Let's have a drink now!

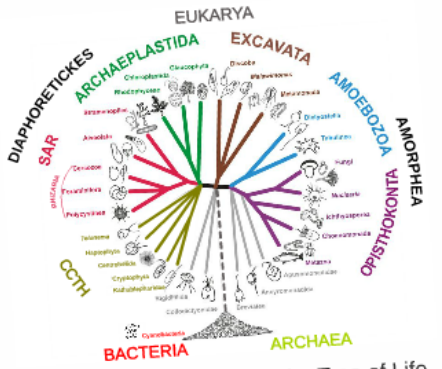
HTC Bayesian Phylogenomics on the Zenobe Tier-1 supercomputer

Denis BAURAIN

InBioS / Dept of Life Sciences
University of Liège

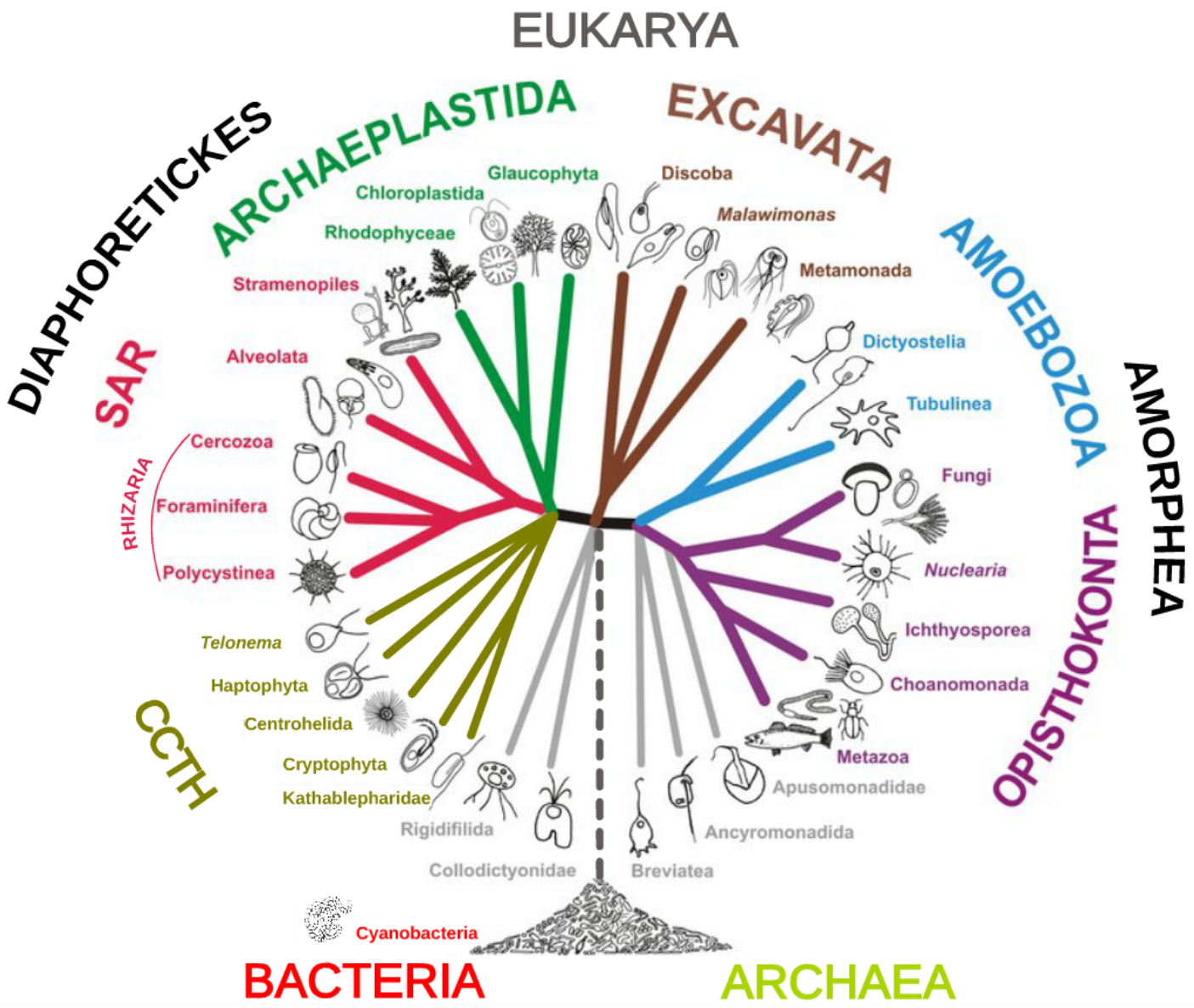
CÉCI scientific day / Apr 22, 2016

Background



We are trying to reconstruct the Tree of Life.

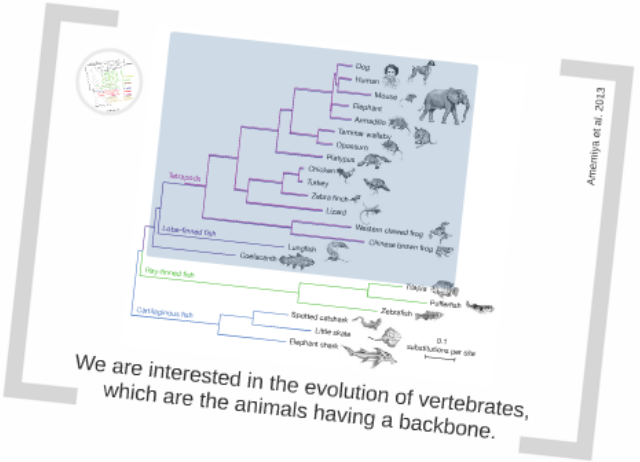
What is the general problem people in your field of research are trying to solve?



adapted from Adl et al. 2012 (MVV)

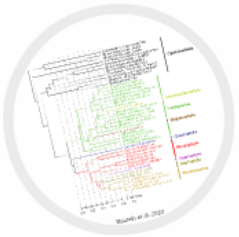
We are trying to reconstruct the Tree of Life.

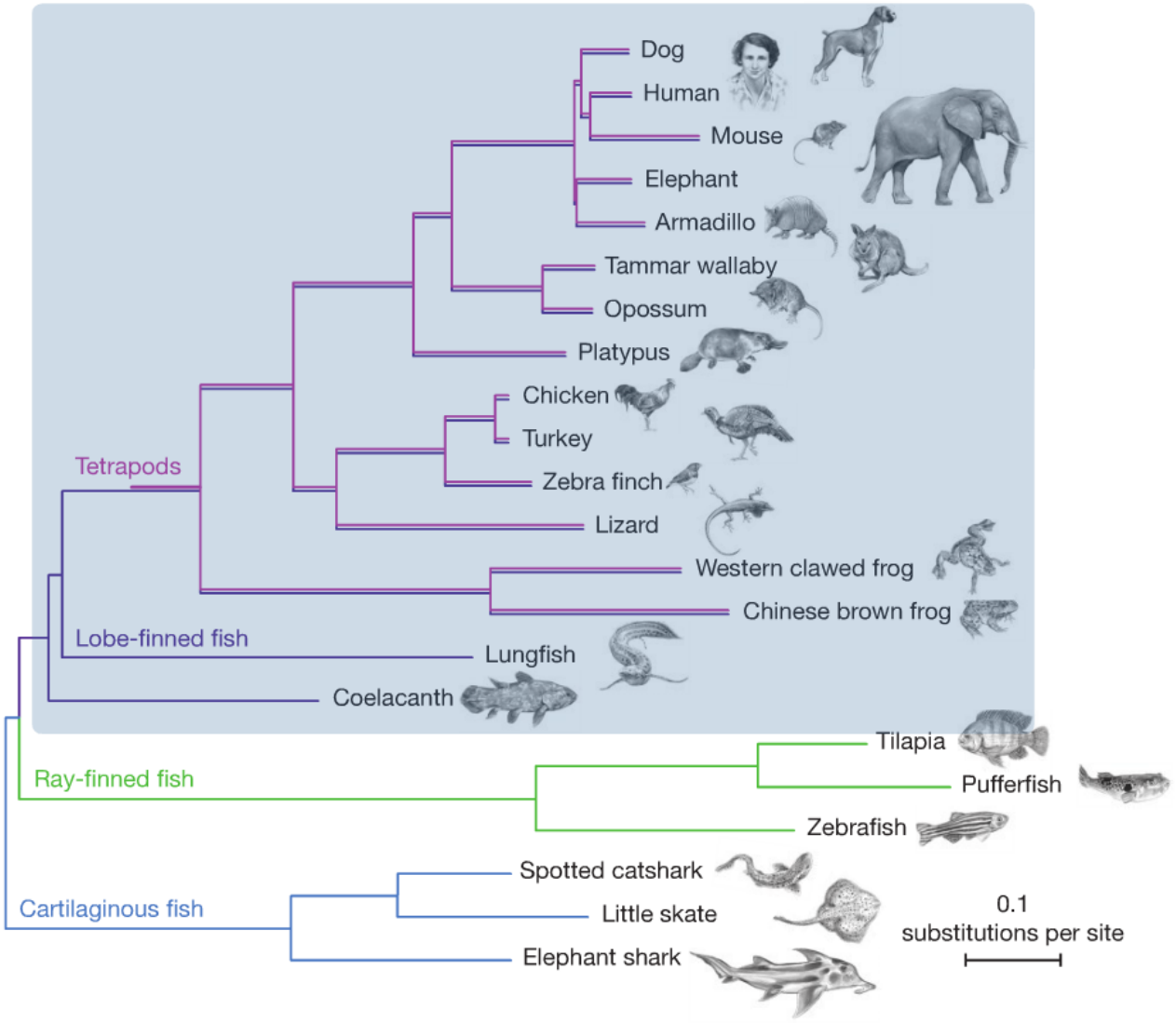
Objectives



We are interested in the evolution of vertebrates, which are the animals having a backbone.

What is the specific problem in that context that you are trying to solve and why does that matter?

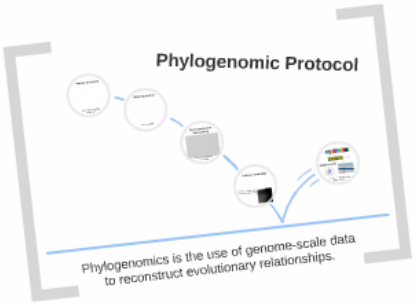


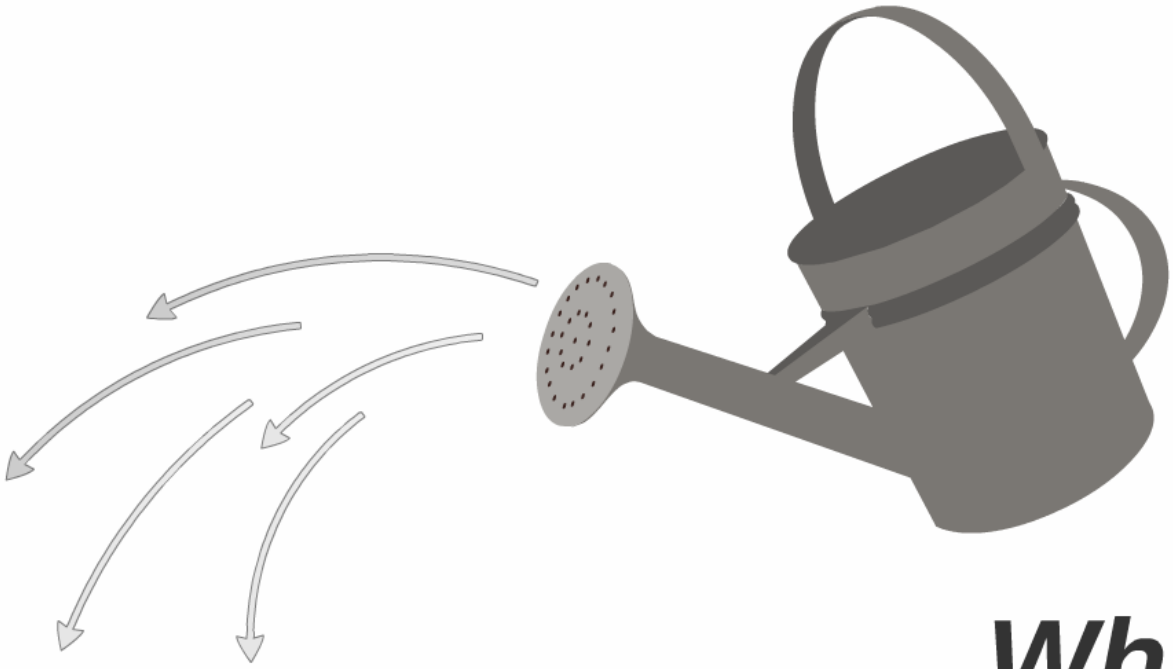


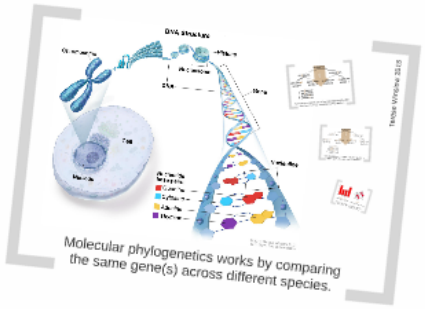
Amemiya et al. 2013

We are interested in the evolution of vertebrates, which are the animals having a backbone.

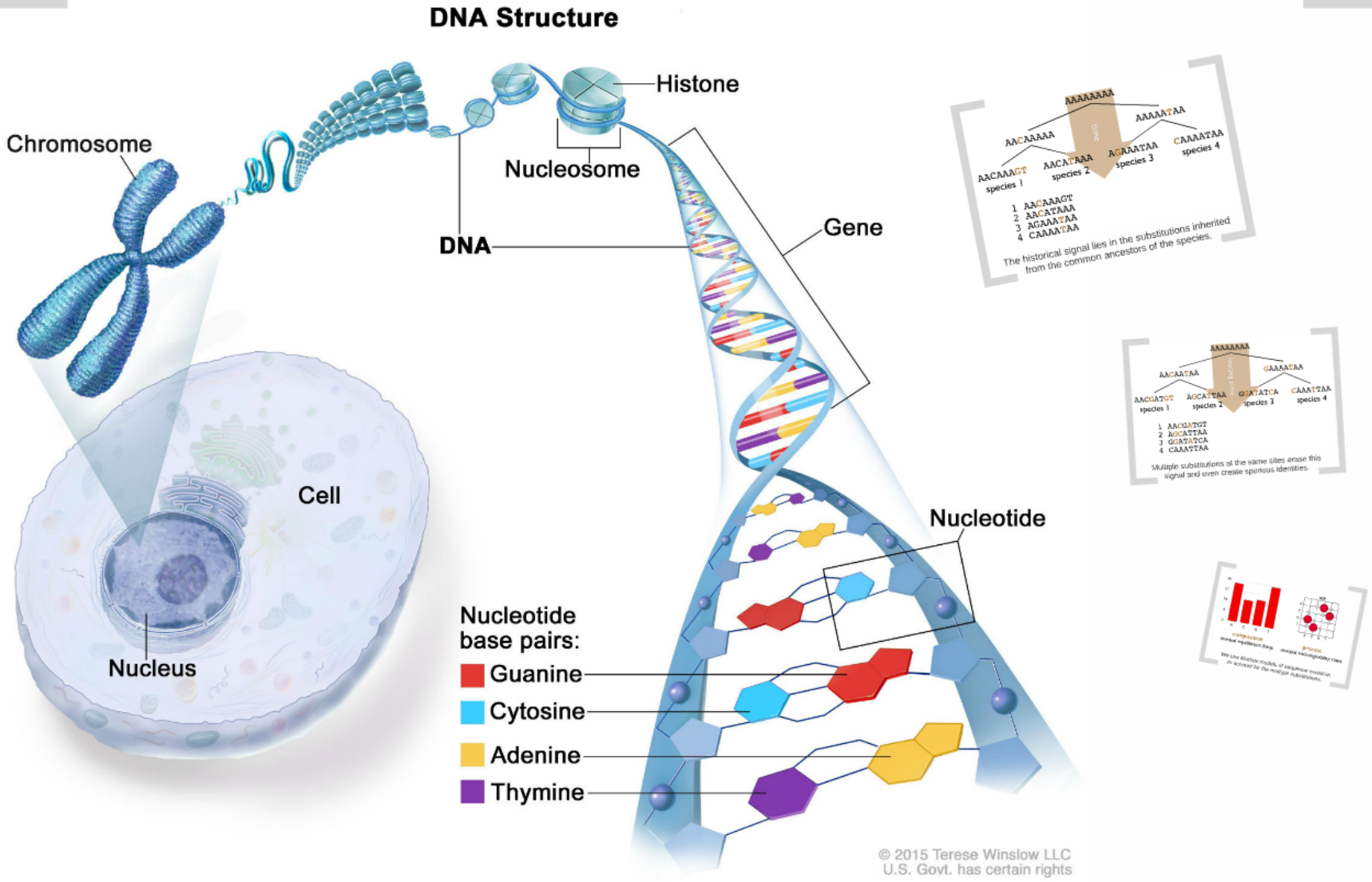
Methods





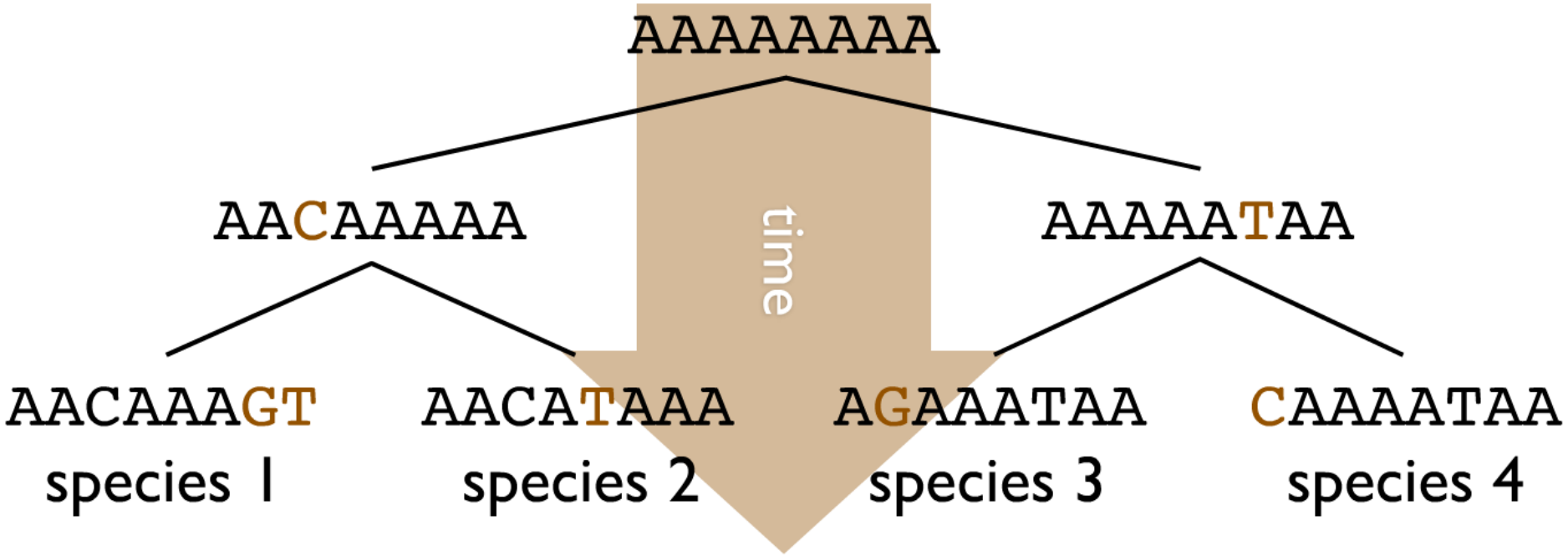


What strategy are you implementing to solve that problem?



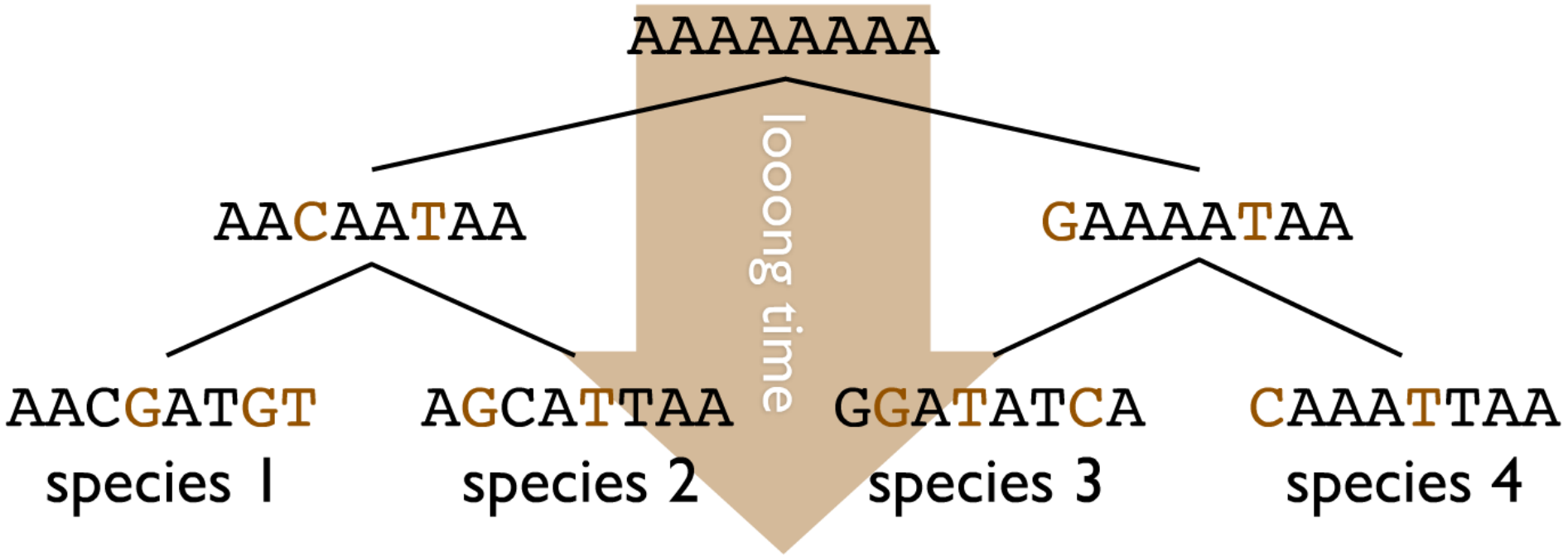
Terese Winslow 2015

Molecular phylogenetics works by comparing the same gene(s) across different species.



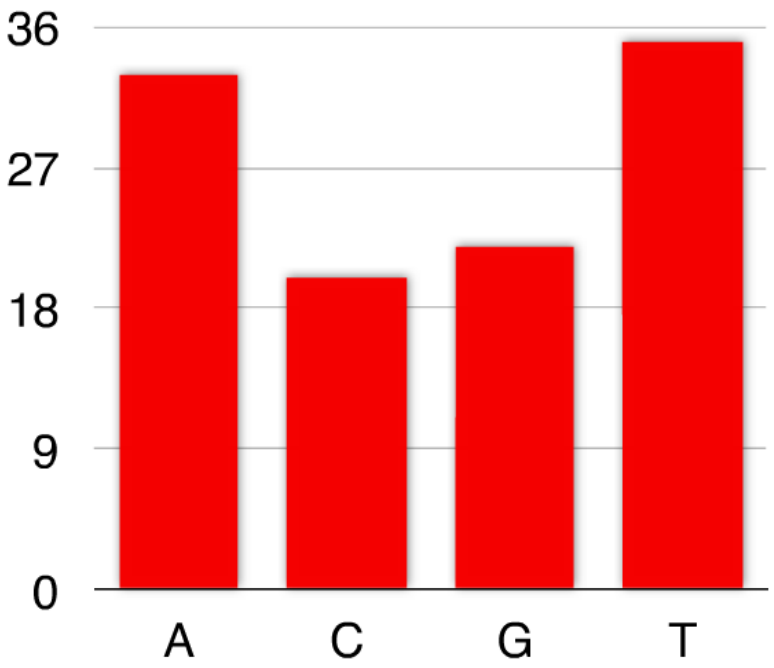
1 AACAAAGT
2 AACATAAA
3 AGAAATAA
4 CAAAATAA

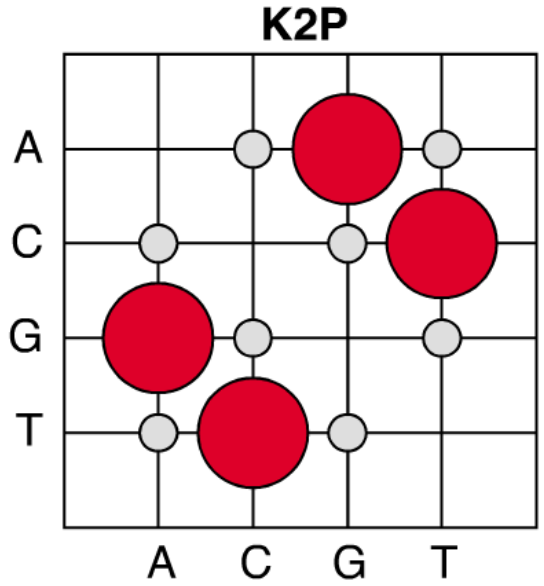
The historical signal lies in the substitutions inherited from the common ancestors of the species.



1 AACGATGT
2 AGCATTAA
3 GGATATCA
4 CAAATTAA

Multiple substitutions at the same sites erase this signal and even create spurious identities.





composition

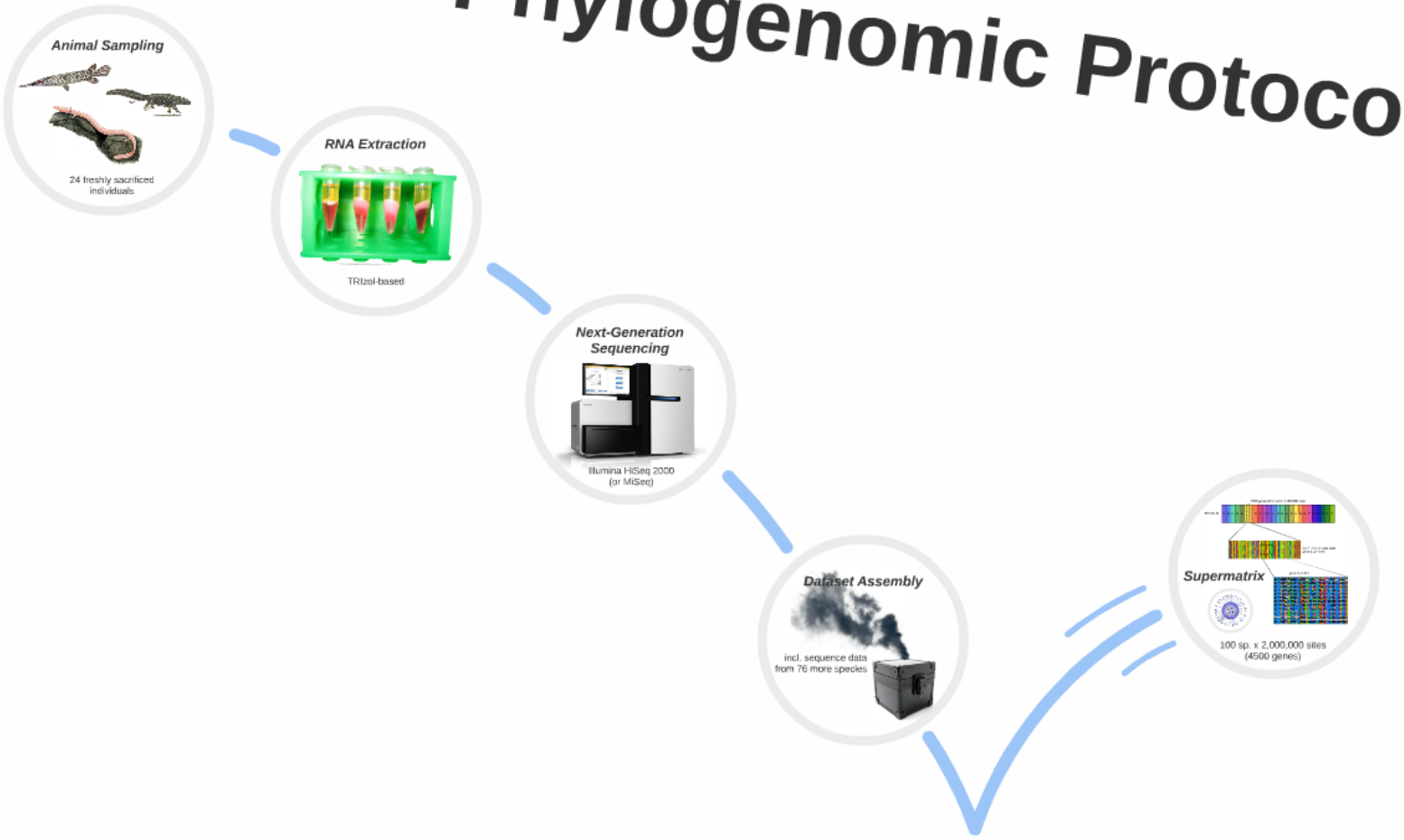
unequal equilibrium freqs

process

unequal exchangeability rates

We use Markov models of sequence evolution to account for the multiple substitutions.

Phylogenomic Protocol



Phylogenomics is the use of genome-scale data to reconstruct evolutionary relationships.

Animal Sampling

24 freshly sacrificed individuals

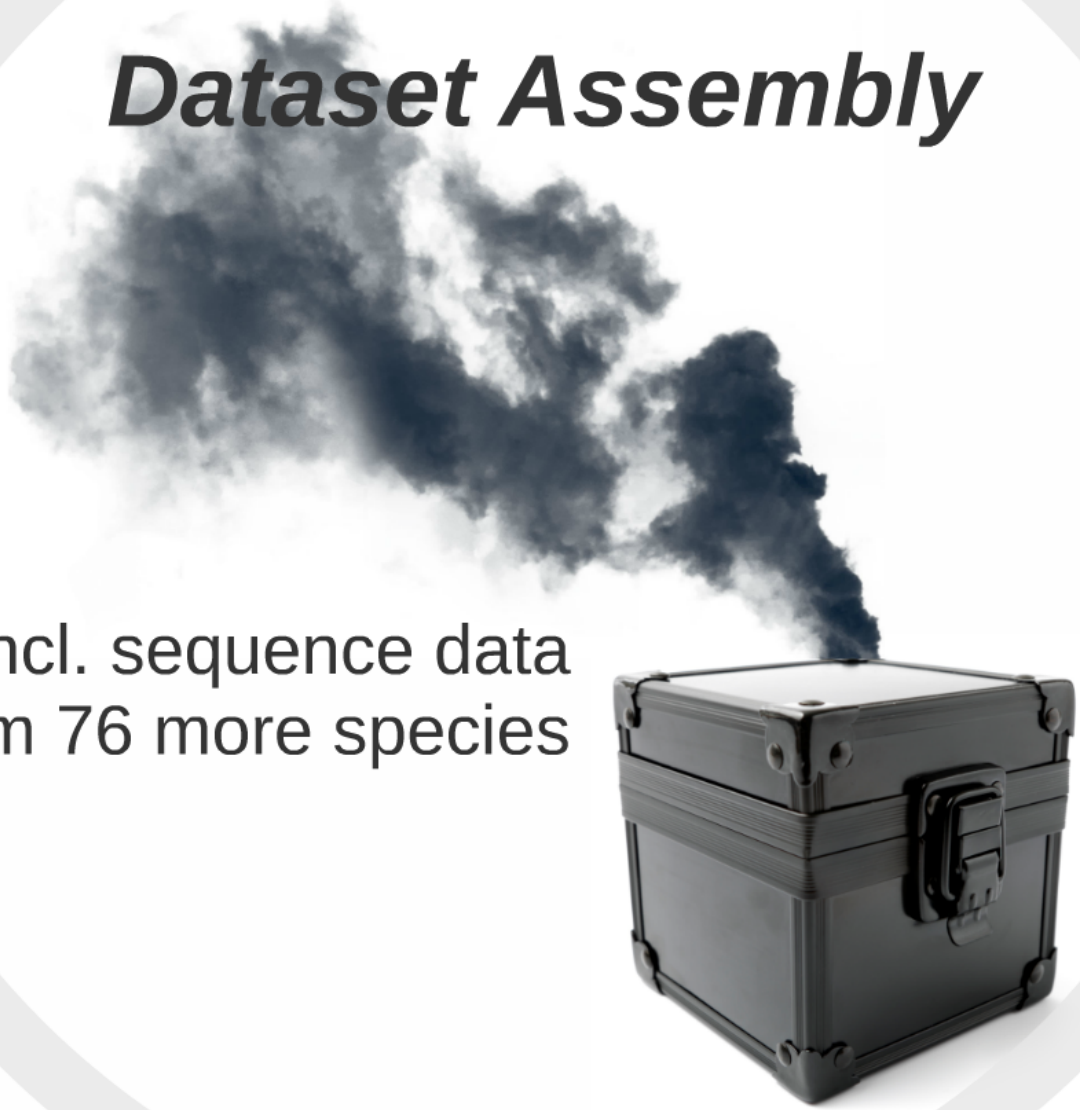
RNA Extraction

TRIzol-based

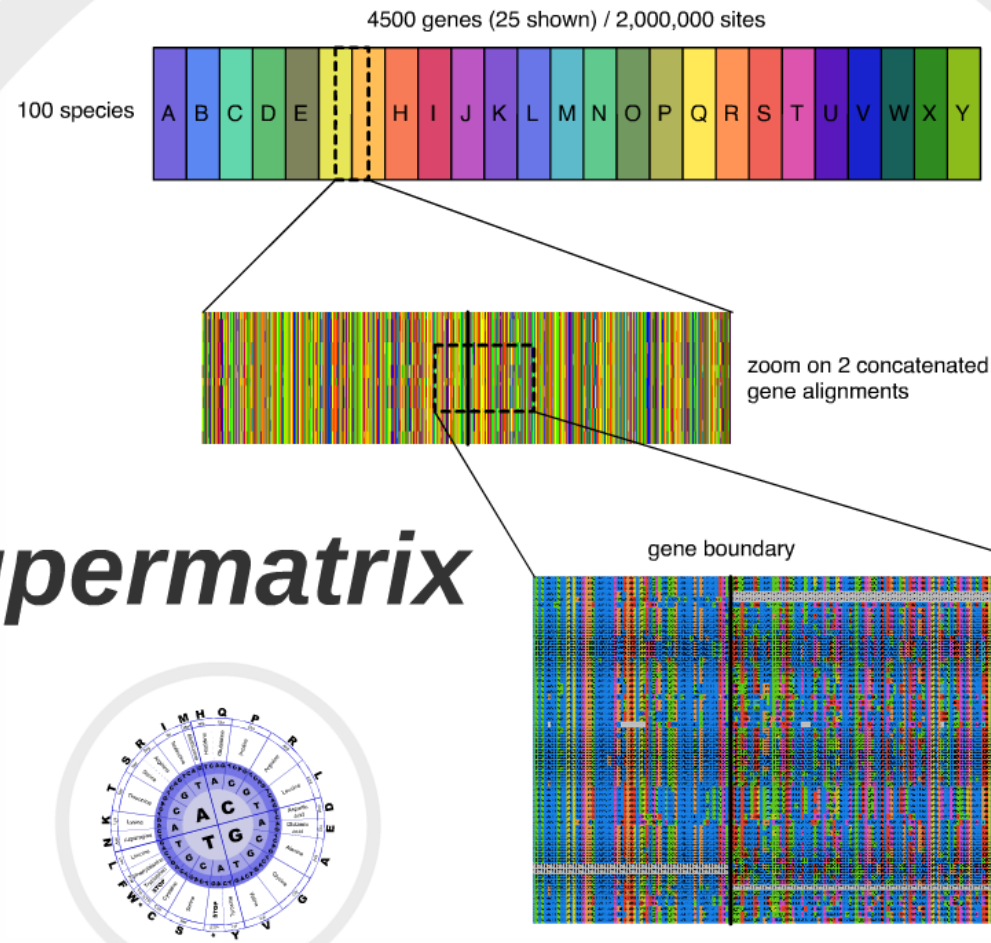
Next-Generation Sequencing

Illumina HiSeq 2000
(or MiSeq)

Dataset Assembly

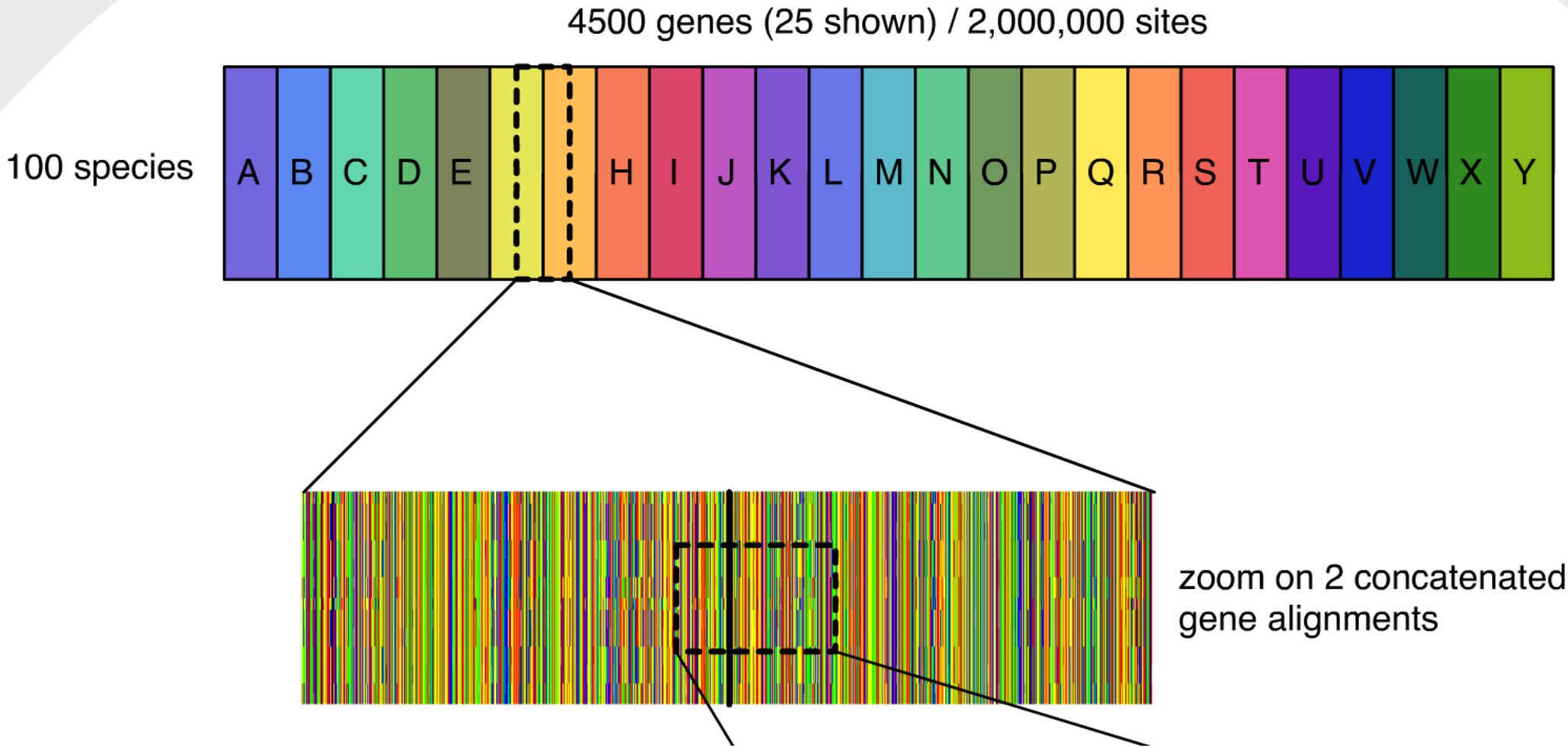


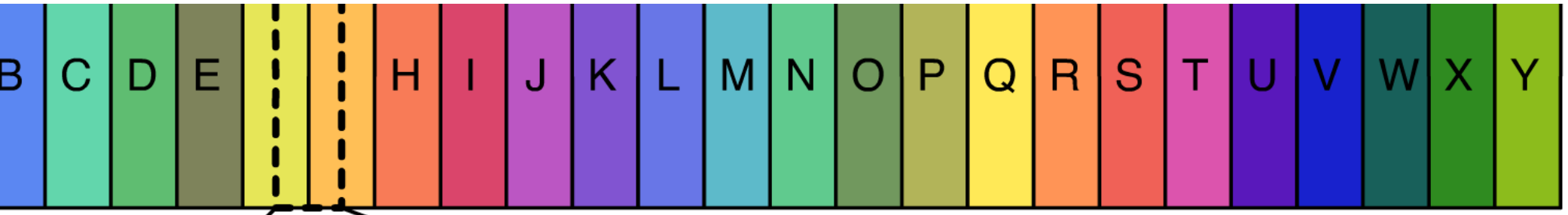
incl. sequence data
from 76 more species

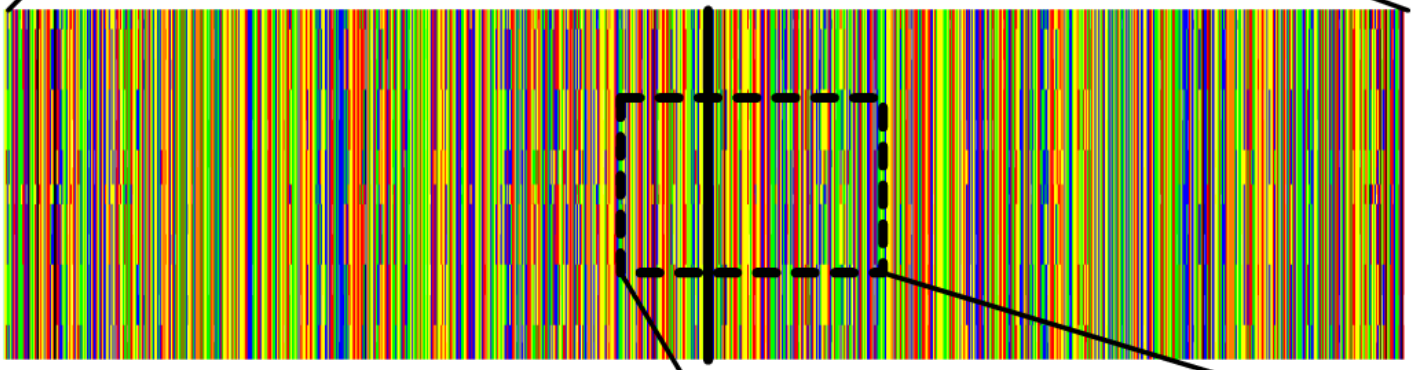


Supermatrix

100 sp. x 2,000,000 sites
(4500 genes)


4500 genes (25 shown) / 2,000,000 sites
100 species
A
B
C
D
E
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
zoom on 2 concatenated
gene alignments

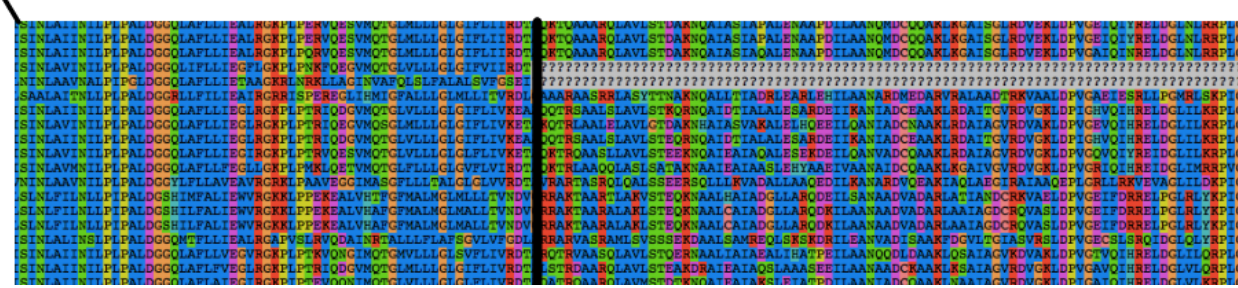


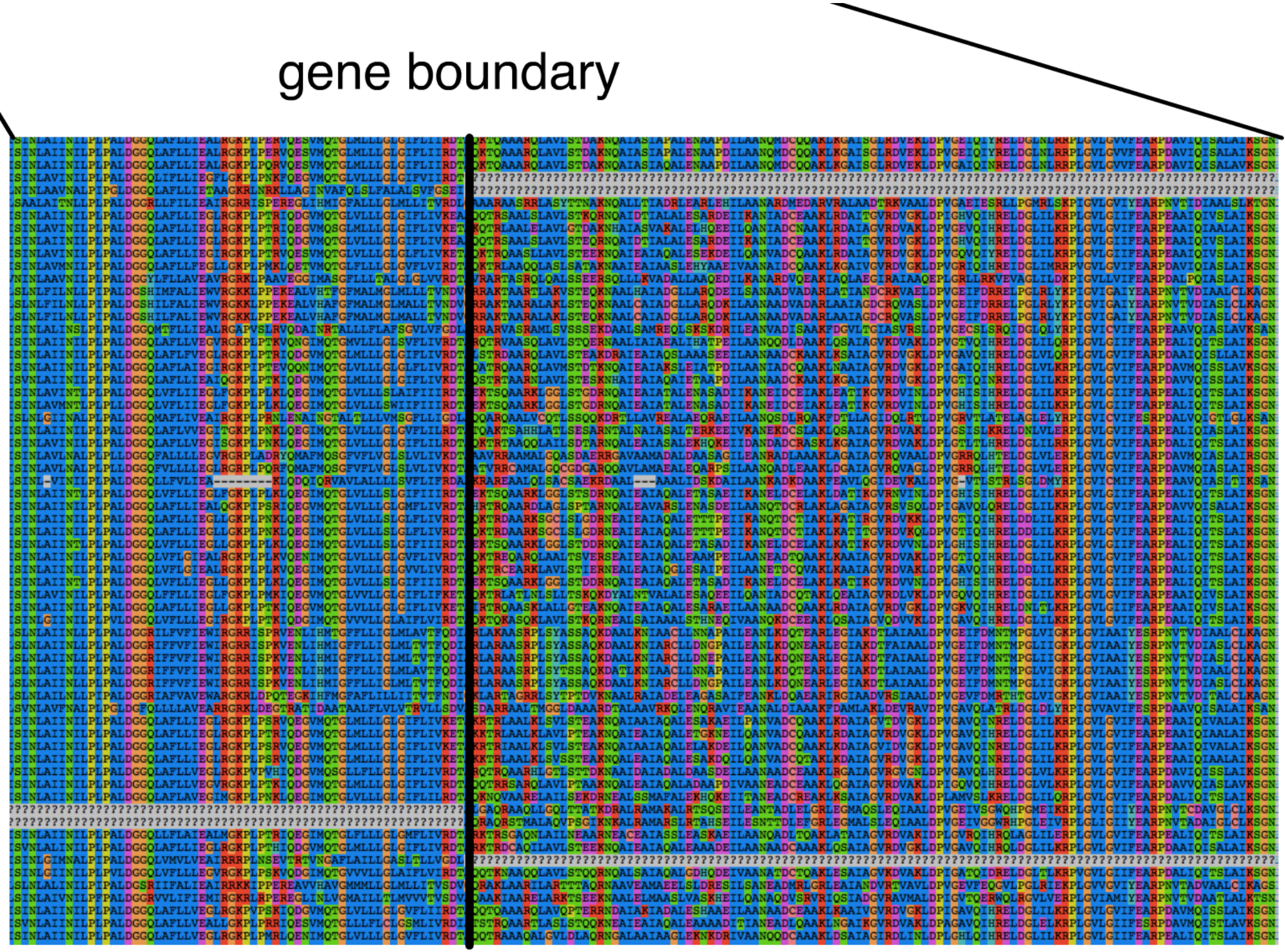


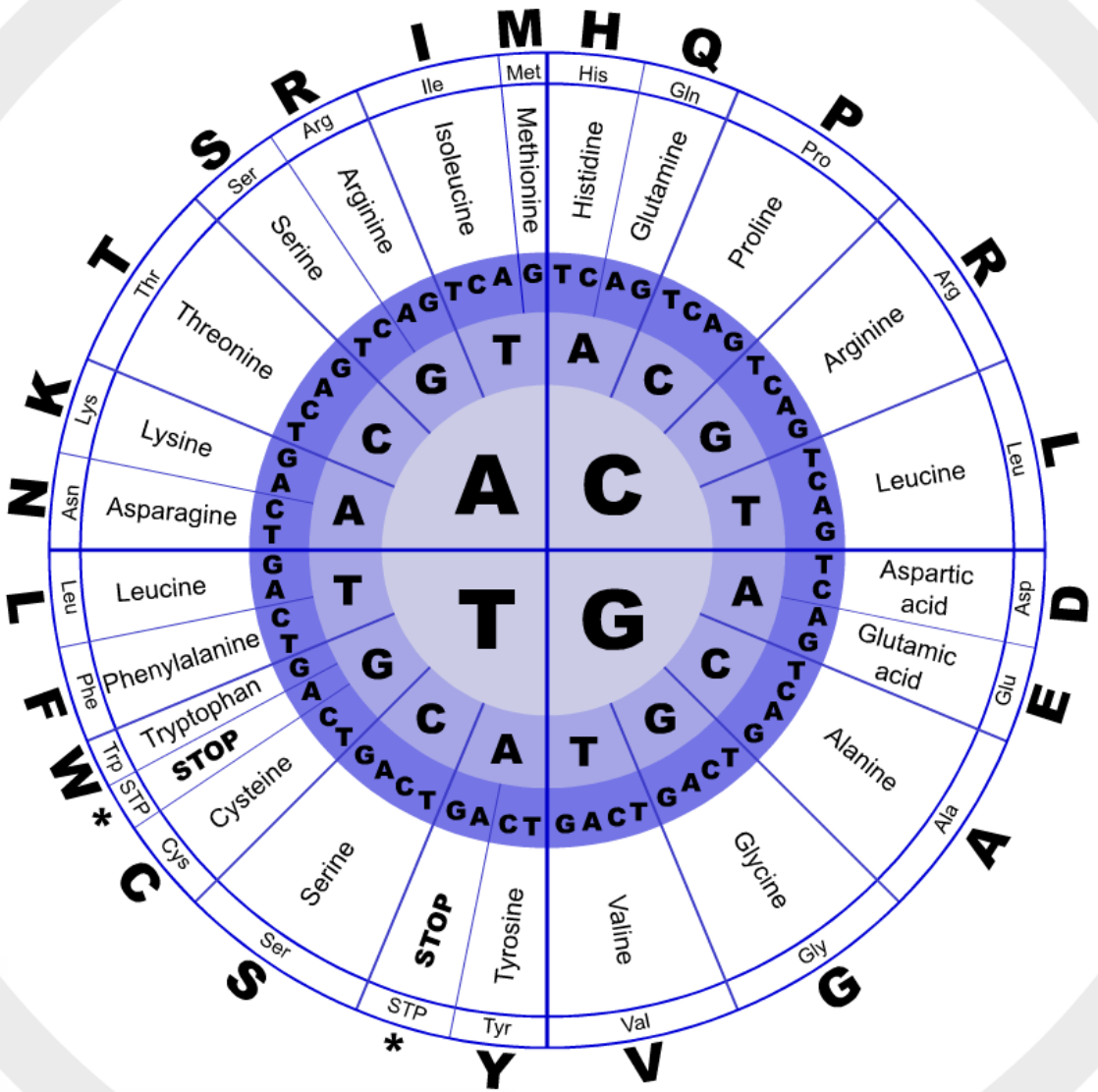
zoom on 2 concatenated gene alignments

matrix

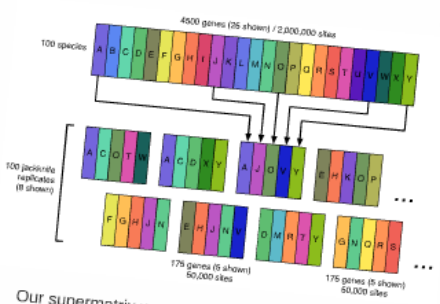
gene boundary




gene boundary


A
C
T
G
Methionine
Isoleucine
Arginine
Serine
Histidine
Glutamine
Proline
Arginine
Threonine
Lysine
Asparagine
Leucine
Phenylalanine
Tryptophan
STOP
Cysteine
Tyrosine
STOP
Serine
Valine
Glycine
Alanine
Aspartic acid
Glutamic acid
Leucine
Met
Ile
Arg
Ser
His
Gln
Pro
Arg
Thr
Lys
Asn
Leu
Phe
Trp
STP
Cys
Ser
STP
Tyr
Val
Gly
Ala
Glu
Asp
Leu

Methods

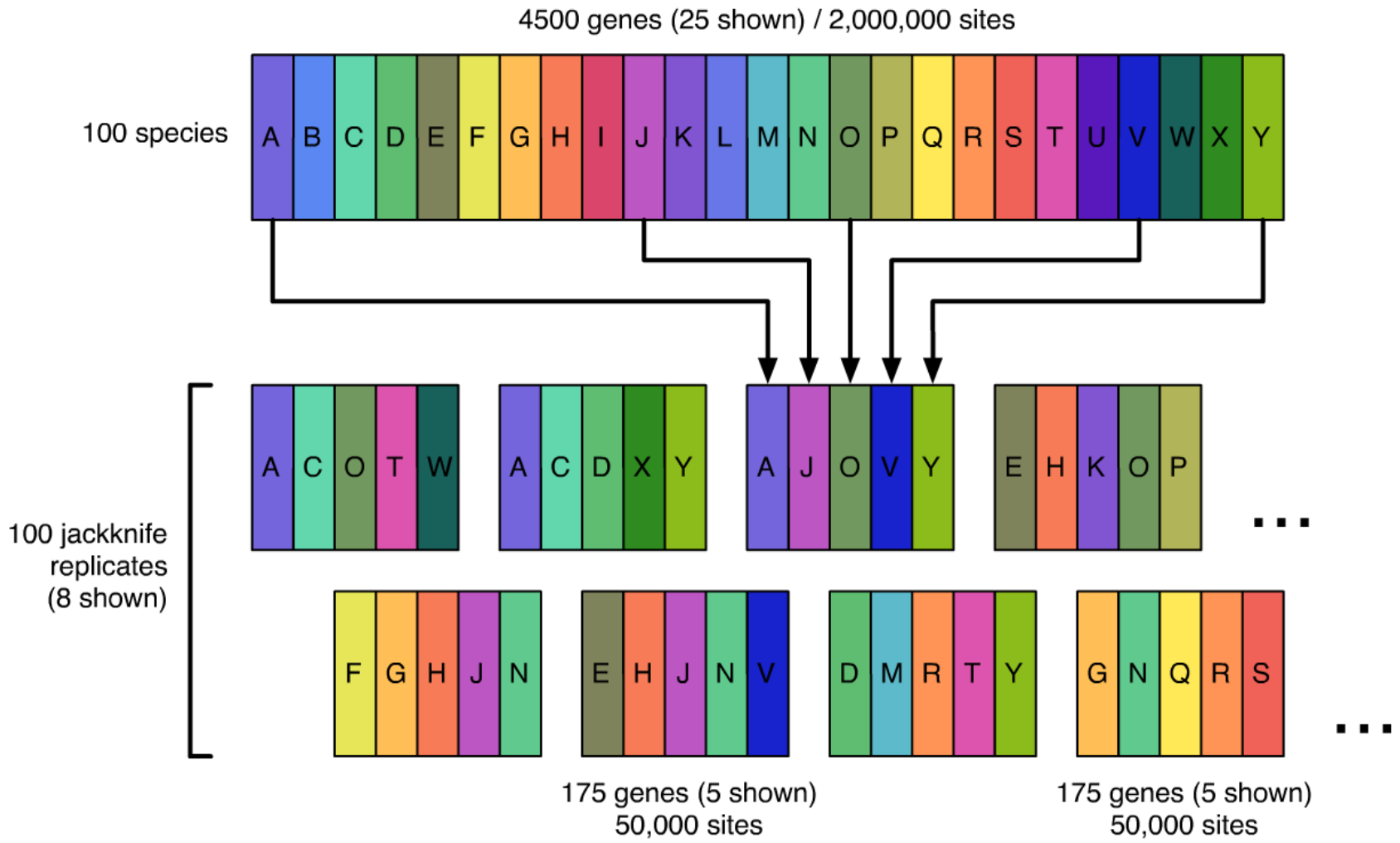


Our supermatrix was too large to be analysed "as is".
We thus used a gene resampling approach.

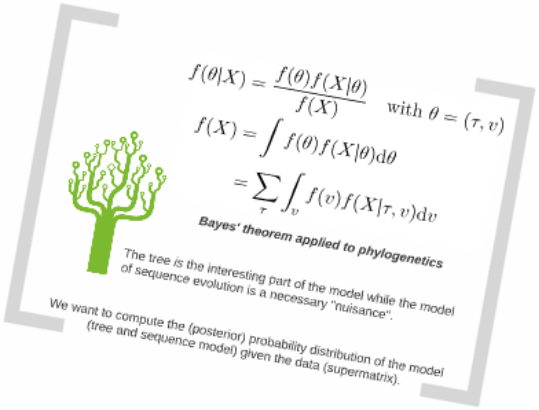
N. Lartillot

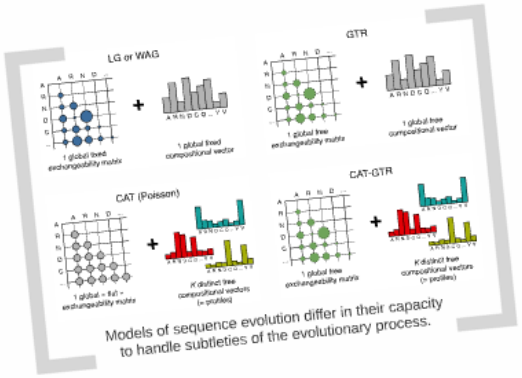
Each jackknife replicate was then analysed using a powerful Bayesian phylogenetic software: PhyloBayes MPI.

Where do computers fit in that strategy? How do you (ab)use them?

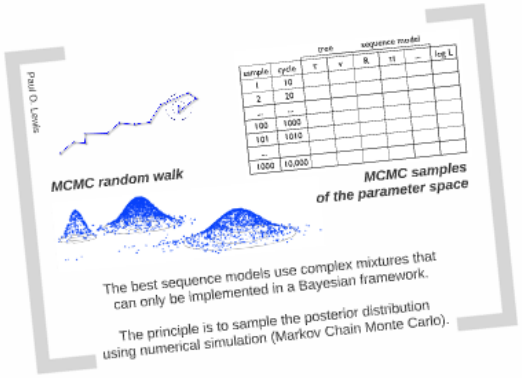


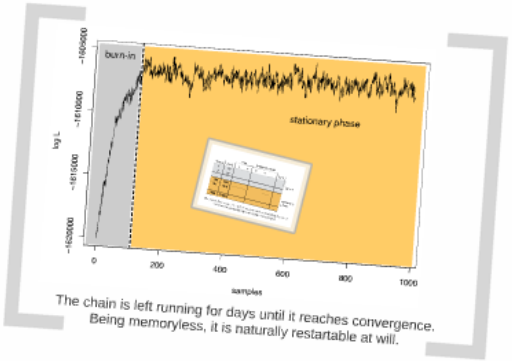
Our supermatrix was too large to be analysed "as is".
We thus used a gene resampling approach.

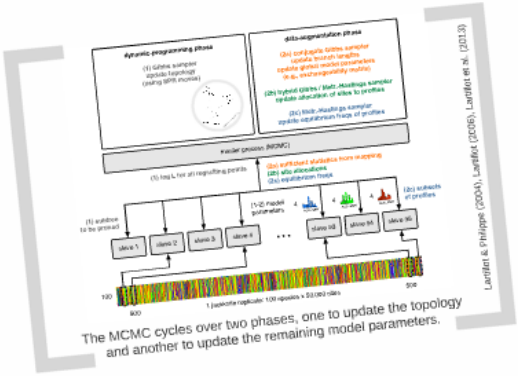




N. Lartillot







Each jackknife replicate was then analysed using a powerful Bayesian phylogenetic software: PhyloBayes MPI.

$$f(\theta|X) = \frac{f(\theta)f(X|\theta)}{f(X)} \quad \text{with } \theta = (\tau, v)$$

$$f(X) = \int f(\theta)f(X|\theta)\mathrm{d}\theta$$

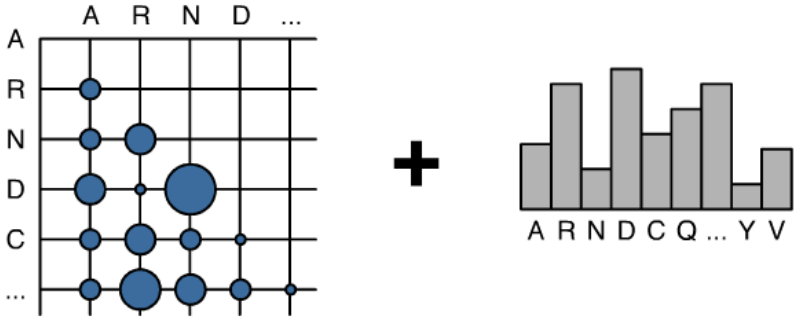
$$= \sum_{\tau} \int_{v} f(v)f(X|\tau, v)\mathrm{d}v$$

Bayes' theorem applied to phylogenetics

The tree *is* the interesting part of the model while the model of sequence evolution is a necessary "nuisance".

We want to compute the (posterior) probability distribution of the model (tree and sequence model) given the data (supermatrix).

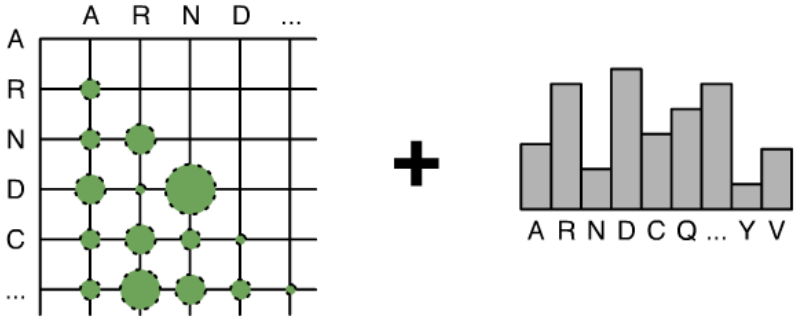
LG or WAG



1 global fixed exchangeability matrix

1 global fixed compositional vector

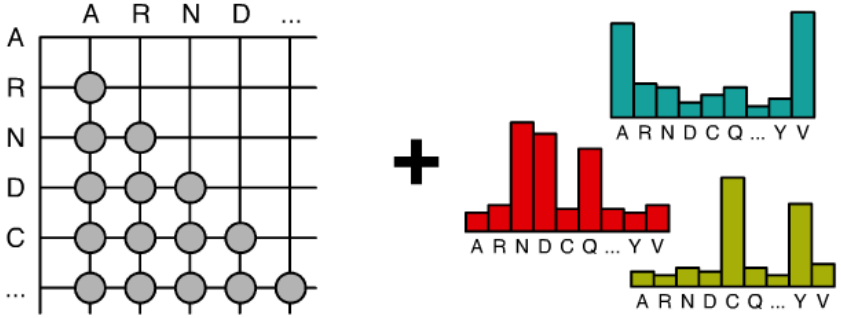
GTR



1 global free exchangeability matrix

1 global free compositional vector

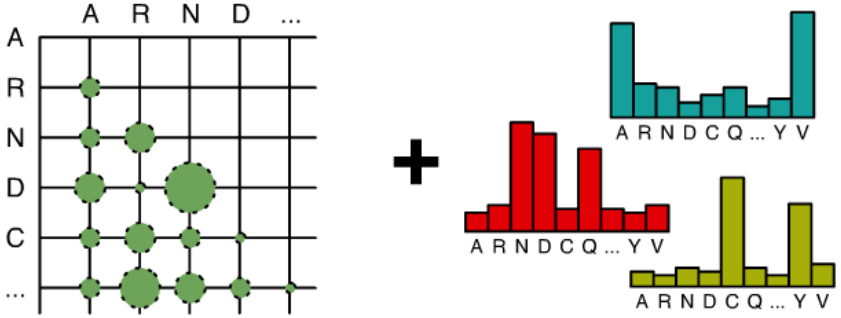
CAT (Poisson)



1 global « flat » exchangeability matrix

K distinct free compositional vectors (= profiles)

CAT-GTR

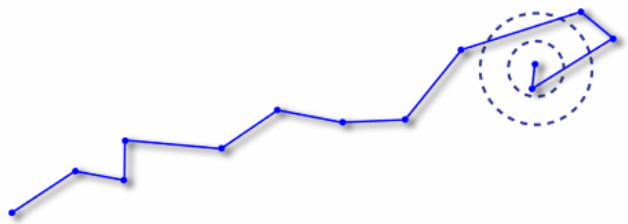


1 global free exchangeability matrix

K distinct free compositional vectors (= profiles)

Models of sequence evolution differ in their capacity to handle subtleties of the evolutionary process.

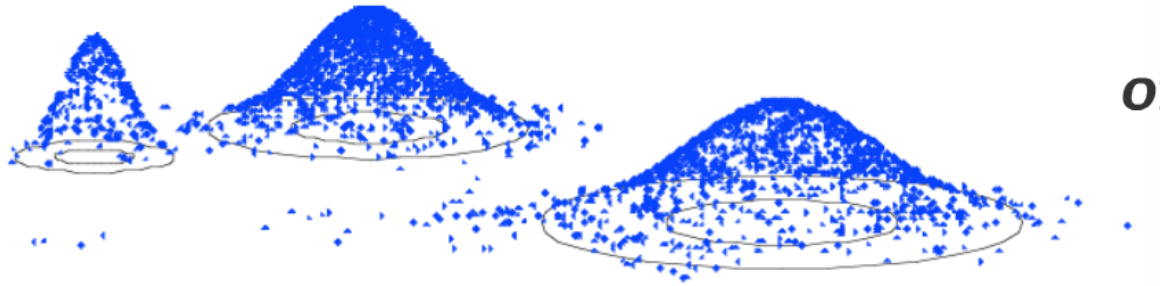
Paul O. Lewis



MCMC random walk

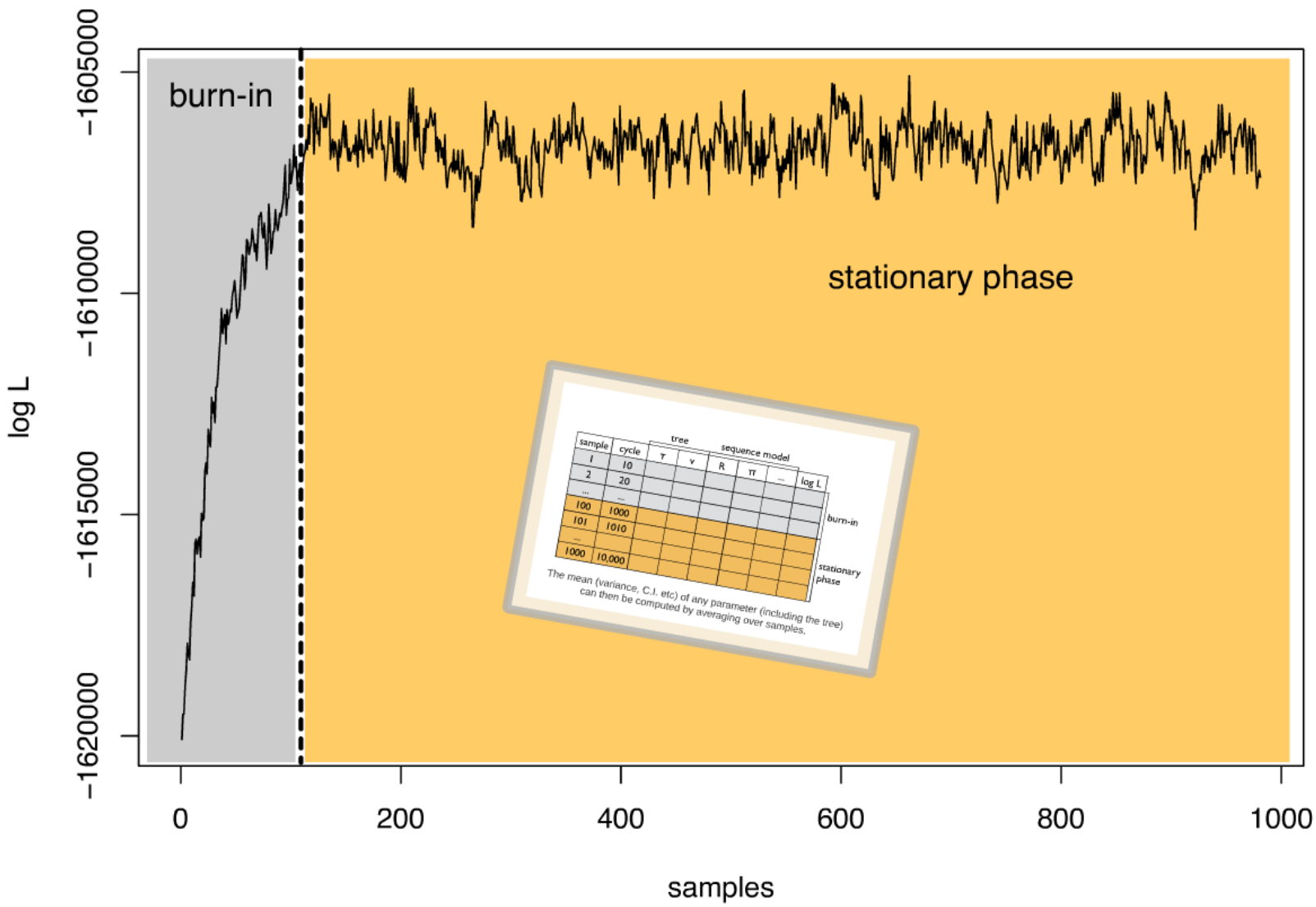
		tree		sequence model			
sample	cycle	τ	v	R	π	...	log L
1	10						
2	20						
...	...						
100	1000						
101	1010						
...							
1000	10,000						

MCMC samples of the parameter space



The best sequence models use complex mixtures that can only be implemented in a Bayesian framework.

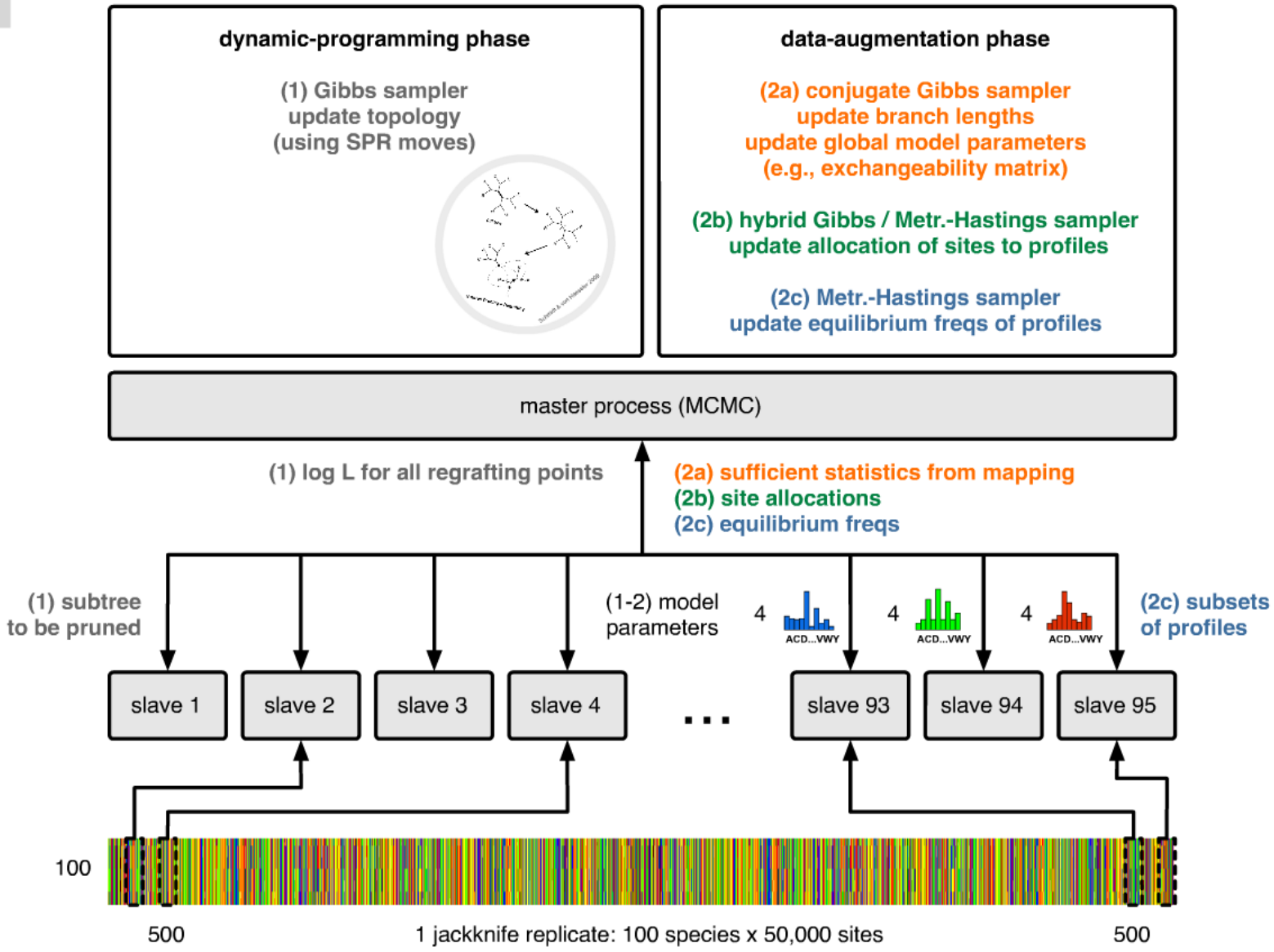
The principle is to sample the posterior distribution using numerical simulation (Markov Chain Monte Carlo).



The chain is left running for days until it reaches convergence.
Being memoryless, it is naturally restartable at will.

		tree		sequence model				
sample	cycle	τ	v	R	π	...	log L	
1	10							burn-in
2	20							
...	...							
100	1000							stationary phase
101	1010							
...								
1000	10,000							

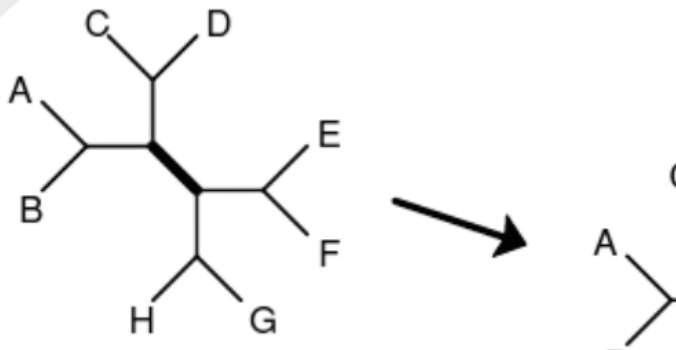
The mean (variance, C.I. etc) of any parameter (including the tree) can then be computed by averaging over samples.



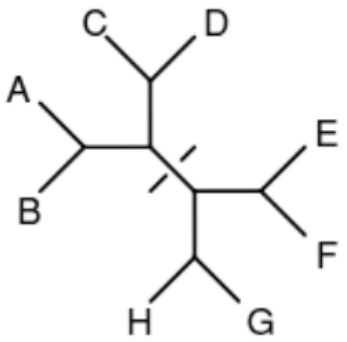
Lartillot & Philippe (2004), Lartillot (2006), Lartillot et al. (2013)

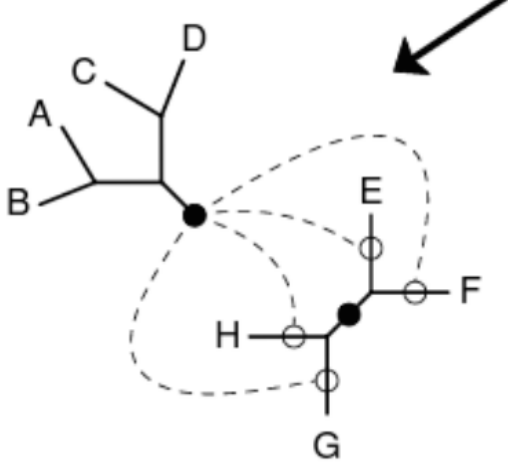
The MCMC cycles over two phases, one to update the topology and another to update the remaining model parameters.

oves)



full tree

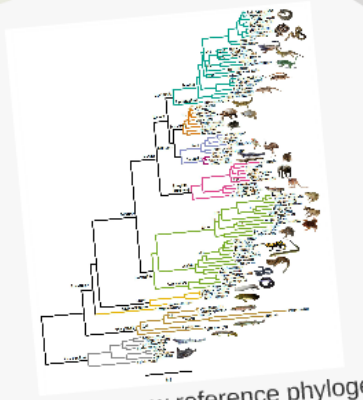




Subtree Pruning + Regrafting

Schmidt & von Haeseler 2009

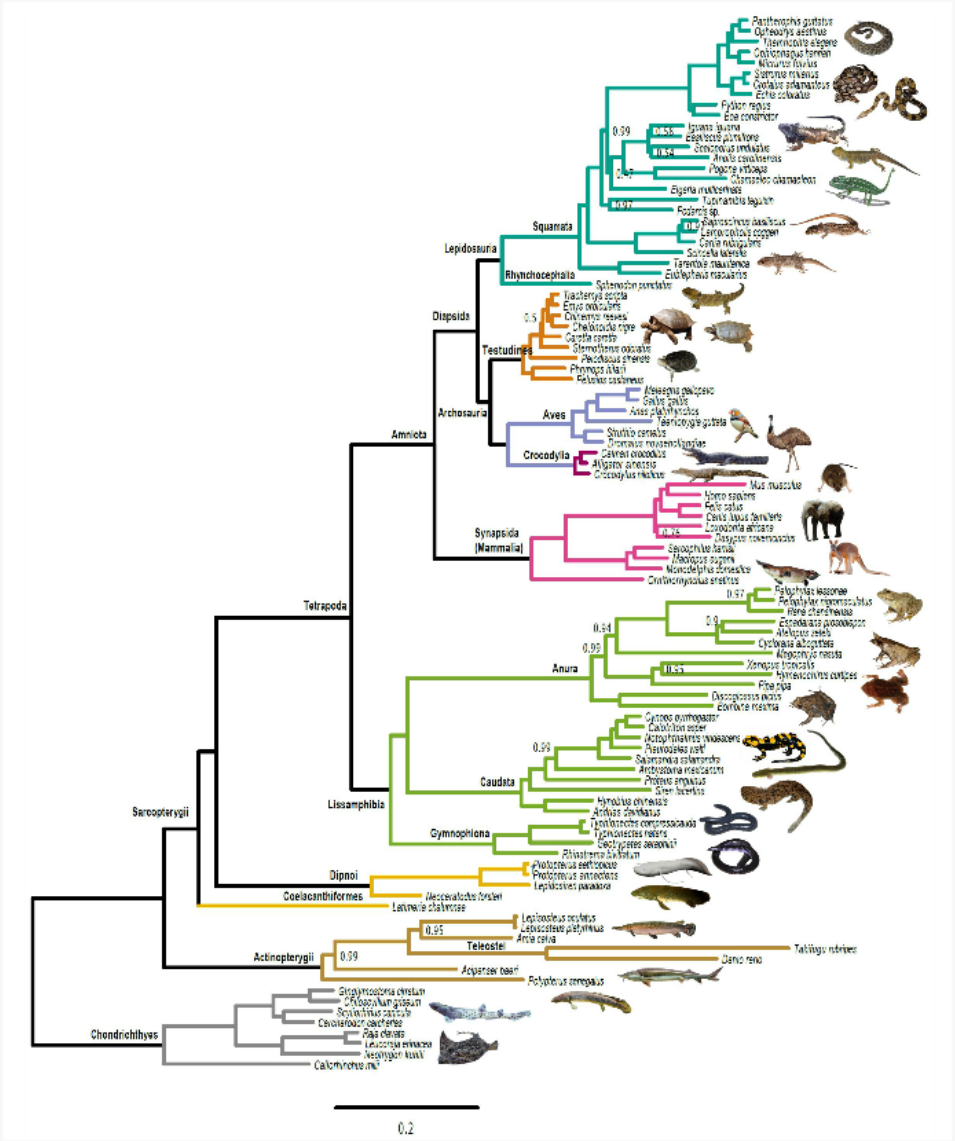
Results



We obtained a new reference phylogenetic framework for the evolution vertebrates.

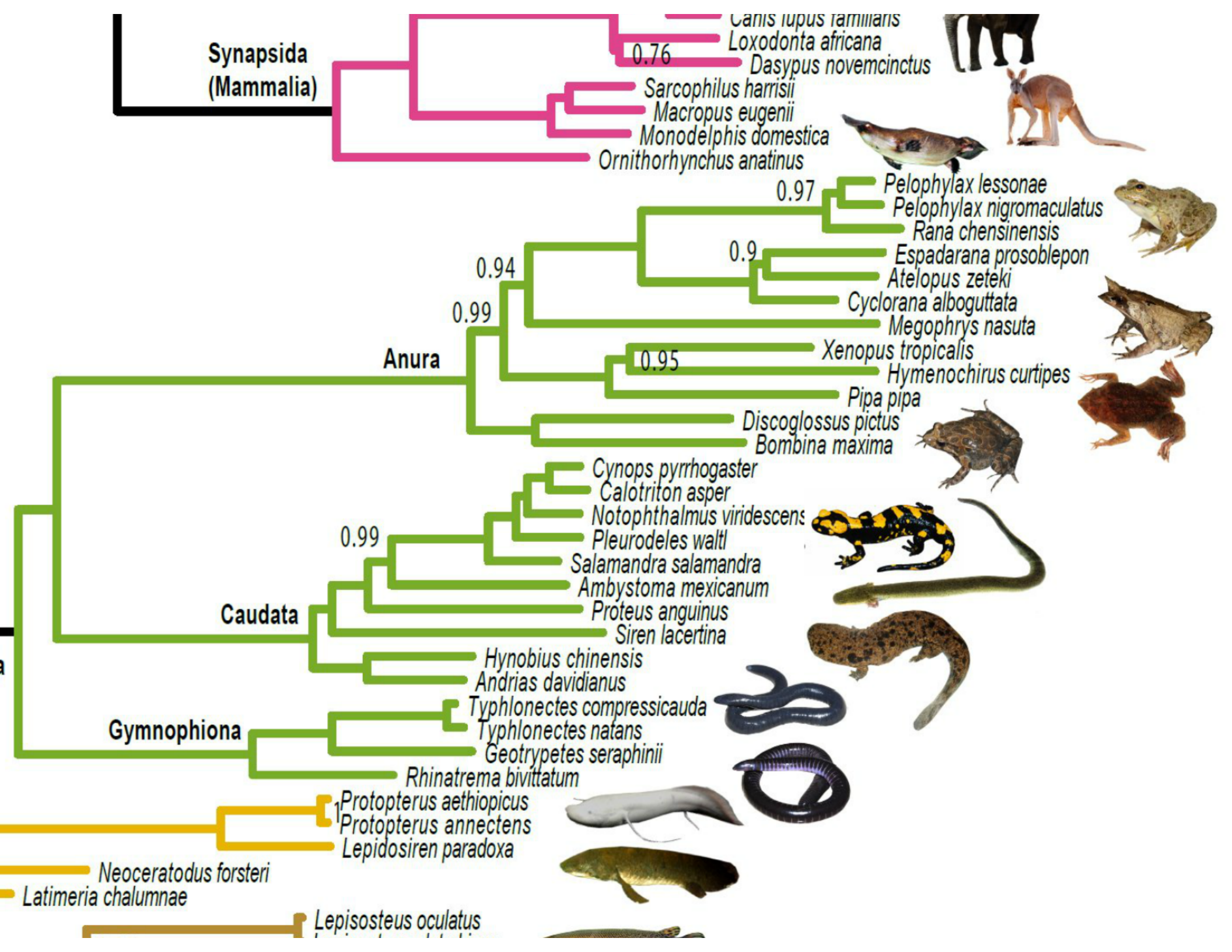
Irisarri, Baurain et al. (in prep.)

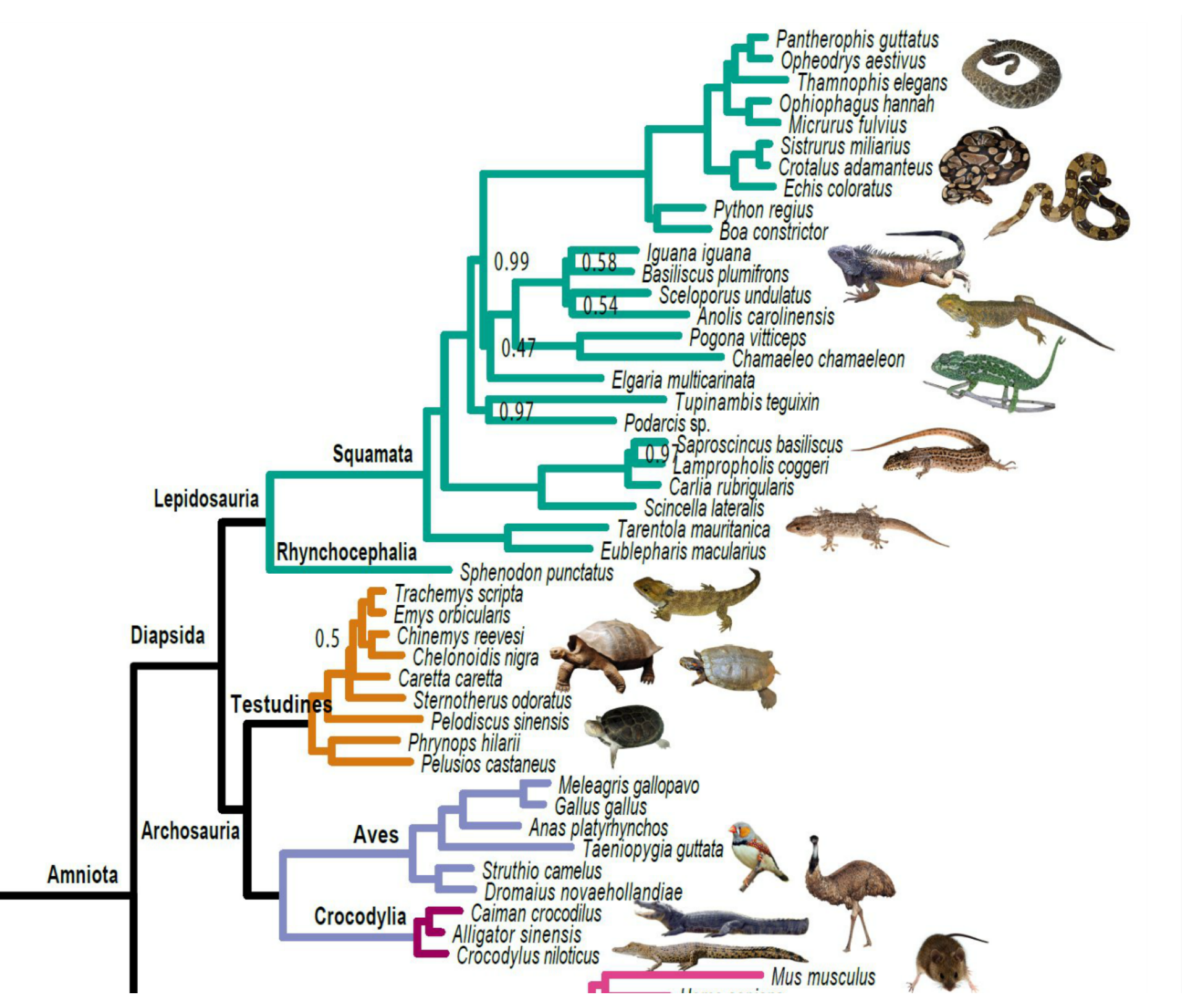
What results have computers enabled you to achieve?



We obtained a new reference phylogenetic framework for the evolution vertebrates.

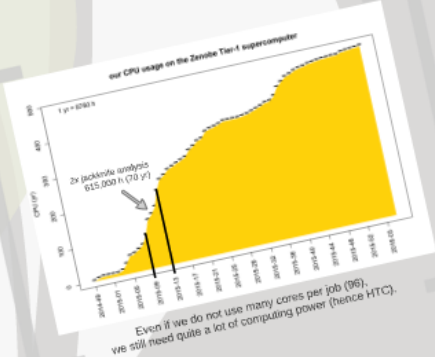
Irisarri, Baurain et al. (in prep.)


Synapsida
(Mammalia)
Canis lupus familiaris
Loxodonta africana
0.76
Dasypus novemcinctus
Sarcophilus harrisii
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
0.97
Pelophylax lessonae
Pelophylax nigromaculatus
Rana chensinensis
0.9
Espadarana prosoblepon
0.94
Atelopus zeteki
Cyclorana alboguttata
0.99
Megophrys nasuta
Anura
Xenopus tropicalis
0.95
Hymenochirus curtipes
Pipa pipa
Discoglossus pictus
Bombina maxima
Cynops pyrrhogaster
Calotriton asper
Notophthalmus viridescens
0.99
Pleurodeles waltl
Salamandra salamandra
Ambystoma mexicanum
Caudata
Proteus anguinus
Siren lacertina
Hynobius chinensis
Andrias davidianus
Typhlonectes compressicauda
Gymnophiona
Typhlonectes natans
Geotrypetes seraphinii
Rhinatrema bivittatum
1
Protopterus aethiopicus
Protopterus annectens
Lepidosiren paradoxa
Neoceratodus forsteri
Latimeria chalumnae
Lepisosteus oculatus


Pantherophis guttatus
Opheodrys aestivus
Thamnophis elegans
Ophiophagus hannah
Micrurus fulvius
Sistrurus miliarius
Crotalus adamanteus
Echis coloratus
Python regius
Boa constrictor
0.99
0.58
Iguana iguana
Basiliscus plumifrons
Sceloporus undulatus
0.54
Anolis carolinensis
Pogona vitticeps
0.47
Chamaeleo chamaeleon
Elgaria multicarinata
Tupinambis teguixin
0.97
Podarcis sp.
Squamata
0.97
Saproscincus basiliscus
Lampropholis coggeri
Carlia rubrigularis
Lepidosauria
Scincella lateralis
Tarentola mauritanica
Eublepharis macularius
Rhynchocephalia
Sphenodon punctatus
Trachemys scripta
Emys orbicularis
Diapsida
0.5
Chinemys reevesi
Chelonoidis nigra
Caretta caretta
Testudines
Sternotherus odoratus
Pelodiscus sinensis
Phrynops hilarii
Pelusios castaneus
Meleagris gallopavo
Gallus gallus
Archosauria
Anas platyrhynchos
Aves
Taeniopygia guttata
Struthio camelus
Dromaius novaehollandiae
Amniota
Crocodylia
Caiman crocodilus
Alligator sinensis
Crocodylus niloticus
Mus musculus

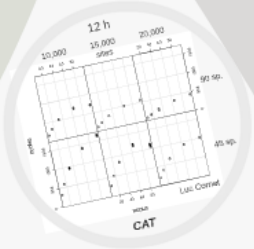
Perspectives

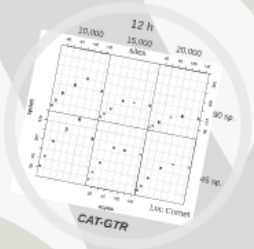
We use templating and job arrays whenever possible to ease up submission and restart.

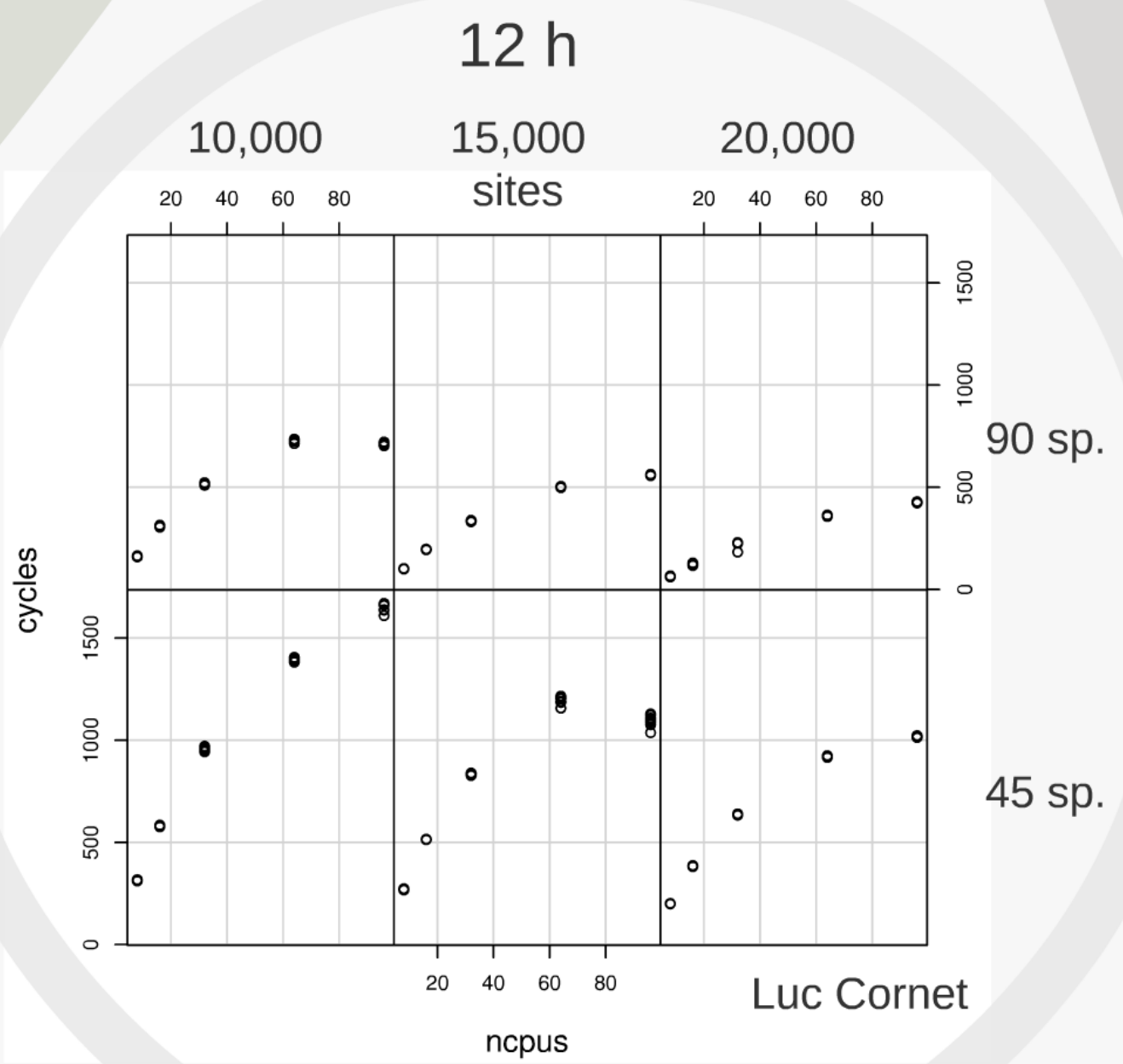


Even if we do not use many cores per job (96), we still need quite a lot of computing power (hence HTC).

What experience do you feel can be transposed to another research field? Can you share any tips and tricks?





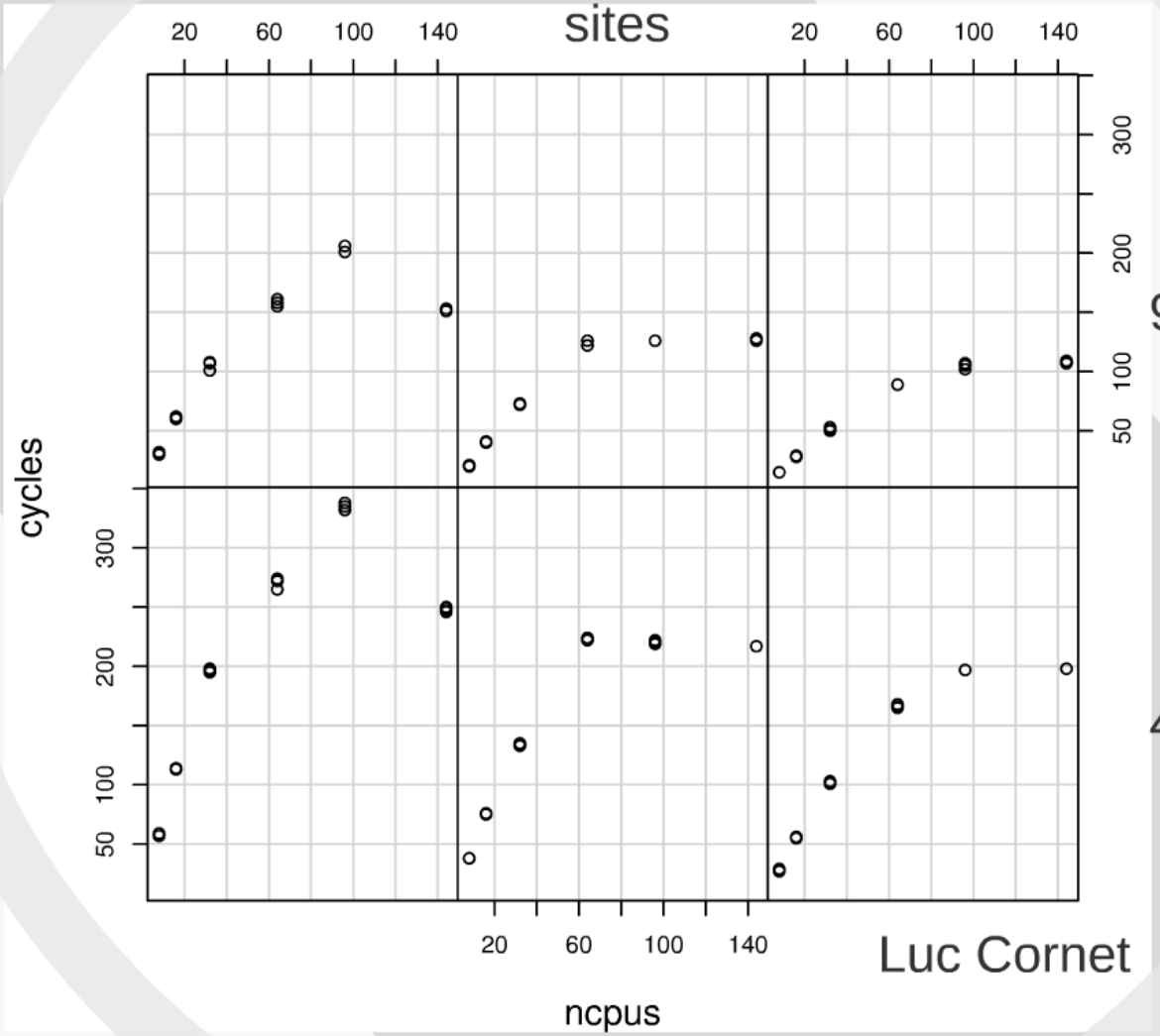

12 h
10,000
15,000
20,000
sites
20
40
60
80
cycles
0
500
1000
1500
90 sp.
45 sp.
ncpus
Luc Cornet

CAT

12 h

10,000 15,000 20,000

sites

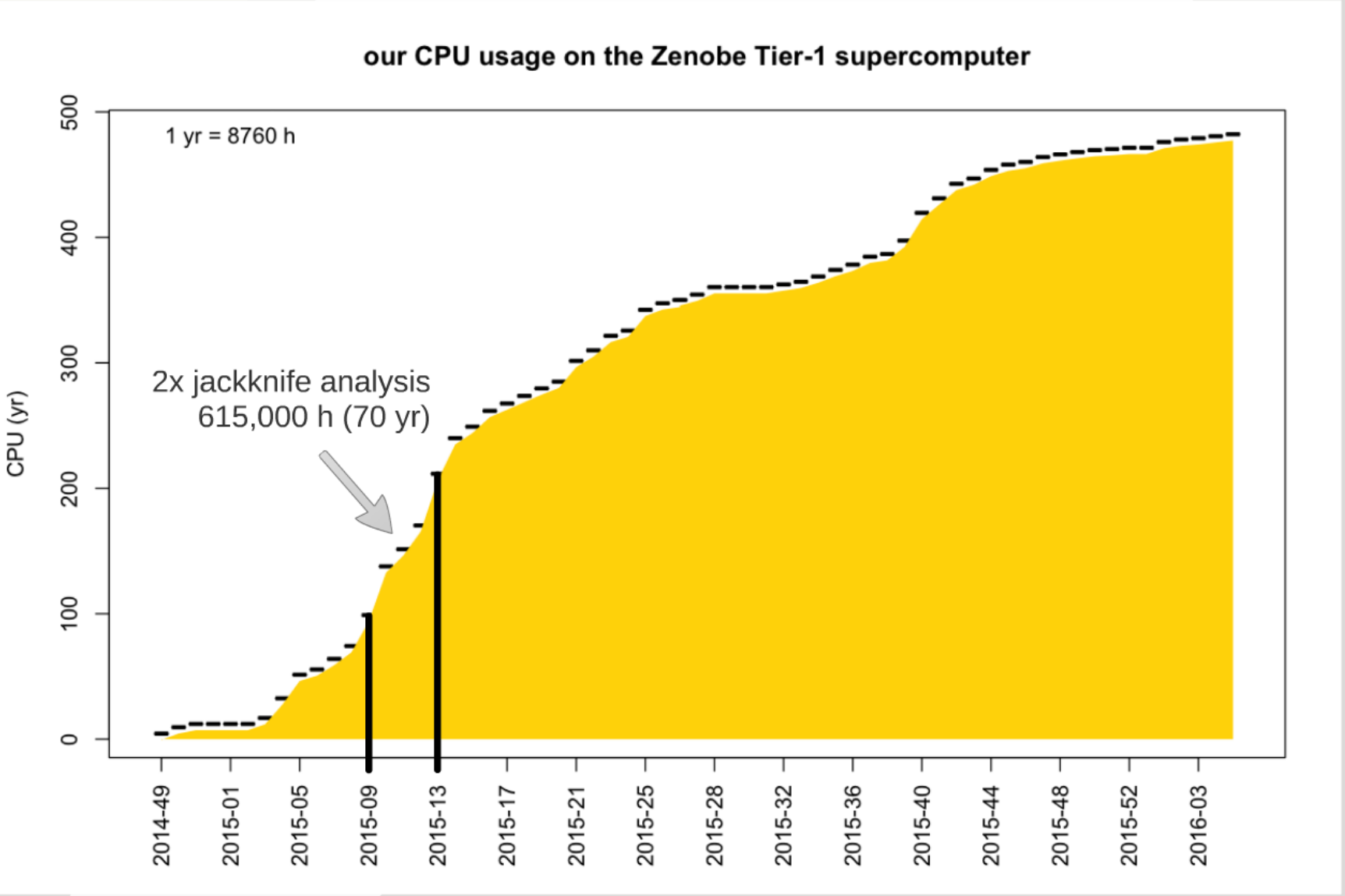


90 sp.

45 sp.

Luc Cornet

CAT-GTR



Even if we do not use many cores per job (96),
we still need quite a lot of computing power (hence HTC).

Acknowledgment

- **Liège**
  - Luc Cornet
  - Raphaël Léonard
  - Damien Sirjacobs
- **CÉCI**
  - David Colignon
- **Cenaero**
  - Danielle Coulon
- **Région wallonne**
  - Gr. agr. 1117545
- **Konstanz**
  - Iker Irisarri
- **Braunschweig**
  - Henner Brinkmann
  - Jörn Petersen
  - Miguel Vences
- **Montpellier**
  - Frédéric Delsuc
- **Moulis / Montreal**
  - Hervé Philippe

Let's have a drink now!