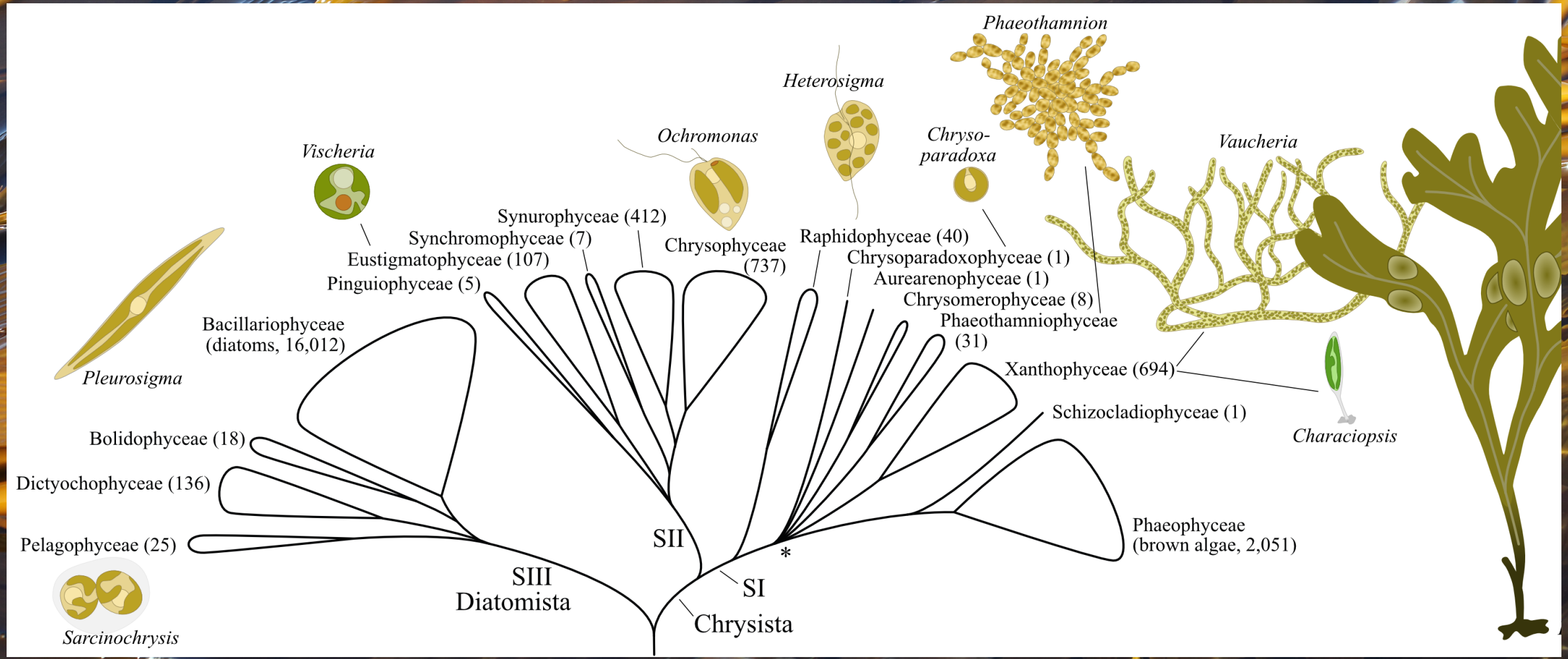
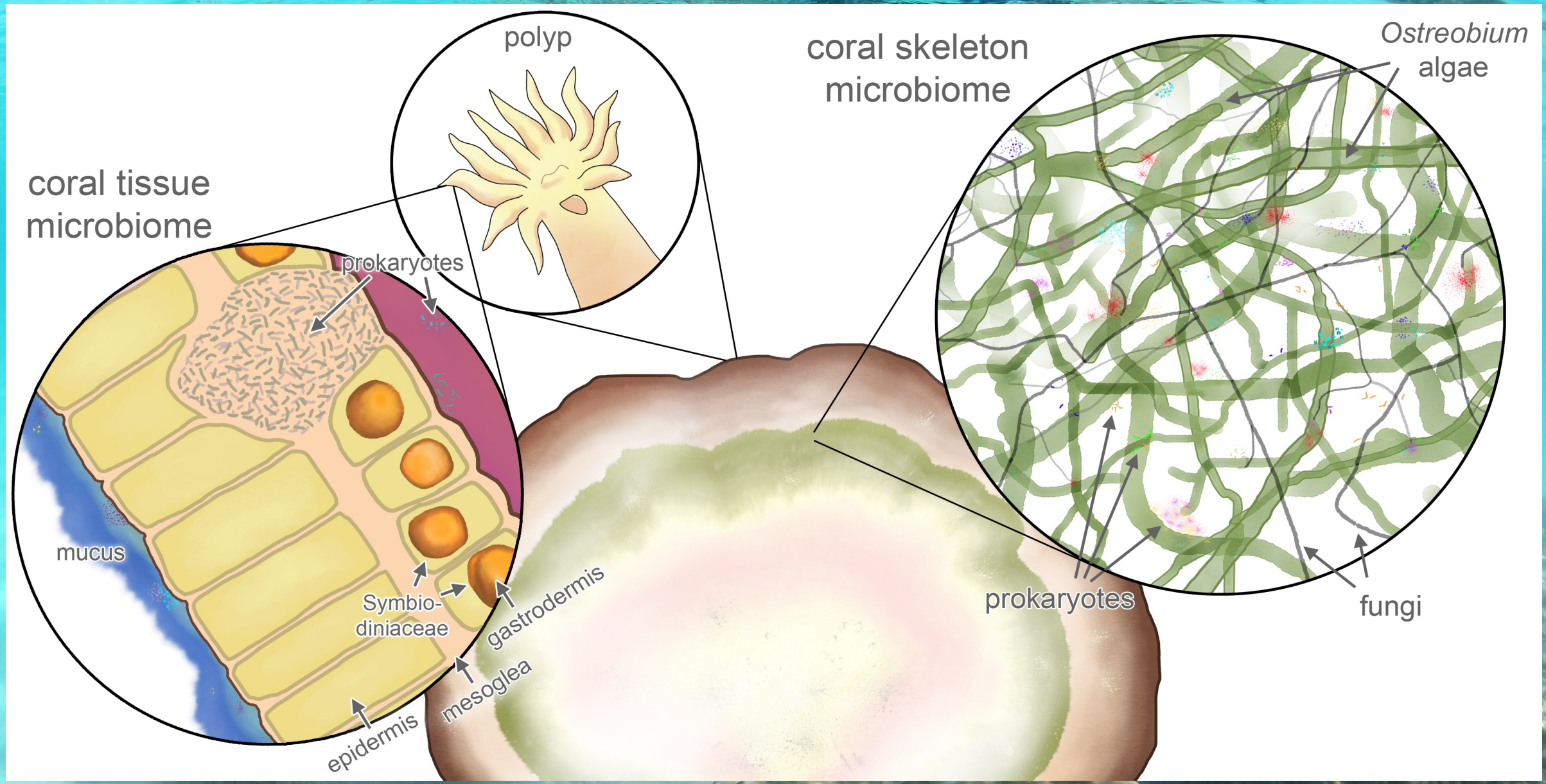
GENE90018: Current Topics in Evolutionary Genetics

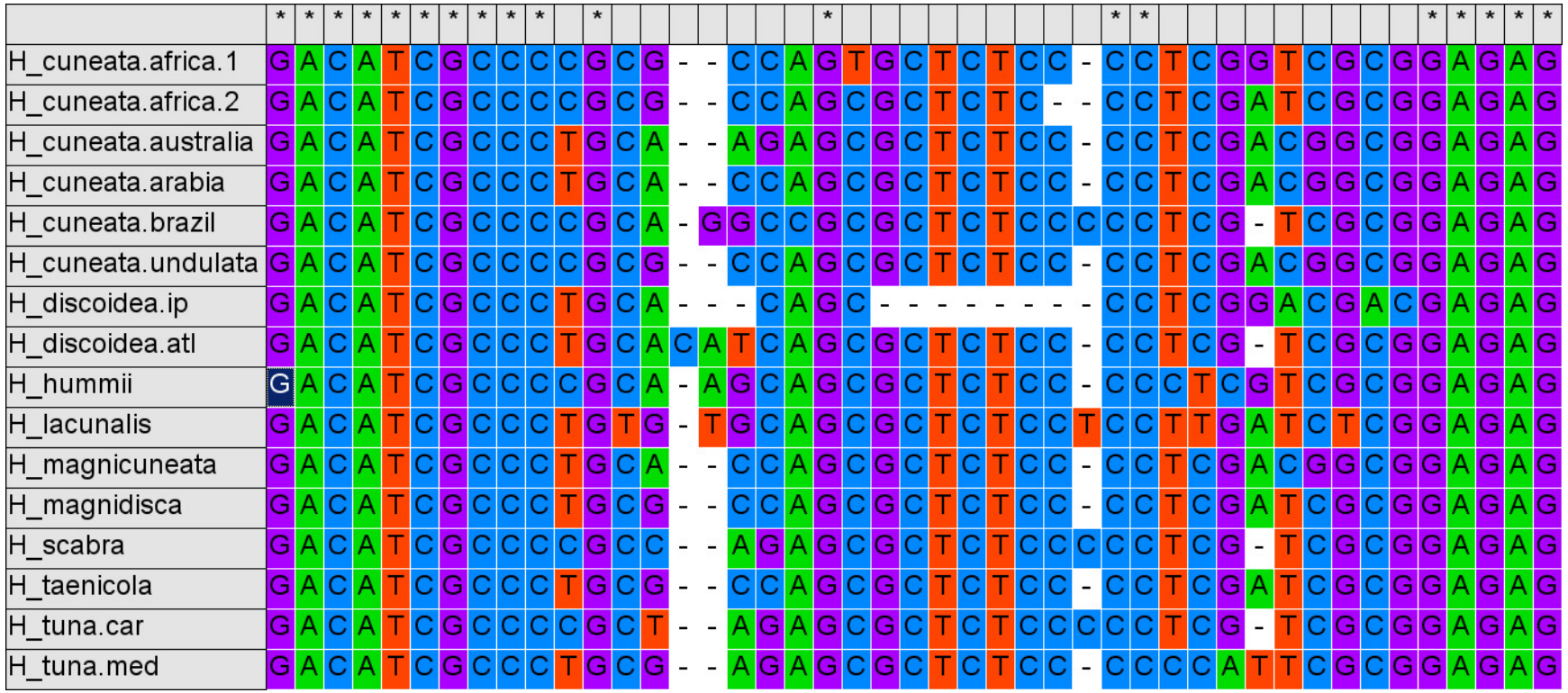
Maximum likelihood phylogenetic inference

Heroen Verbruggen
School of BioSciences
heroen@unimelb.edu.au


Phaeothamnion
Heterosigma
Ochromonas
Vischeria
Chryso-
paradoxa
Vaucheria
Synurophyceae (412)
Synchromophyceae (7)
Eustigmatophyceae (107)
Pinguiophyceae (5)
Chrysophyceae
(737)
Raphidophyceae (40)
Chrysoparadoxophyceae (1)
Aurearenophyceae (1)
Chrysomerophyceae (8)
Phaeothamniophyceae
(31)
Bacillariophyceae
(diatoms, 16,012)
Pleurosigma
Xanthophyceae (694)
Schizocladiophyceae (1)
Characiopsis
Bolidophyceae (18)
Dictyochophyceae (136)
Pelagophyceae (25)
SII
*
Phaeophyceae
(brown algae, 2,051)
SIII
Diatomista
SI
Chrysista
Sarcinochrysis


polyp
coral tissue microbiome
prokaryotes
mucus
Symbio-diniaceae
gastrodermis
mesoglea
epidermis
coral skeleton microbiome
Ostreobium algae
prokaryotes
fungi

Building a phylogeny starts from a sequence alignment



Overview of phylogenetic methods

- Distance trees
- Counting changes: maximum parsimony
- Modeling evolution: maximum likelihood
- Bayesian inference

Maximum likelihood

Likelihood is based on models

Model describes mathematically how molecular sequences evolve

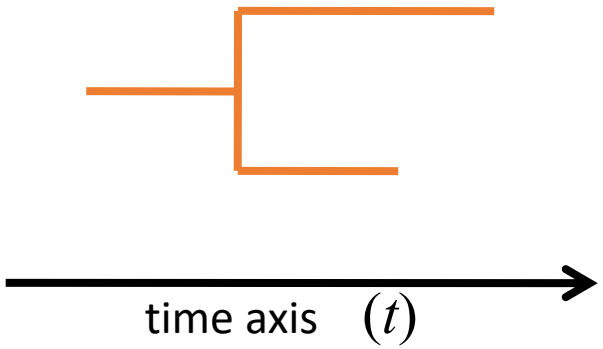
Why likelihood?

- Mathematically explicit: no hidden assumptions
- Results expressed in terms of probability
- Superior performance in phylogenetics

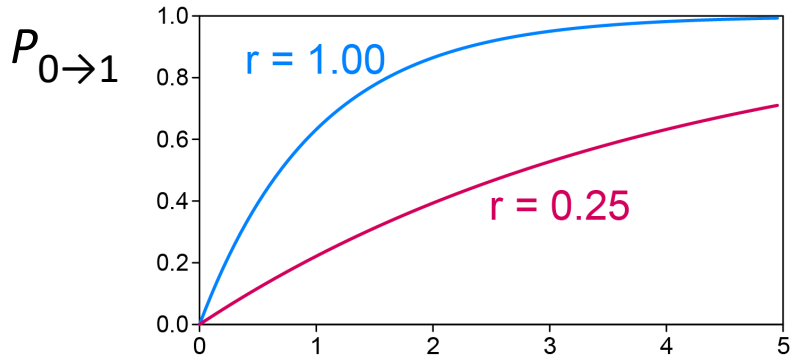
Markov models

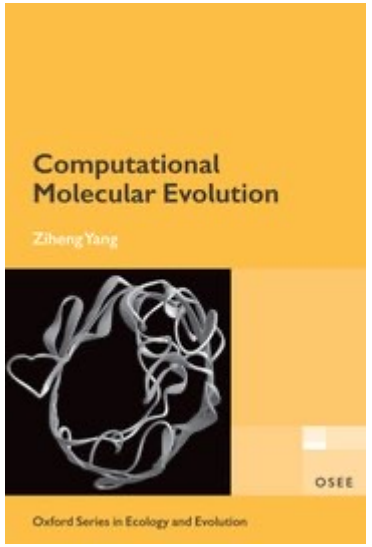
- Simple stochastic process in which the distribution of future states depends only on the present state
- Simplest case: binary character with states 0 and 1

$$P_{0\to 1}(\Delta t) = 1 - e^{-r_{0\to 1}.\Delta t}$$



probability of state transition over time





Markov models: instantaneous rate matrix (Q)

Binary trait: 2-state

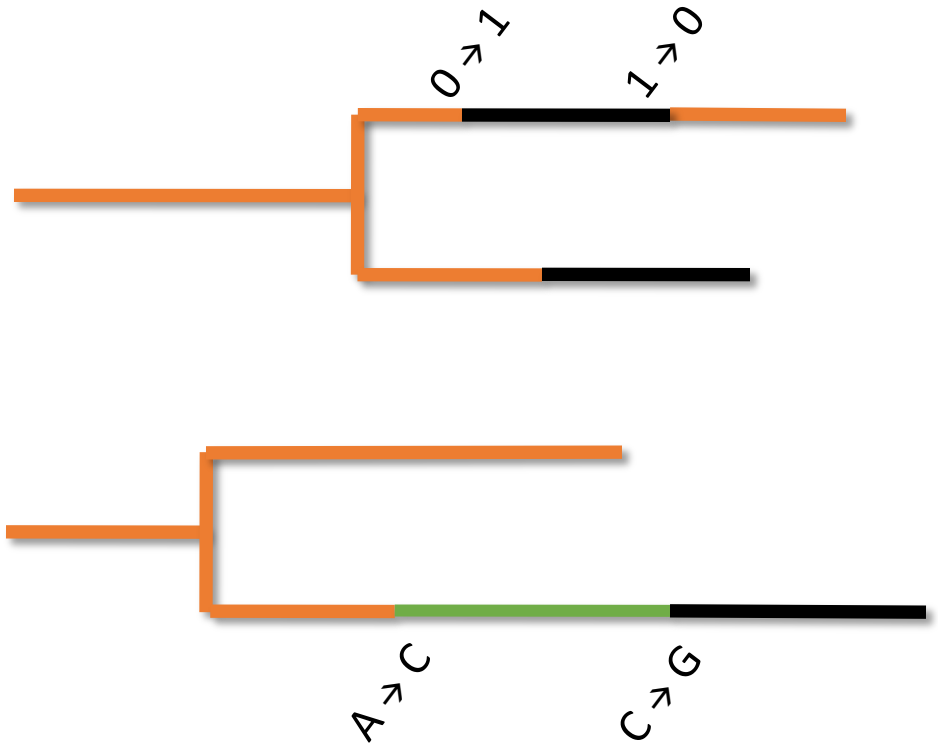
	0	1
0	·	$r_{0\to1}$
1	$r_{1\to0}$	·

DNA data: 4-state

	A	C	G	T
A	·	$r_{A\to C}$	$r_{A\to G}$	$r_{A\to T}$
C	$r_{C\to A}$	·	$r_{C\to G}$	$r_{C\to T}$
G	$r_{G\to A}$	$r_{G\to C}$	·	$r_{G\to T}$
T	$r_{T\to A}$	$r_{T\to C}$	$r_{T\to G}$	·

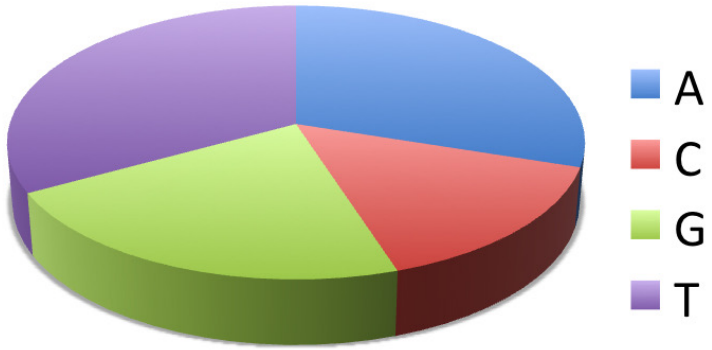
Why models?

- Reversions and multiple changes

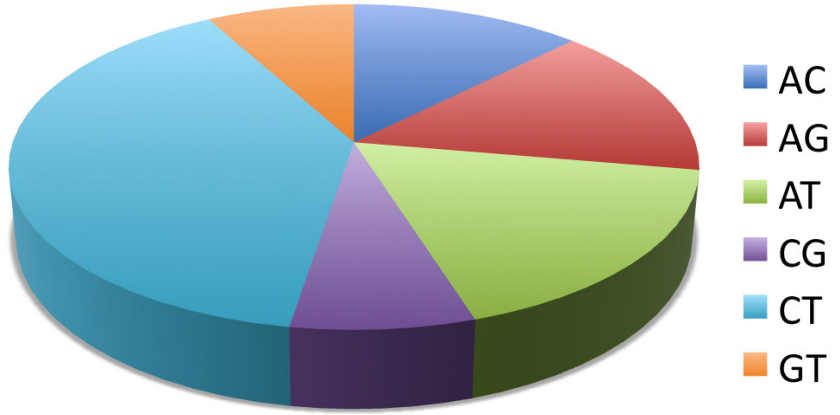


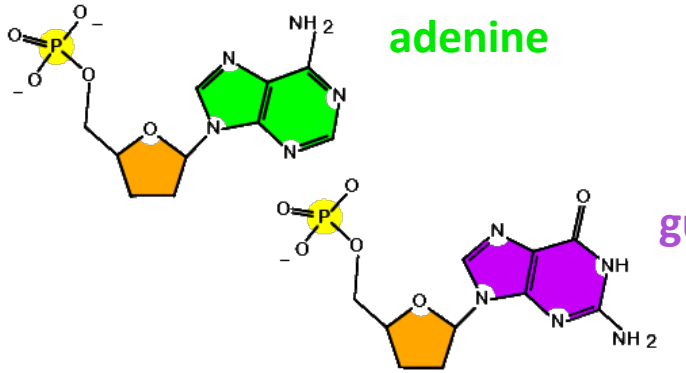
Why models?

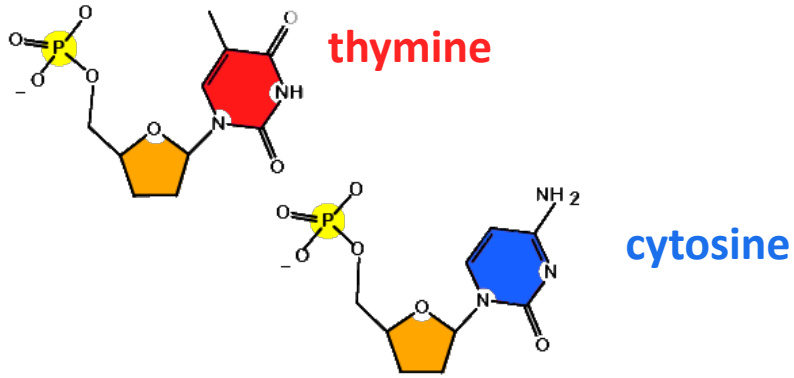
Base frequencies



Rates of nucleotide substitution







Basic elements of General Time Reversible (GTR) model

- Rate matrix
  - 12 rates

$$Q = \begin{bmatrix} \cdot & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & \cdot & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & \cdot & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & \cdot \end{bmatrix}$$

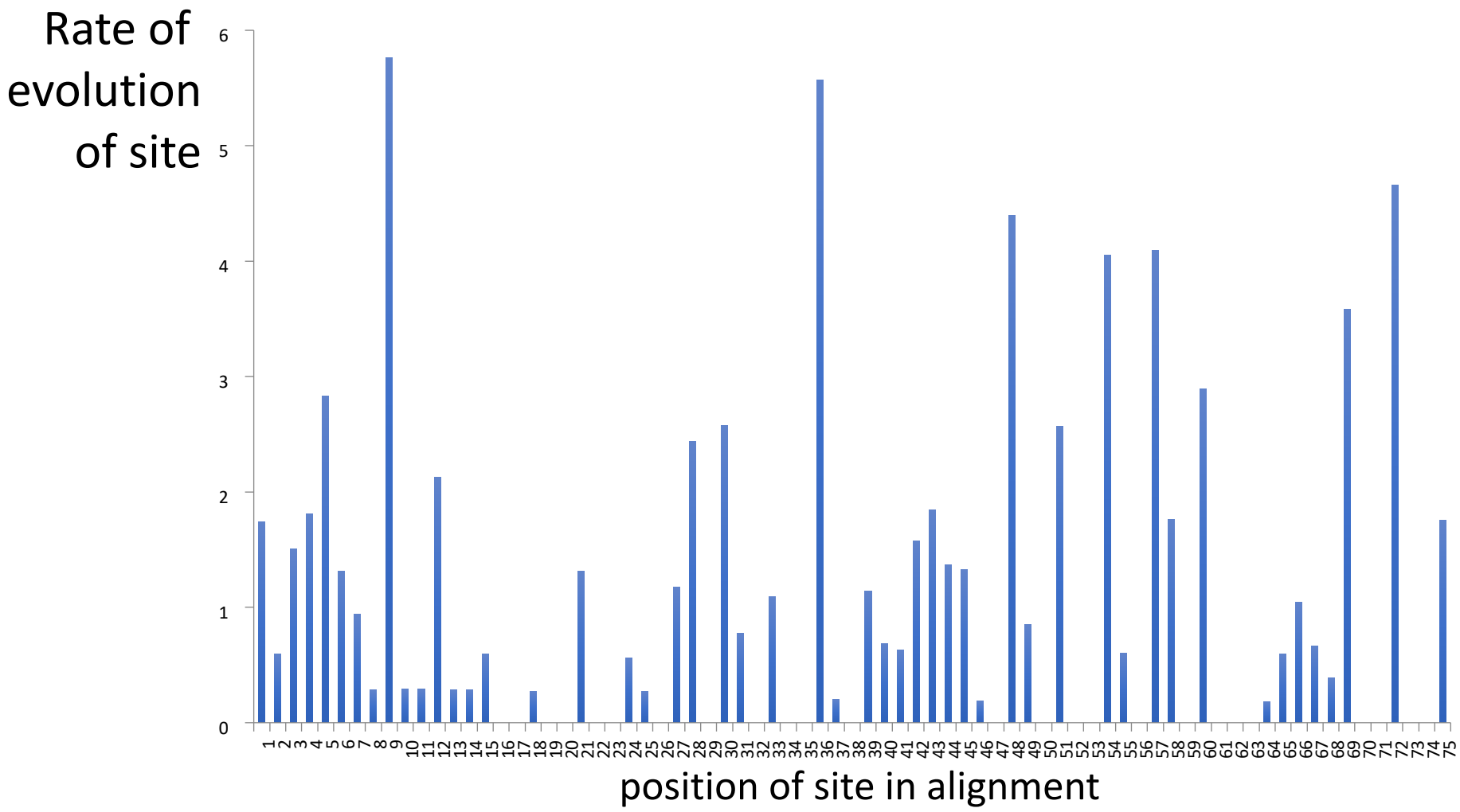
  - Time reversible $r_{XY} = r_{YX}$
  - 5 model parameters:
    all relative to $r_{GT} = 1$

- Base frequencies
  - 4 base frequencies $F = \begin{bmatrix} \pi_A & \pi_C & \pi_G & \pi_T \end{bmatrix}$
  - 3 model parameters $\pi_T = 1 - \left(\pi_A + \pi_C + \pi_G\right)$

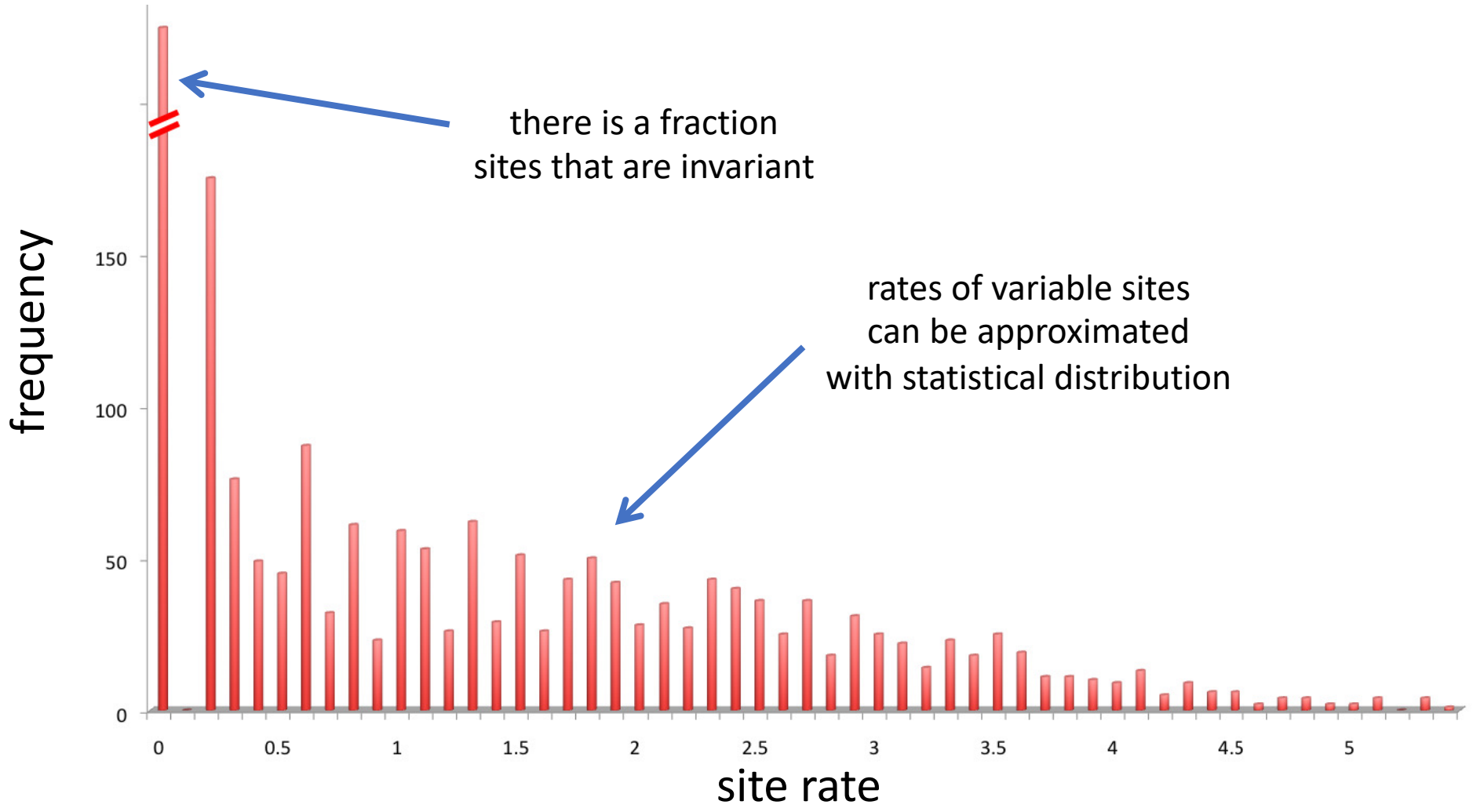
GTR model

$$Q = \begin{bmatrix} & [A] & [C] & [G] & [T] \\ [A] & \cdot & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ [C] & r_{AC}\pi_A & \cdot & r_{CG}\pi_G & r_{CT}\pi_T \\ [G] & r_{AG}\pi_A & r_{CG}\pi_C & \cdot & r_{GT}\pi_T \\ [T] & r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & \cdot \end{bmatrix}$$

Among-site rate variation



Among-site rate variation

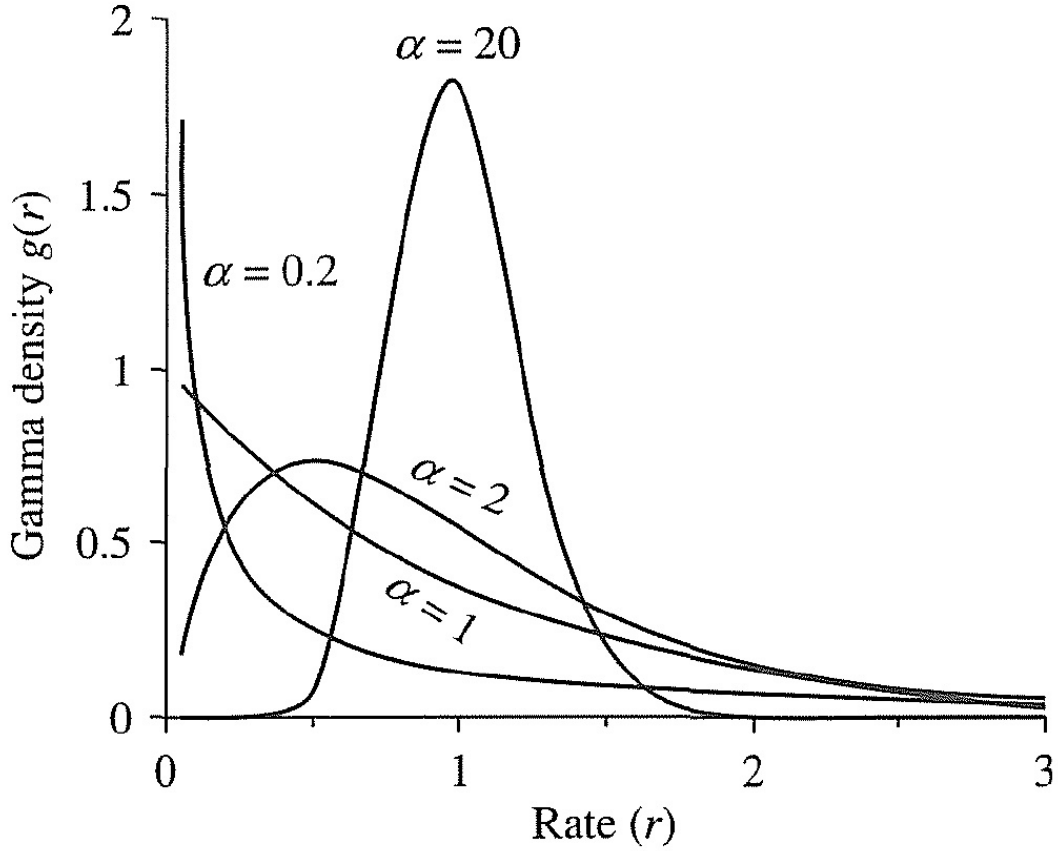


Among-site rate variation

- Gamma distribution: $+\Gamma$

  (mixture model)

- Proportion of invariant sites: $+\mathbf{I}$



Advantages of model-based methods

- Deals better than MP with differences in rates between lineages (but do remain cautious of long branches clustering together)
- ML handles this better because it takes the possibility of homoplasies into consideration (parallel, convergent and back substitutions)
- Major advantage that different types of substitutions (and different sites) carry a more realistic weight thanks to use of models
- Reasonably robust against different types of model violations as long as model is a decent approximation of actual process of molecular evolution
- But, it is computationally more expensive

Likelihood computation

- The likelihood **L** is the probability of the data given the model.

- The data consist of an alignment of sequences.
- The **model** is a hypothesis of how the data were generated.
  - topology
  - branch lengths
  - other model parameters (base frequencies, rate matrix, gamma shape, ...)

- Different topologies are evaluated one at a time.
- Each time, branch lengths and other parameter are optimized.

} **computationally expensive !!**

- Each site has a likelihood, the total likelihood is the product of all site likelihoods.

- The ML tree is the topology corresponding to the model that yielded the highest overall likelihood.

- Likelihoods are reported on a logarithmic scale for mathematical convenience: **ln L**

Searching tree space: a walk through the forest

Which trees are evaluated?

- All trees: **exhaustive** search
- Selection of trees: **heuristic** search

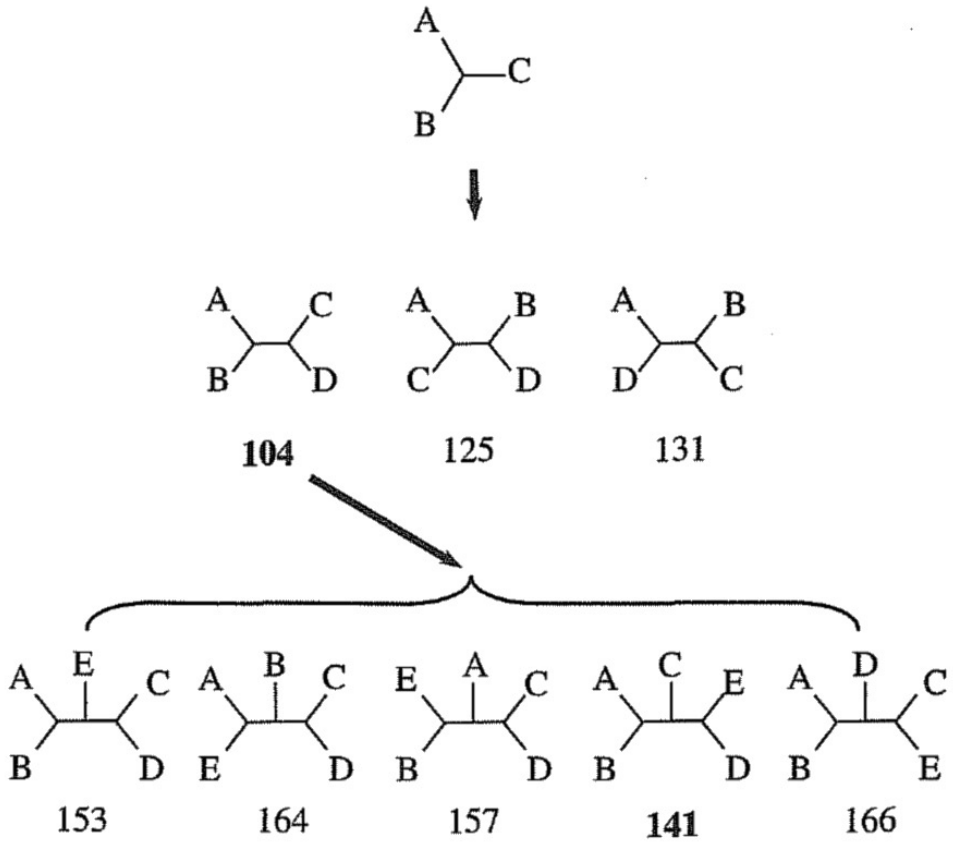
Basic principles for heuristics

- Getting a start tree
- Modifying that tree and evaluating criterion of choice

Species	Number of trees
2	1
3	4
4	26
5	236
6	2,752
7	39,208
8	660,032
9	12,818,912
10	282,137,824
11	6,939,897,856
12	188,666,182,784
13	5,617,349,020,544
14	181,790,703,209,728
15	6,353,726,042,486,272
16	238,513,970,965,257,728
17	9,571,020,586,419,012,608
18	408,837,905,660,444,010,496
19	18,522,305,410,364,986,906,624
20	887,094,711,304,119,347,388,416
30	7.0717×10^{41}
40	1.9037×10^{61}
50	6.85×10^{81}
100	3.3388×10^{195}

Getting a start tree

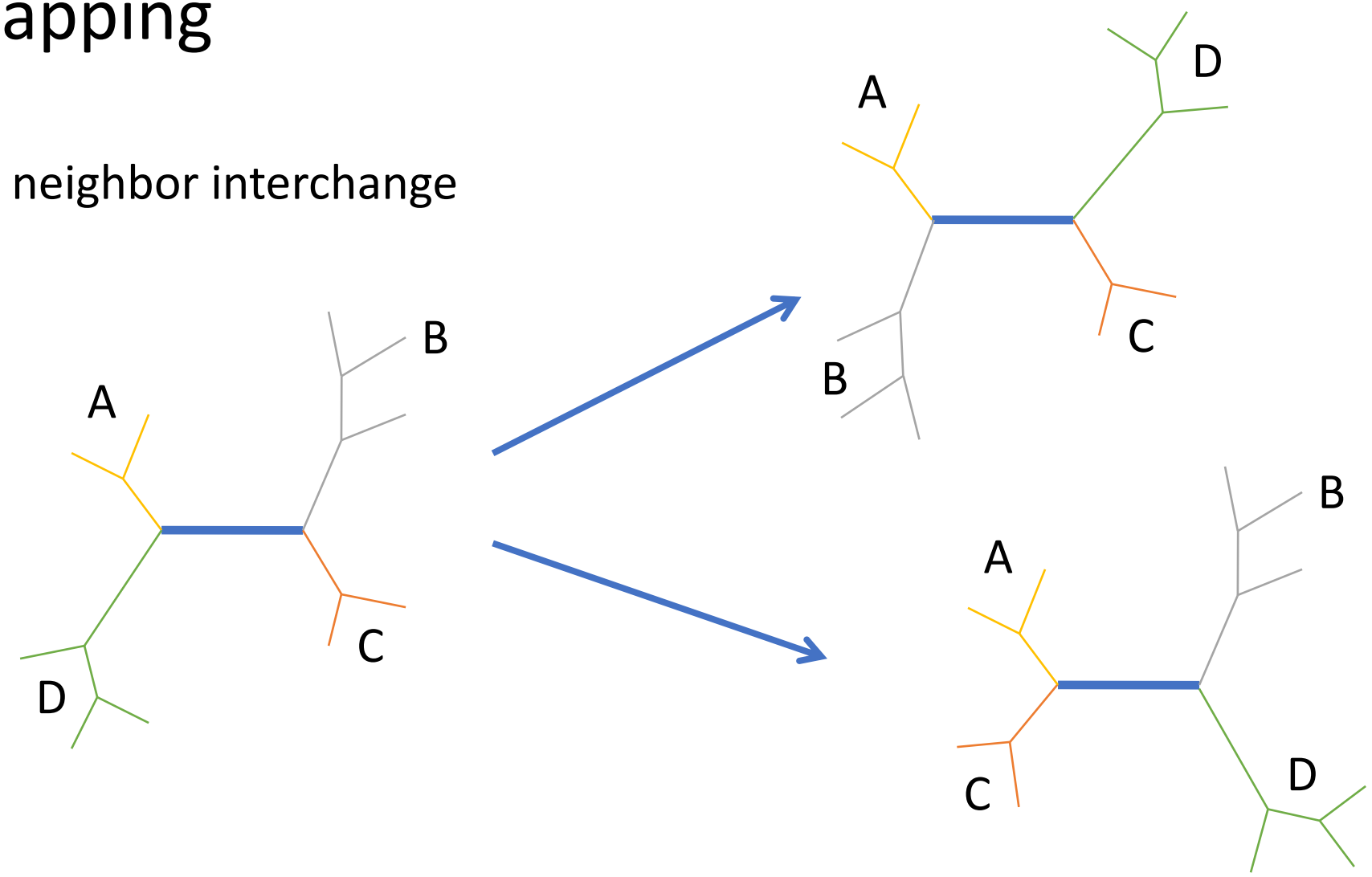
Stepwise addition:



Other options: Star decomposition, Distance method, MP search

Branch swapping

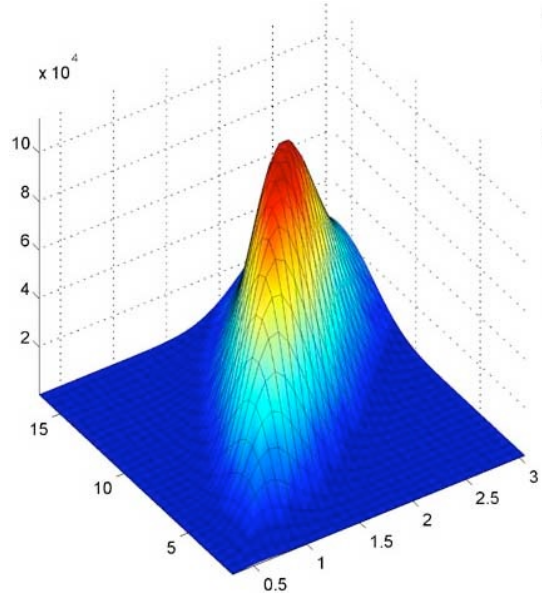
NNI: nearest neighbor interchange

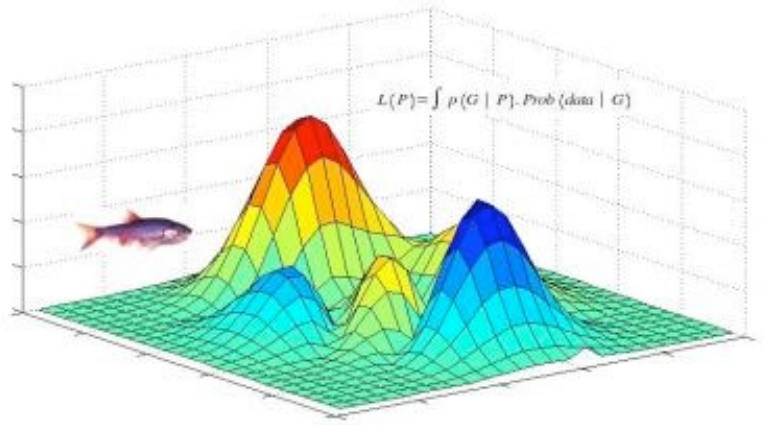


Other options: SPR (subtree pruning & regrafting) and TBR (tree bisection & reconnection)

Heuristic search

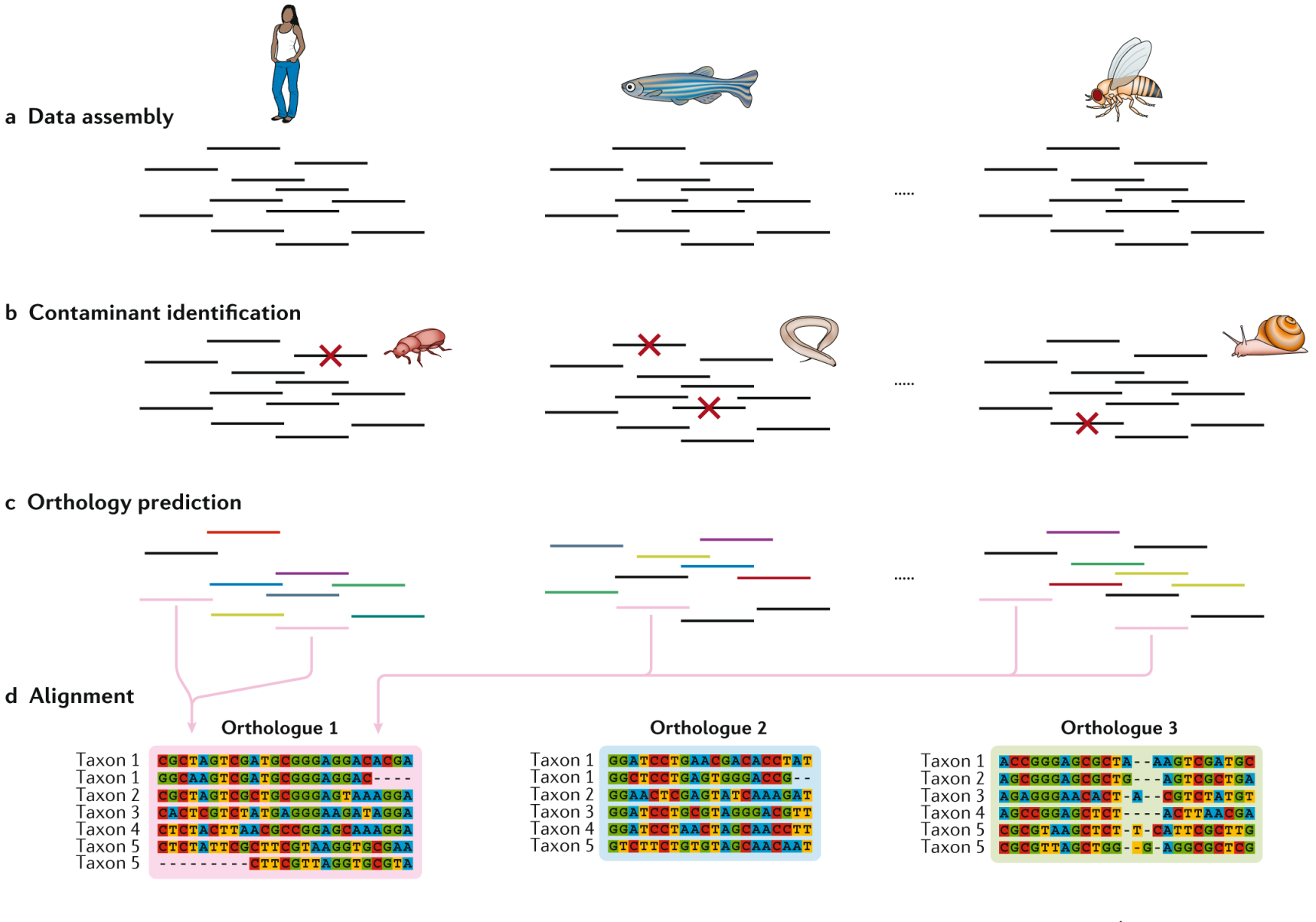
- Hill climbing algorithms will not always climb the tallest hill



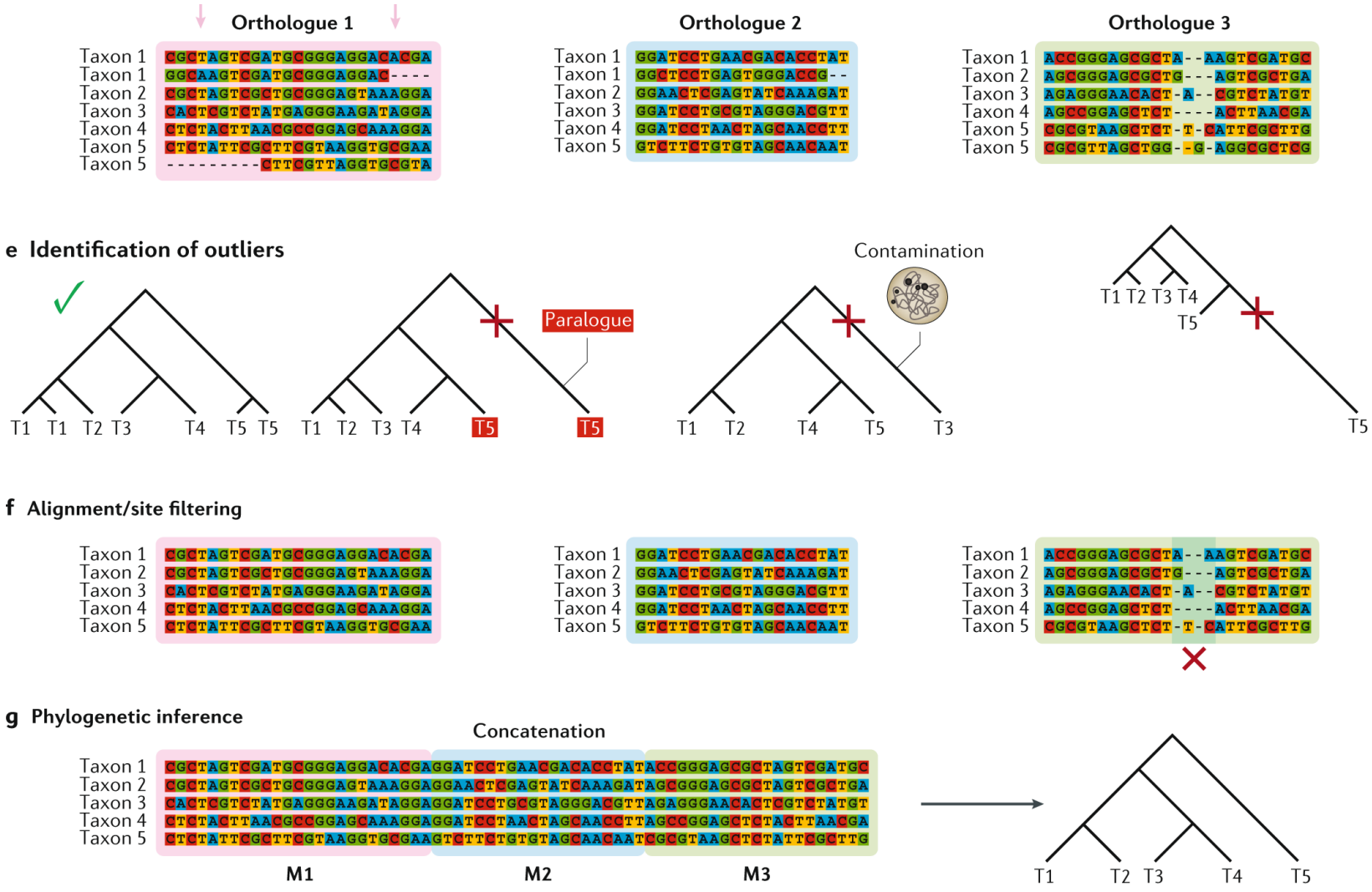


- Software will differ in how likely they are to find the optimal tree
- Use multiple start trees to start hill climbing process

Genome-scale phylogenetics

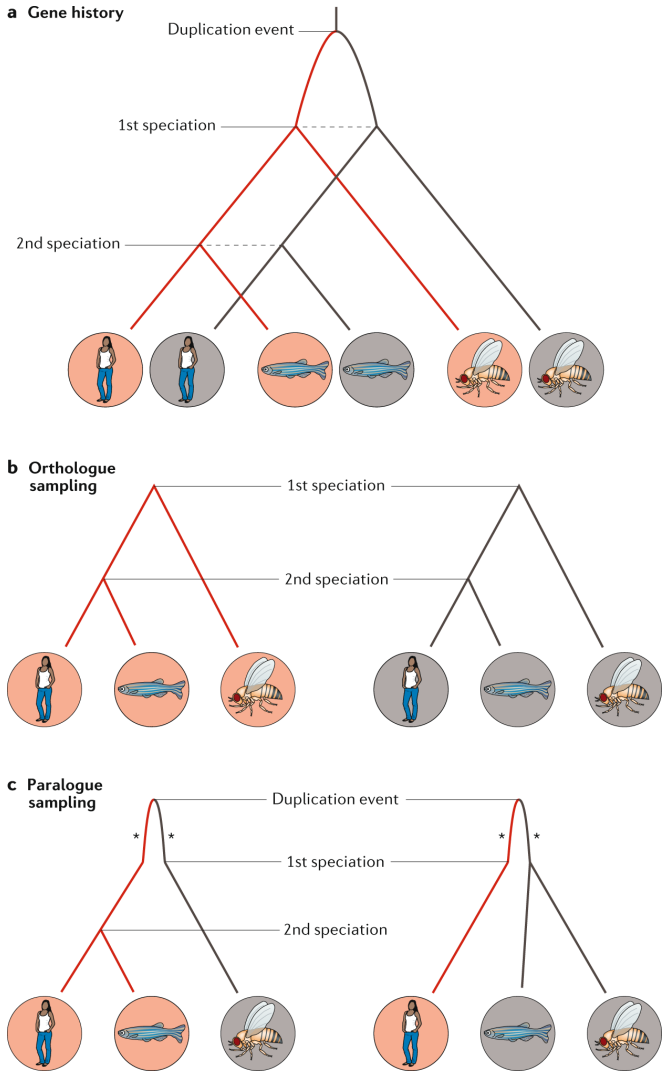


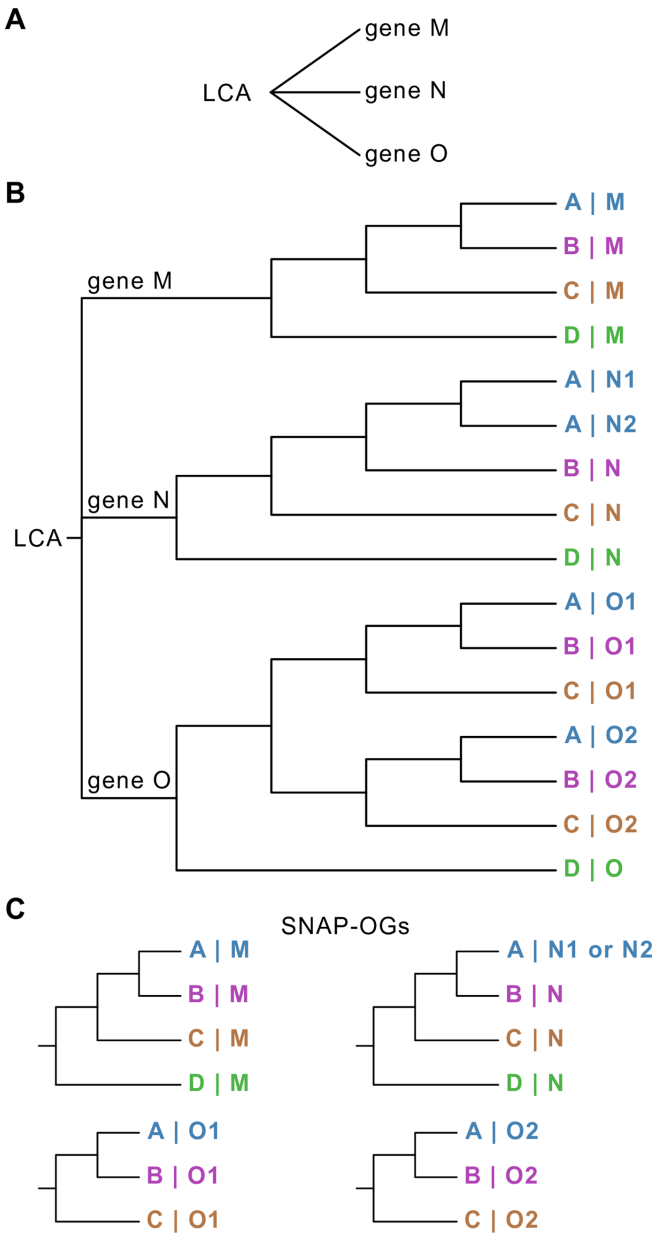
Genome-scale phylogenetics



Kapli et al. (2020) https://doi.org/10.1038/s41576-020-0233-0

Dealing with paralogs





Steenwyk et al. (2022) https://doi.org/10.1371/journal.pbio.3001827