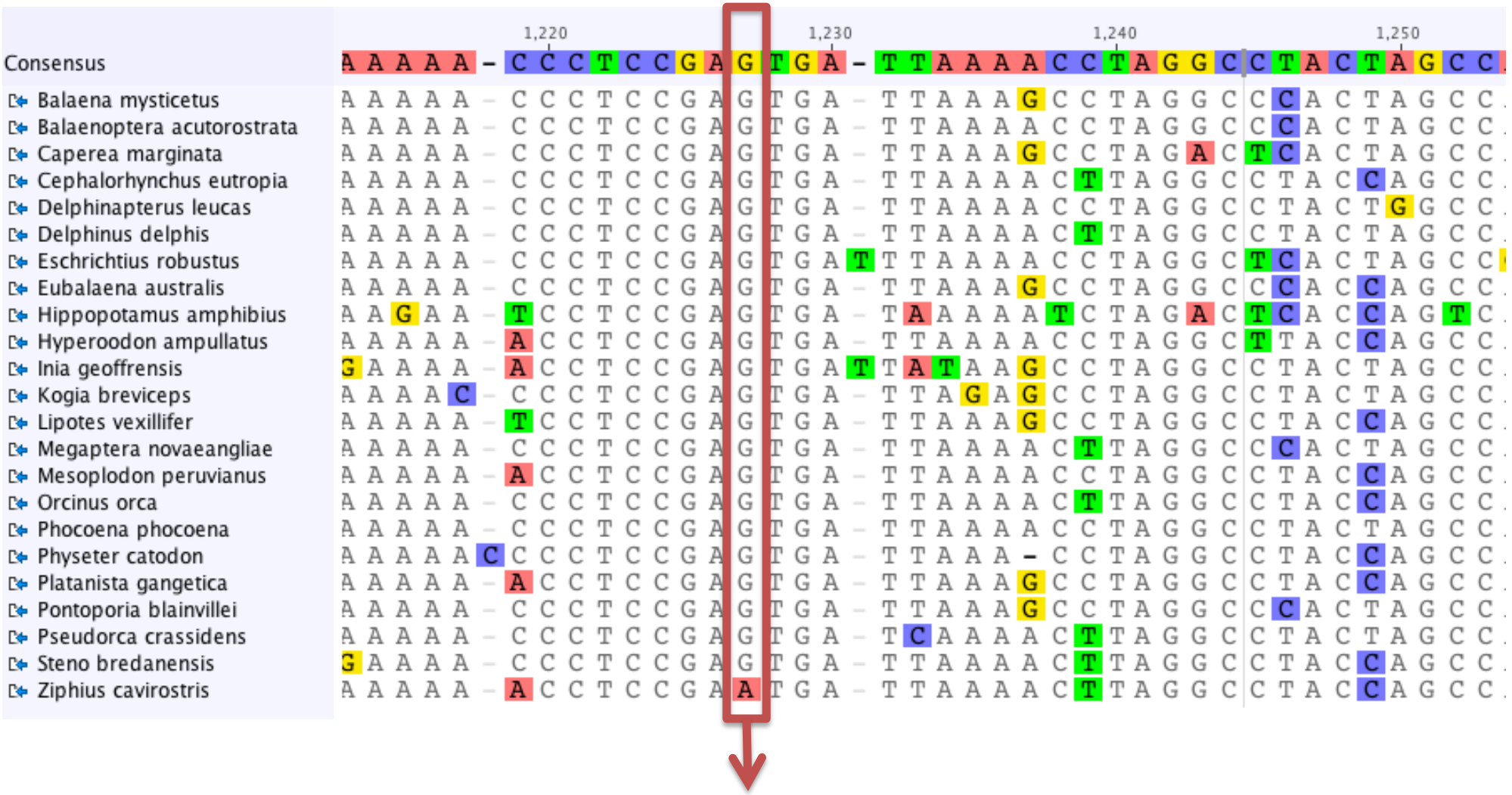
Where do turtles fit in the tree of life?

- Sequence the same gene from all taxa
- Align the gene sequences to each other
- Estimate a phylogeny



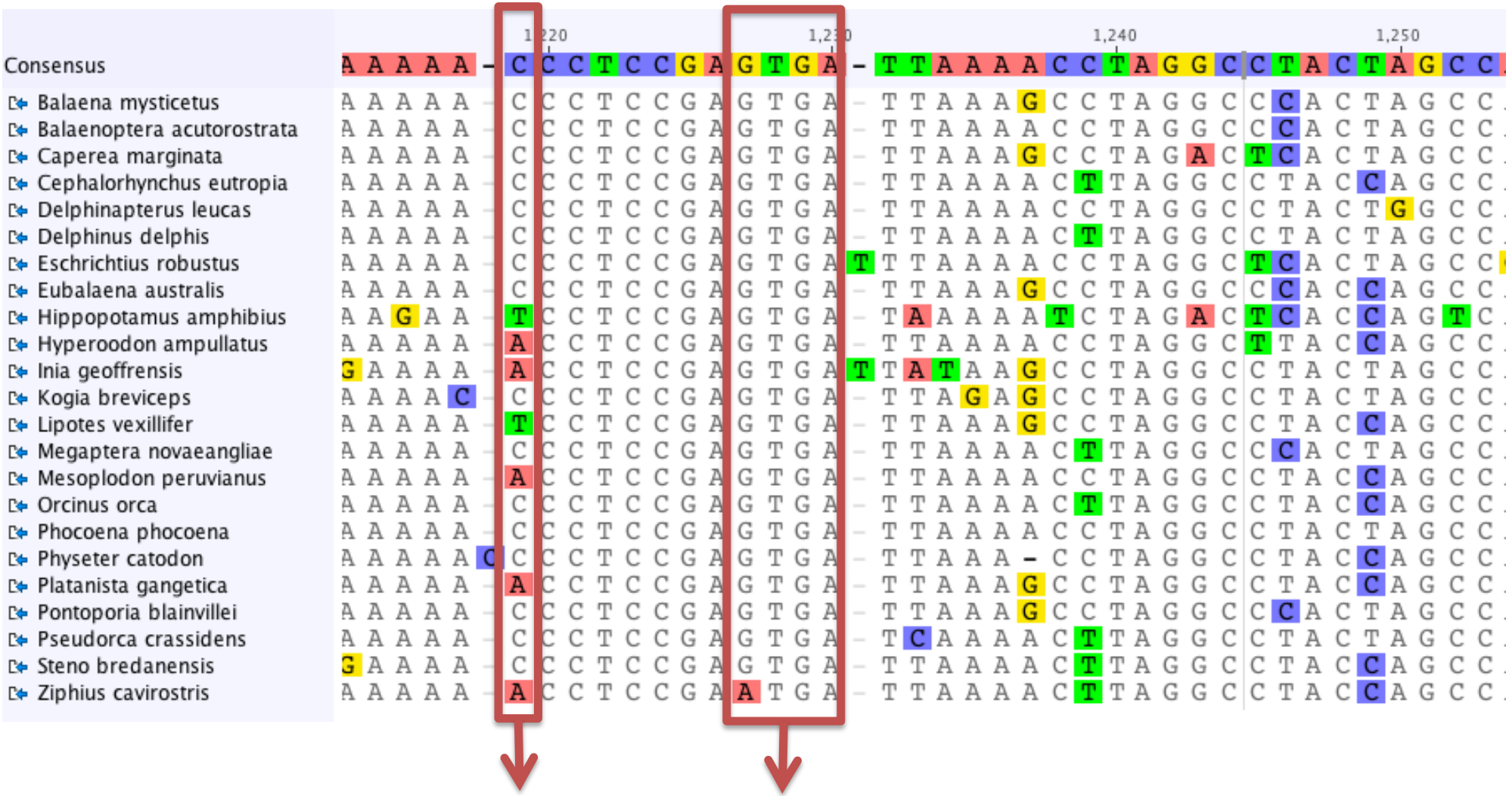

1. Bird-chicken
2. Bird-zebra_finch
3. Crocodile-Caiman
4. Crocodile-alligator
5. Turgle-loggerhead
6. Turtle-bearded_toadhead
7. Turtle-galapagos_tortoise
8. Human
9. Opussum
10. Platypus
11. Frog
12. Wall_lizard
13. Lizard
14. Lungfish
1
50
72
100
150
200
250
300
350
400
450
500
550
570





A homologous site
(one that was inherited from a common ancestor of all the sequences in the alignment)

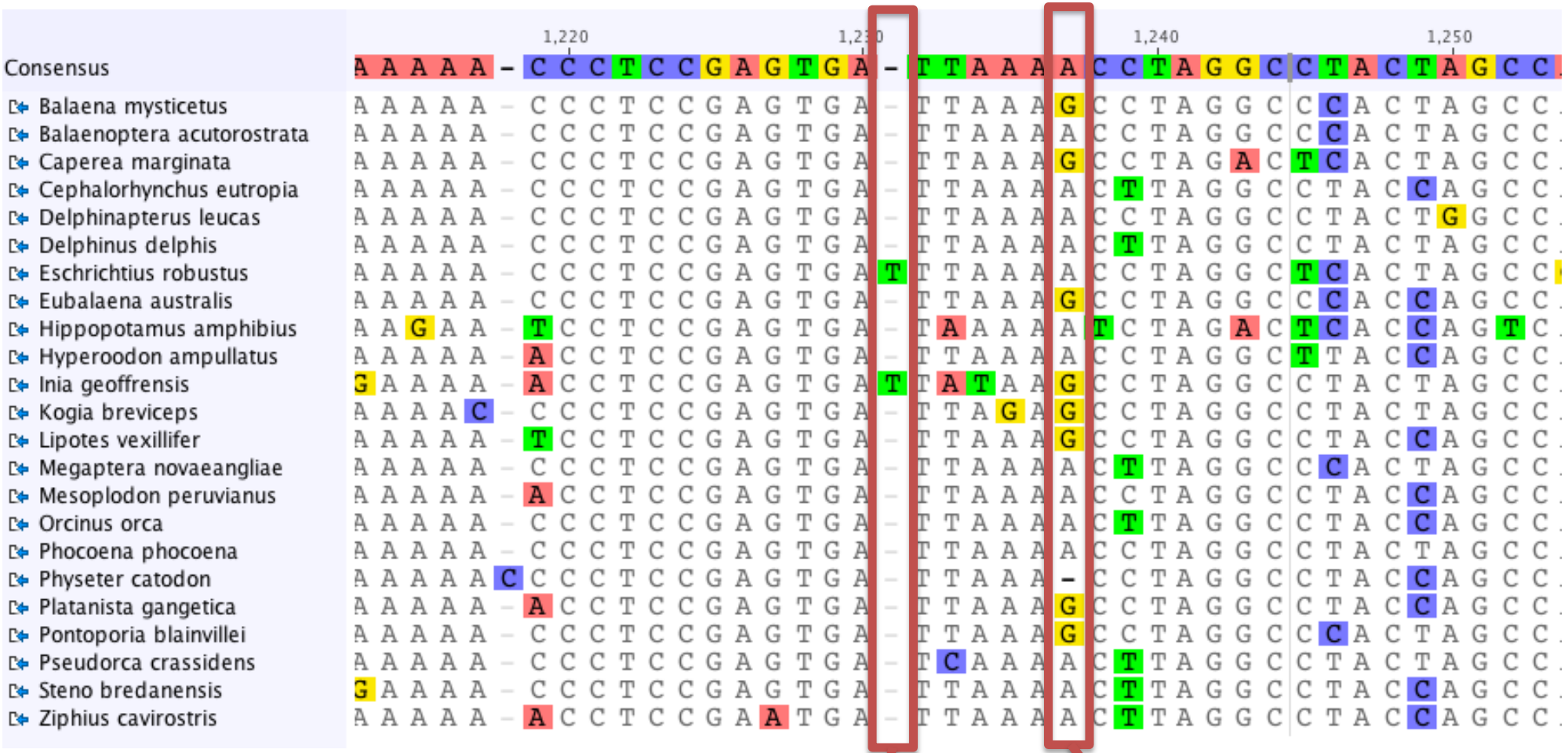
The point of alignment is to maximise the number of sites (or columns) for which you can confidently infer homology



Informative for topology and amount of evolution

Not informative for inferring the tree topology
Informative for inferring amount of evolution (i.e. branch lengths, times, evolutionary rates, substitution model parameters)

Both of these kinds of sites are useful in phylogenetics



Probably a deletion

An Indel

(A site where an insertion or a deletion has happened)

Indels are informative, but most ML and Bayesian phylogenetics don't use them sensibly

Spp1	A		Spp1	A		Spp1	A		Spp1	A		Spp1	A
Spp2	A		Spp2	A		Spp2	A		Spp2	A		Spp2	A
Spp3	A	=	Spp3	A	+	Spp3	A	+	Spp3	A	+	Spp3	A
Spp4	–		Spp4	T		Spp4	C		Spp4	G		Spp4	A
Spp5	A		Spp5	A		Spp5	A		Spp5	A		Spp5	A

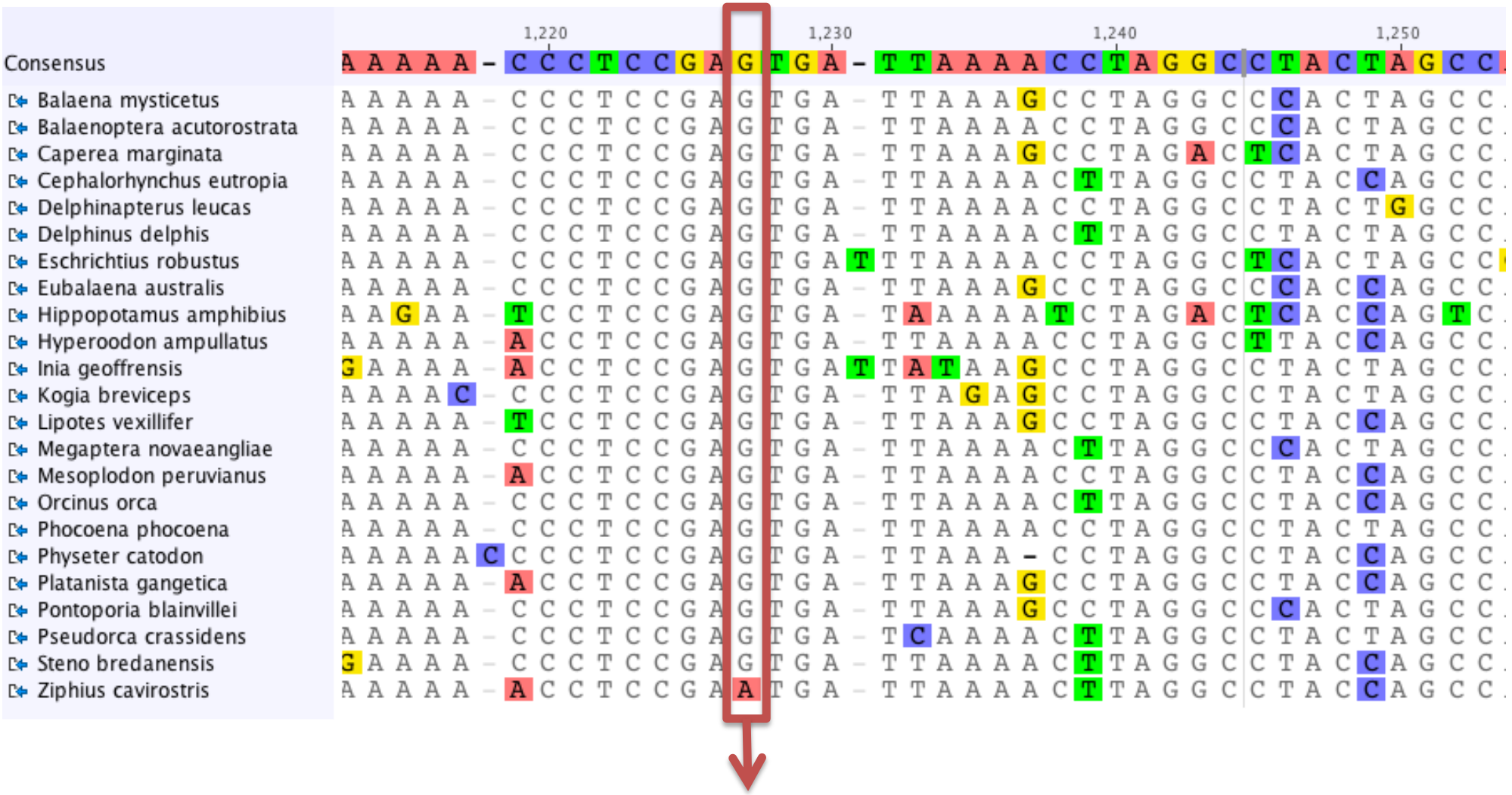
$$\ln L_{\_} = \ln L_T + \ln L_C + \ln L_G + \ln L_A$$

Spp1	A		Spp1	A
Spp2	A		Spp2	A
Spp3	A	=	Spp3	A
Spp4	–		Spp4	N
Spp5	A		Spp5	A

$$\ln L_{\_} = \ln L_N$$

Most ML and Bayesian methods treat indels as missing data (N's)

Indels are informative, but most ML and Bayesian phylogenetics don't use them sensibly



A homologous site

(one that was inherited from a common ancestor of all the species in the alignment

The point of alignment is to maximise the number of sites (or columns) for which you can confidently infer homology

(while bearing in mind the limitations of the downstream methods…)

A practical approach to alignment

1. Align sequences with automated methods (e.g. MAFFT)
2. Check alignments by eye (e.g. Geneious)
3. Fix major errors
    - Remove alignments without useful information
    - Remove sequences with uncertain homology
4. Go back to step 1 until you find no more major errors
5. Optional – fix minor errors by hand
    - Can depend on the question and amount of data
    - E.g. realign certain regions
    - E.g. delete poorly aligned columns
6. Use automated methods to clean final alignments

Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D.Thompson, Desmond G.Higgins+ and Toby J.Gibson*
European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

BMC Bioinformatics

BioMed Central

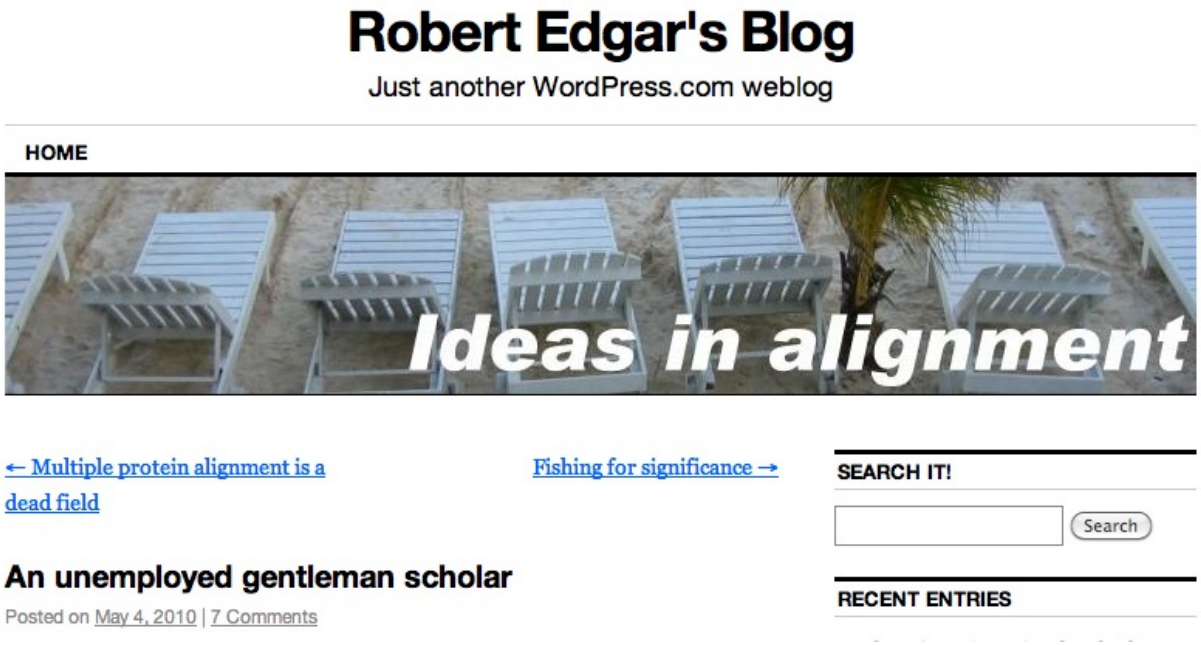
Software

Open Access

MUSCLE: a multiple sequence alignment method with reduced time and space complexity

Robert C Edgar*

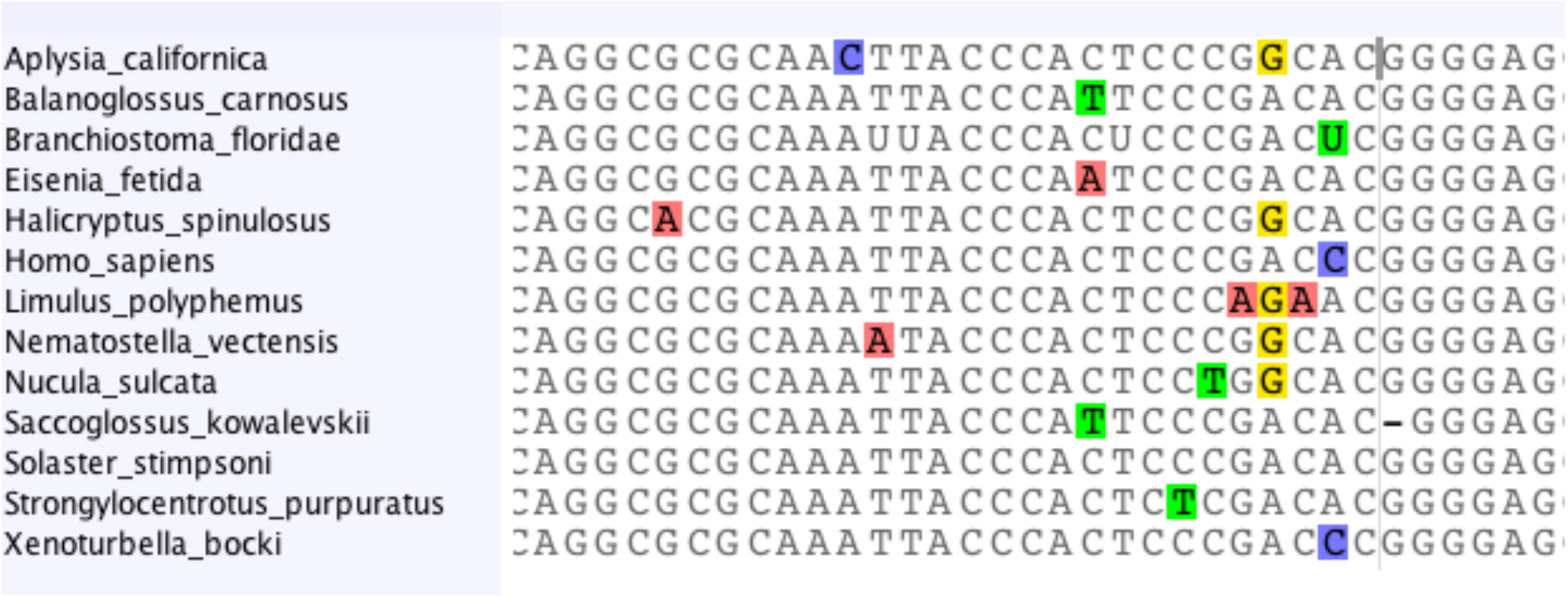
ALWAYS check and edit your alignments by eye!



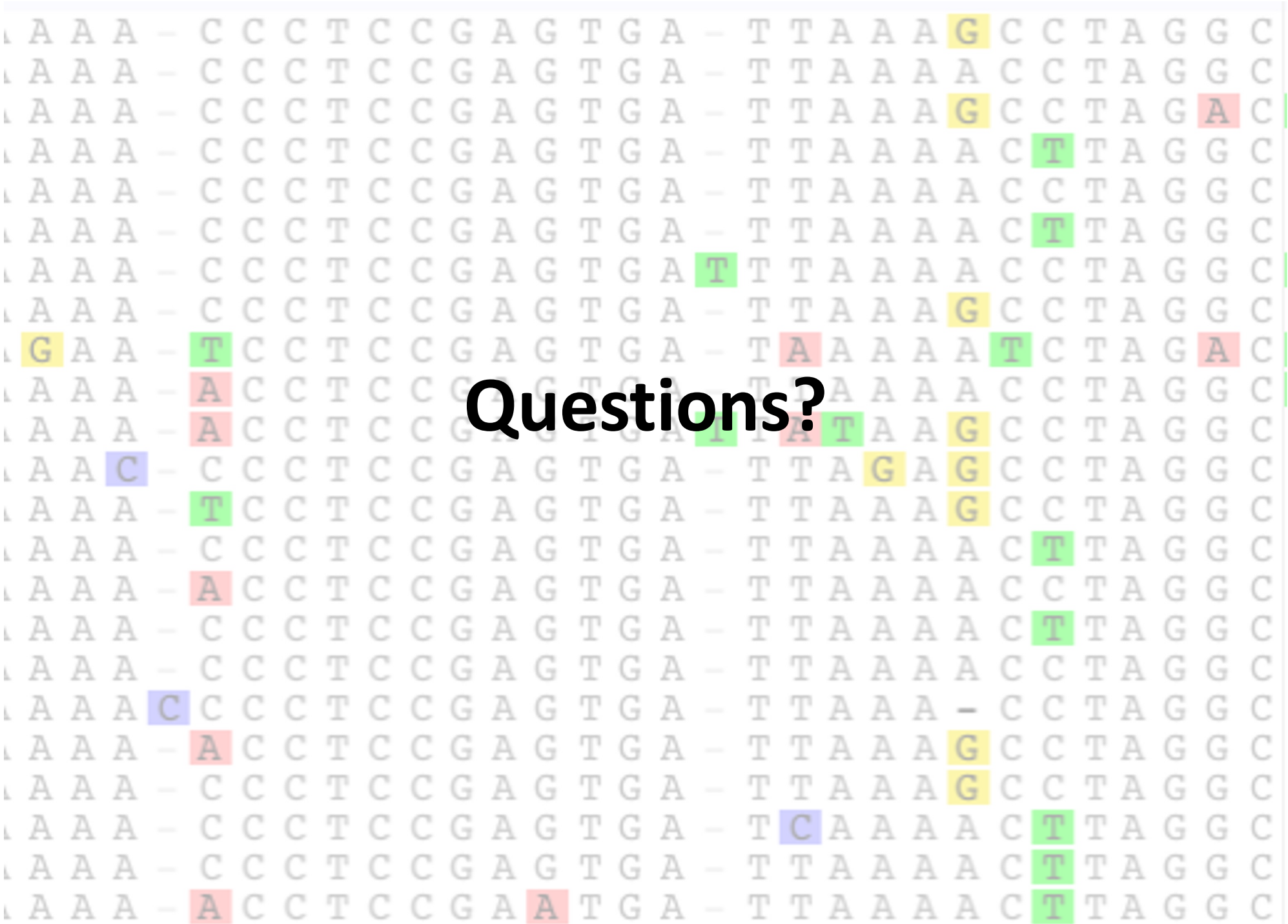
The key question is, **can an [automated] multiple alignment represent homology between letters accurately enough to enable robust inferences to be made by downstream tools?** …

The answer is, **probably not**…

The Aim



A collection of alignments for which you are confident that every column is homologous


Questions?

Download the workshop practicals here

https://github.com/roblanf/Workshop-MIG/releases/latest

Unzip, then open up the HTML file:

1_Alignment/Alignment.html