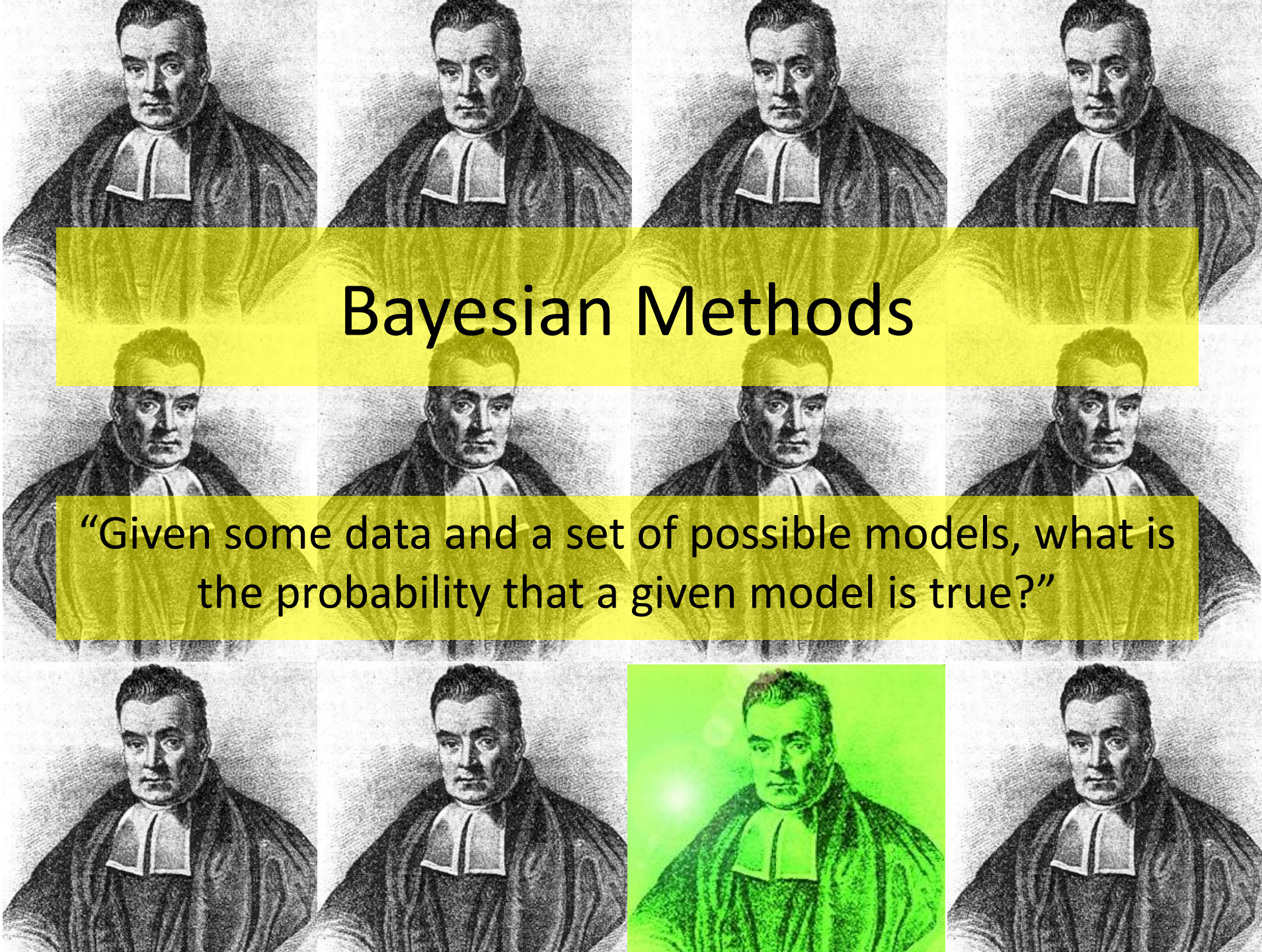
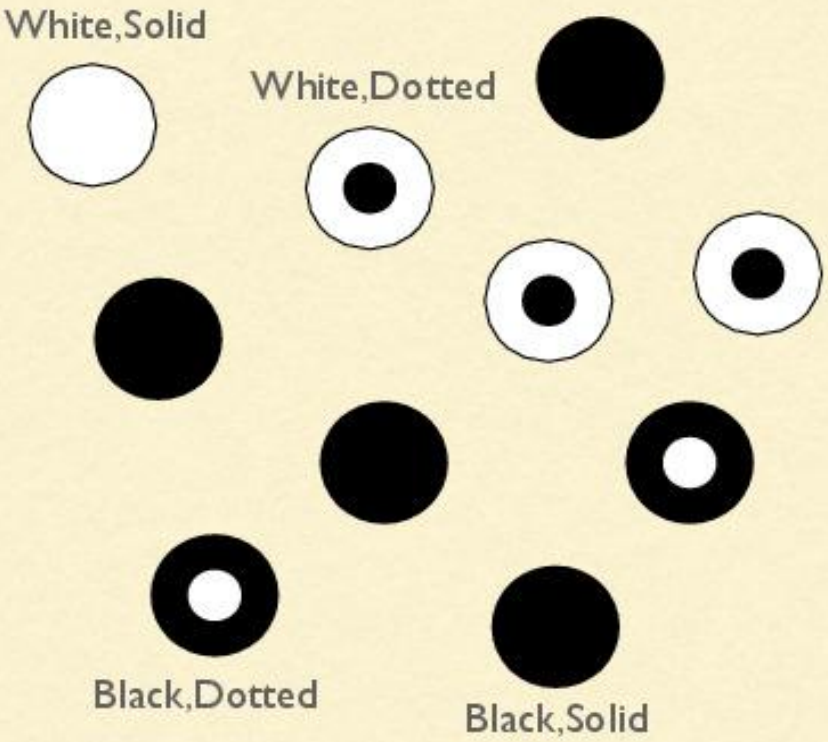
Bayesian Methods

"Given some data and a set of possible models, what is the probability that a given model is true?"



Joint probabilities



10 marbles in a bag
Sampling with replacement

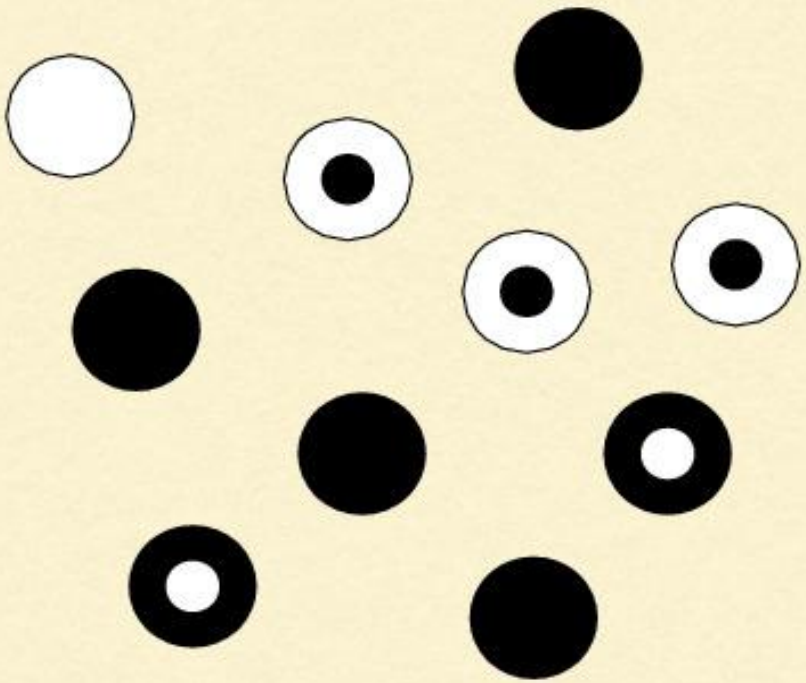
Pr(B,S) = 0.4

Pr(W,S) = 0.1

Pr(B,D) = 0.2

Pr(W,D) = 0.3

Conditional probabilities



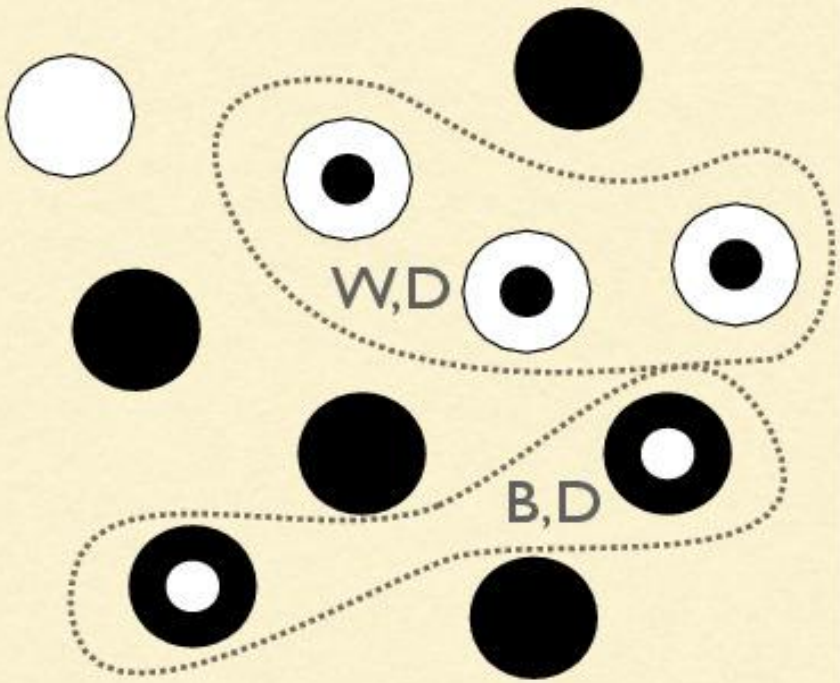
What's the probability that a marble is black given that it is dotted?

5 marbles satisfy the condition (D)

$$\Pr(B|D) = \frac{2}{5}$$

2 remaining marbles are black (B)

Marginal probabilities



Marginalizing over color yields the total probability that a marble is dotted (D)

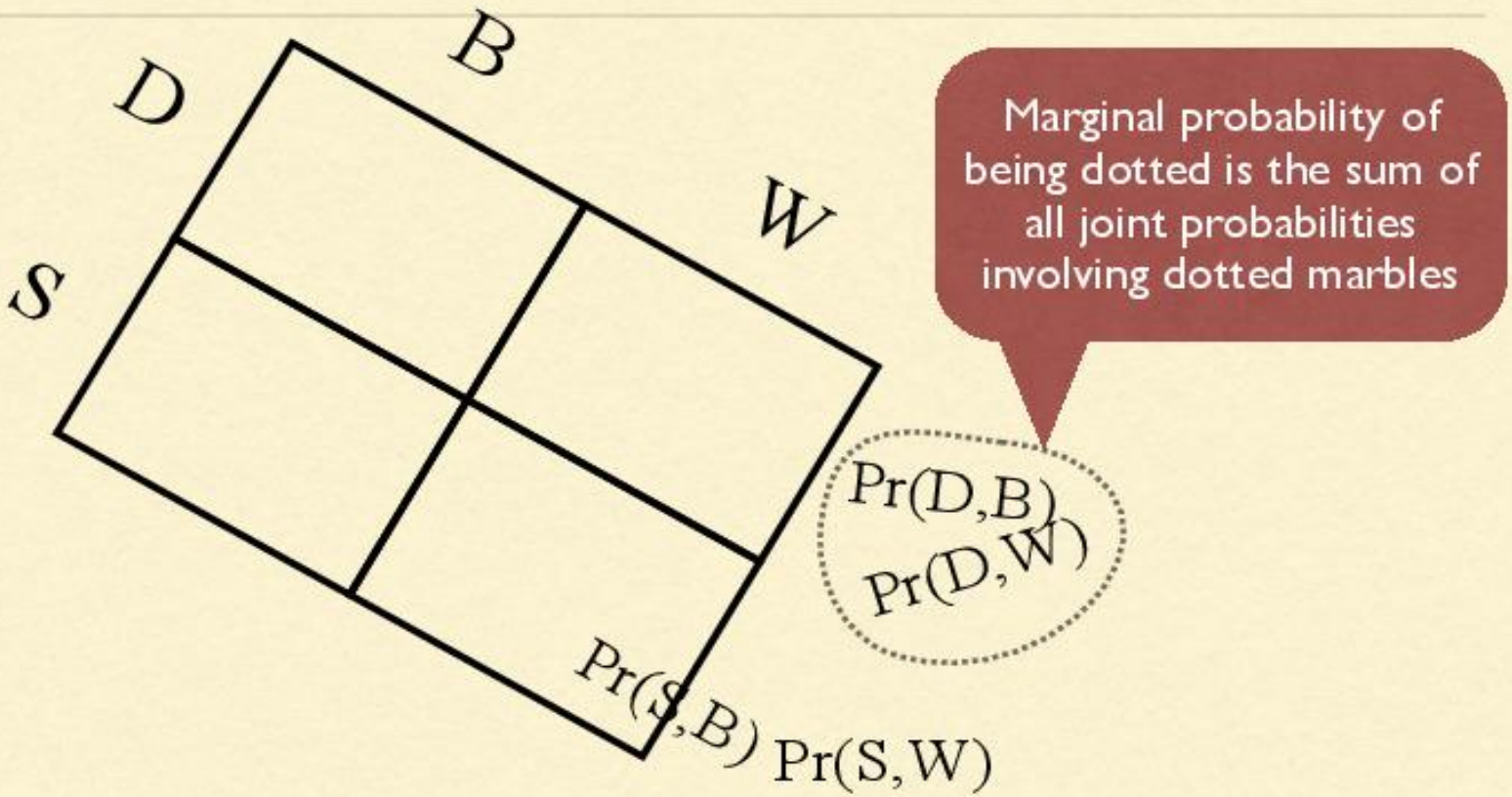
$$\Pr(\mathbf{D}) = \Pr(B,\mathbf{D}) + \Pr(W,\mathbf{D})$$
$$= 0.2 + 0.3$$
$$= 0.5$$

Marginalization involves summing all joint probabilities containing D

Marginalization

	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

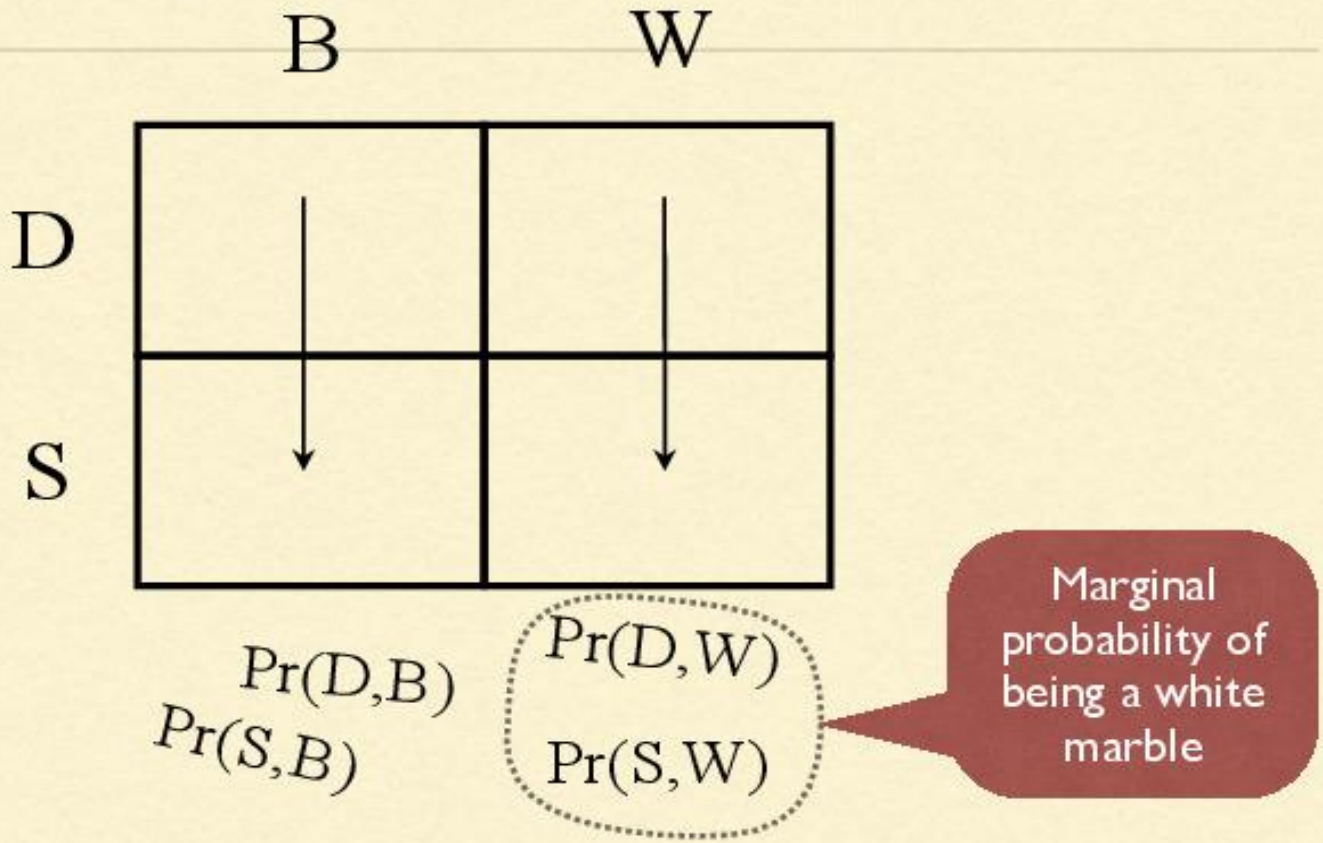
Marginalizing over colors



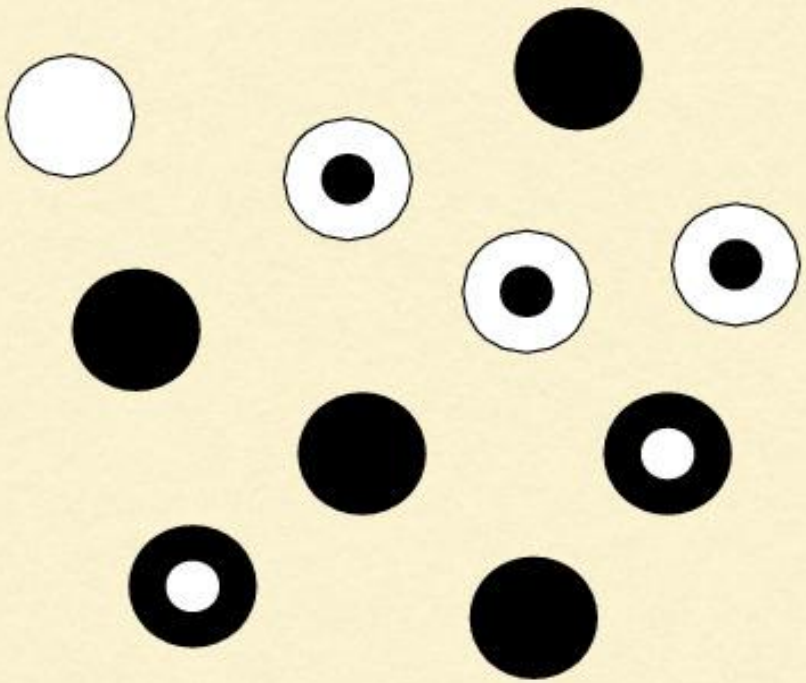
Joint probabilities

	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

Marginalizing over "dottedness"



Bayes' rule



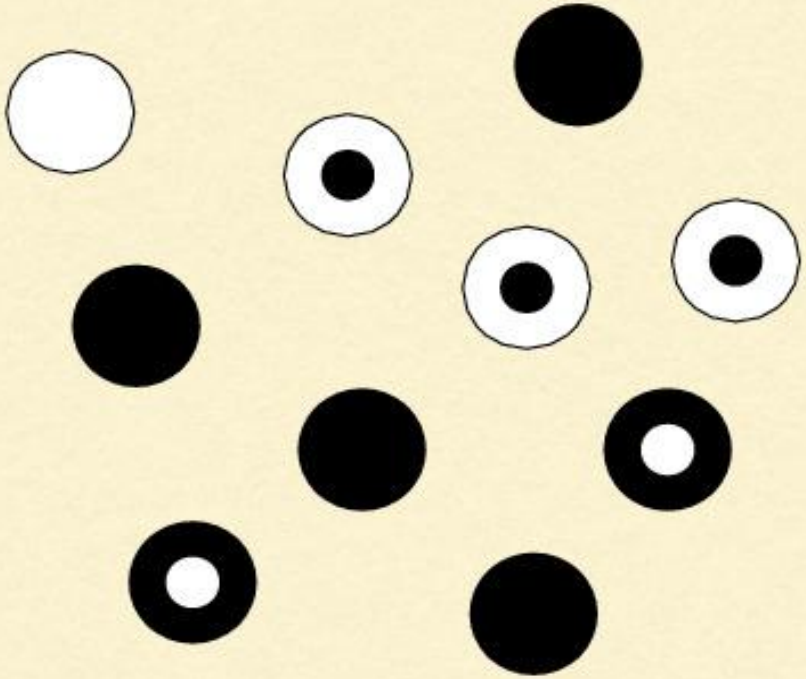
The joint probability Pr(B,D) can be written as the product of a *conditional probability* **and the** *probability of that condition*

$$\Pr(B,D) \rightarrow \Pr(B|D)\,\Pr(D)$$

$$\Pr(B,D) \rightarrow \Pr(D|B)\,\Pr(B)$$

Either B or D can be the condition

Bayes' rule



Equate the two ways of writing Pr(B,D)

$$\Pr(B|D)\ \Pr(D) = \Pr(D|B)\ \Pr(B)$$

Divide both sides by Pr(D)

$$\frac{\Pr(B|D)\ \cancel{\Pr(D)}}{\cancel{\Pr(D)}} = \frac{\Pr(D|B)\ \Pr(B)}{\Pr(D)}$$

Bayes' rule

$$\Pr(B|D) = \frac{\Pr(D|B)\ \Pr(B)}{\Pr(D)}$$

Bayes' rule (variations)

$$\Pr(B|D) = \frac{\Pr(D|B)\Pr(B)}{\Pr(D)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\Pr(B,D) + \Pr(W,D)}$$

Pr(*D*) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**

Bayes' rule (variations)

$$\Pr(B|D) = \frac{\Pr(D|B)\Pr(B)}{\Pr(B,D) + \Pr(W,D)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\Pr(D|B)\Pr(B) + \Pr(D|W)\Pr(W)}$$

$$= \frac{\Pr(D|B)\Pr(B)}{\sum_{\theta \in \{B,W\}} \Pr(D|\theta)\Pr(\theta)}$$

Bayes' rule in statistics

Likelihood of hypothesis θ

Prior probability of hypothesis θ

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

The Bayesian Gist

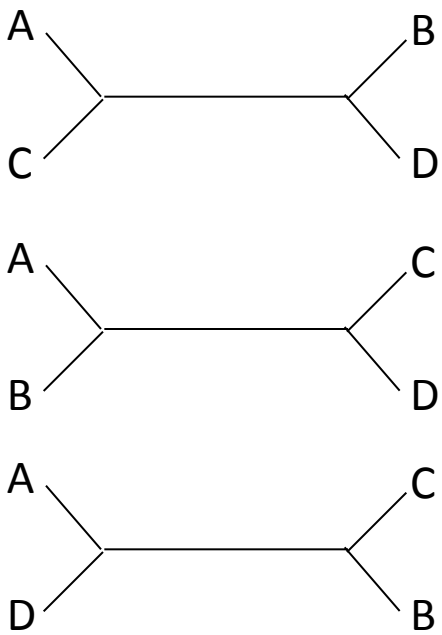
- Take your prior beliefs about the model (substitution model, topology, branch lengths)

- Observe the likelihood of your data given this model

- Update your prior beliefs based on this to get your posterior

Prior Probability

- What is the initial weighting of models?

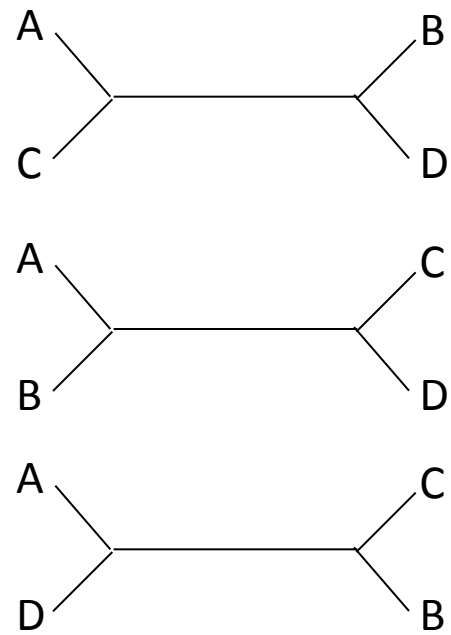


1/3

1/3

1/3

Flat



2/3

1/6

1/6

Informative

Why Bayesian?

All of the advantages of other model-based methods, plus:

(1) Explicit incorporation of prior hypotheses concerning models

(2) Calculation of posterior probabilities: the relative ‘goodness’ of models are taken into account

Bayes' rule in statistics

Likelihood of hypothesis θ

Prior probability of hypothesis θ

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

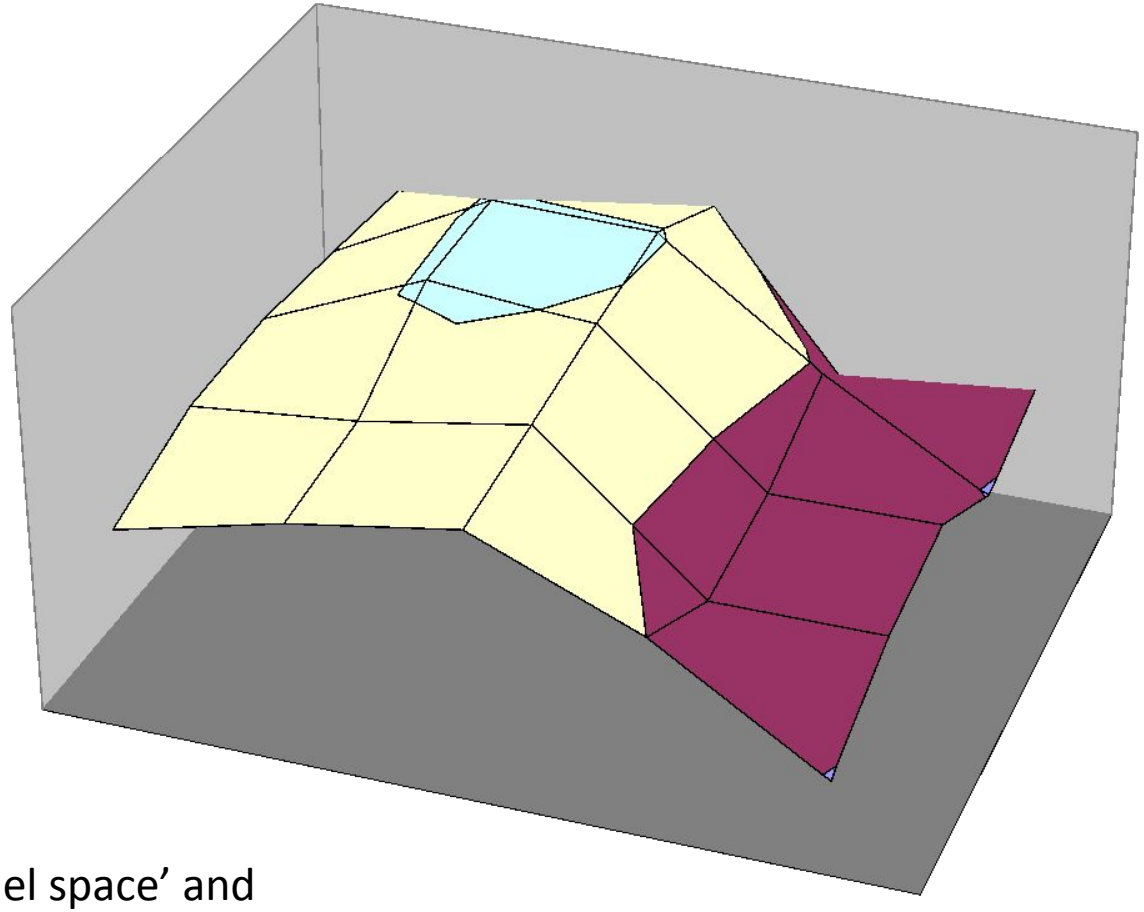
Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

The Likelihood Surface

For simple distributions (e.g. binomials for coin-flipping), we can analytically integrate over the entire likelihood function

For horrendously complex distributions (e.g. likelihoods for all trees), we cannot do this



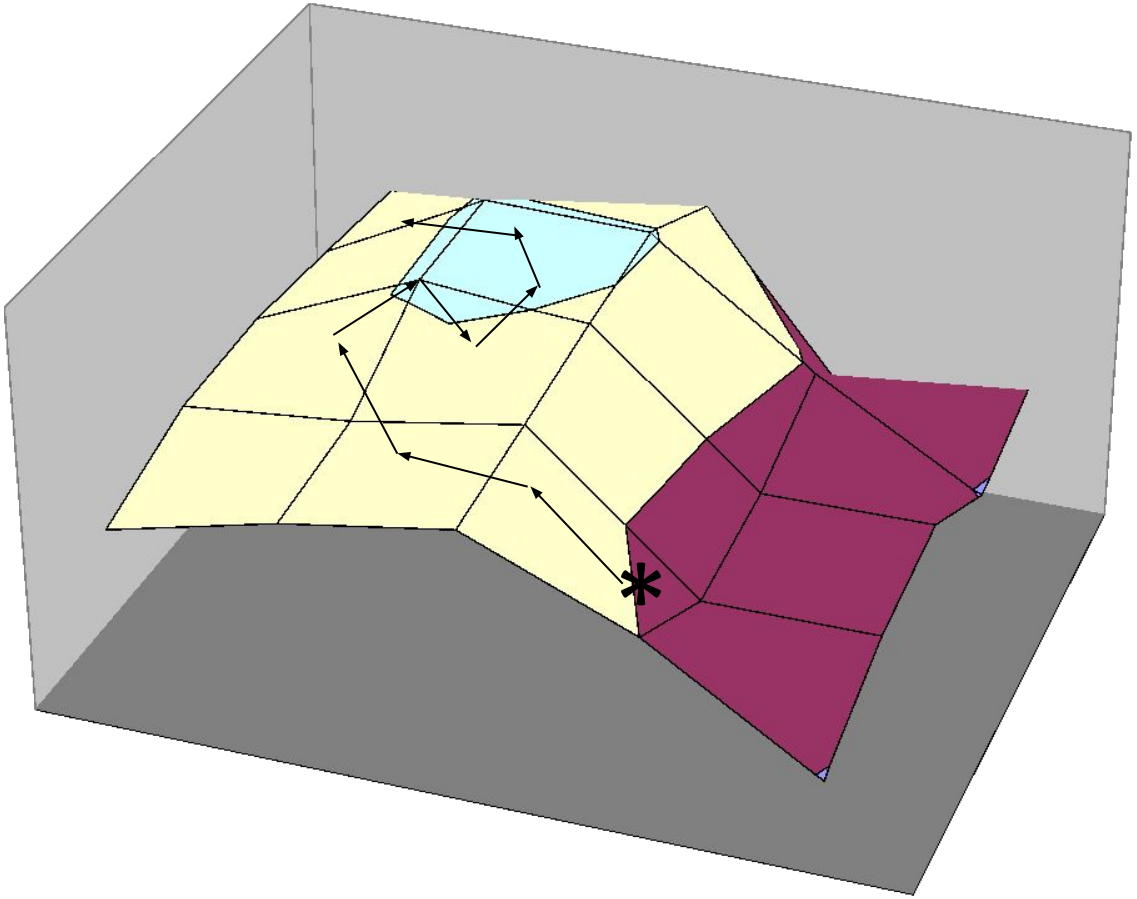
We could visit every point in ‘model space’ and evaluate the likelihood. For many datasets this is a “not before the heat death of the universe” problem

Iterative Integration

The solution is a random walk through model space

Random steps can be accepted or rejected, with a preference for steps that increase the likelihood

But we CAN DESCEND THE HILL



Why does this help?

Cancellation of marginal likelihood

When calculating the ratio (R) of posterior densities, the marginal probability of the data cancels.

$$\frac{p(\theta^* \mid D)}{p(\theta \mid D)} = \frac{\dfrac{p(D \mid \theta^*)\, p(\theta^*)}{\cancel{p(D)}}}{\dfrac{p(D \mid \theta)\, p(\theta)}{\cancel{p(D)}}} = \frac{p(D \mid \theta^*)\, p(\theta^*)}{p(D \mid \theta)\, p(\theta)}$$

Posterior odds

Apply Bayes' rule to both top and bottom

Likelihood ratio

Prior odds

High-level Difference

Maximum likelihood: optimisation method

Bayesian: sampling method

Markov chain Monte Carlo

The past does not influence the future

Keep a record of where we've been

Steps in model space are proposed randomly

Procedure

(1) Start with a random model ψ

(2) Propose a change to a new model ψ′

(3) Accept the change from ψ to ψ′ with probability

$$= \min\left[1, \underbrace{\frac{f(X \mid \Psi')}{f(X \mid \Psi)}}_{\text{likelihood ratio}} \times \underbrace{\frac{1}{1}}_{\text{prior ratio}} \times \underbrace{\frac{1}{1}}_{\text{proposal ratio}}\right]$$

(Huelsenbeck et al., 2002)

(4) Add the current tree to the growing chain

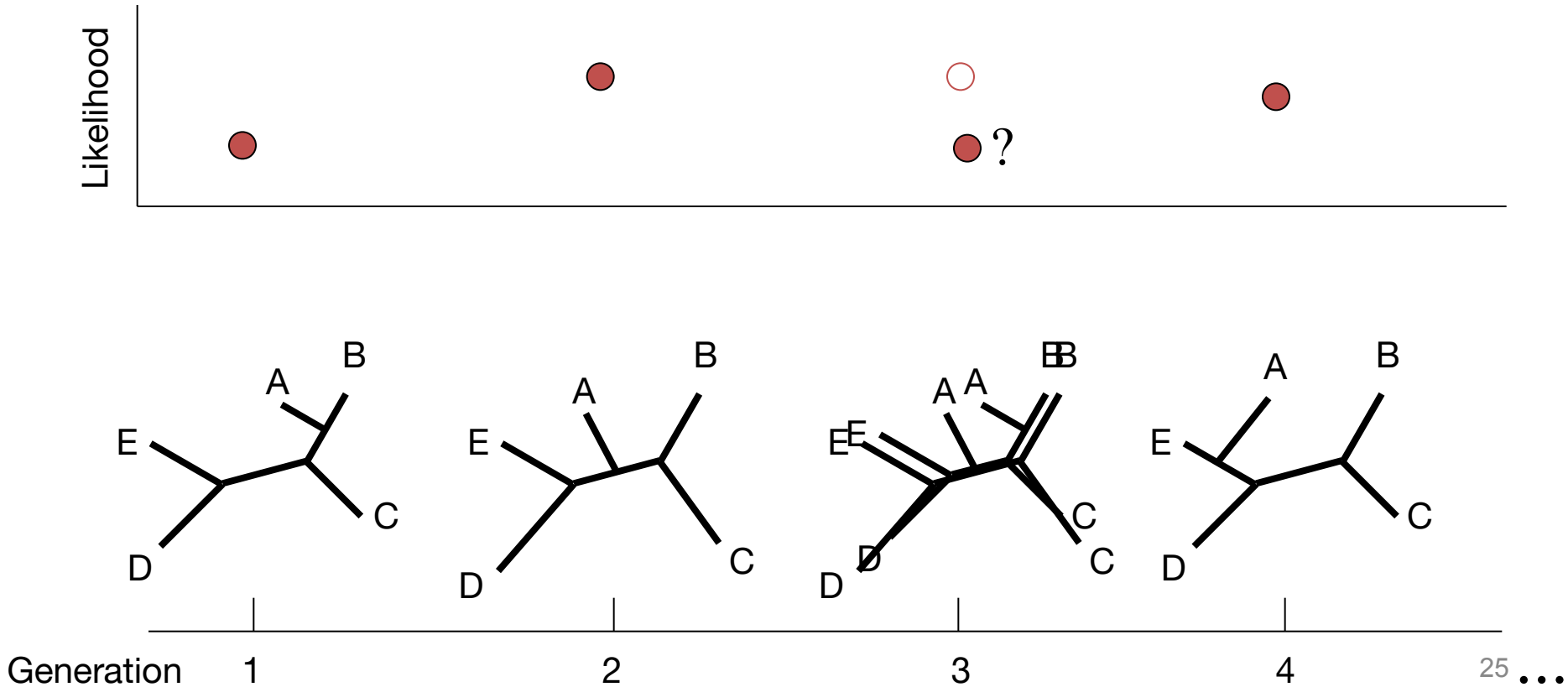
(5) Goto 2

Goto 2???

- In theory (assuming certain basic properties of the chain), MCMC will sample **every** point in likelihood space in proportion to its posterior probability

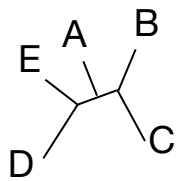
- IF the chain is run for an infinite number of iterations

MCMC in Practice

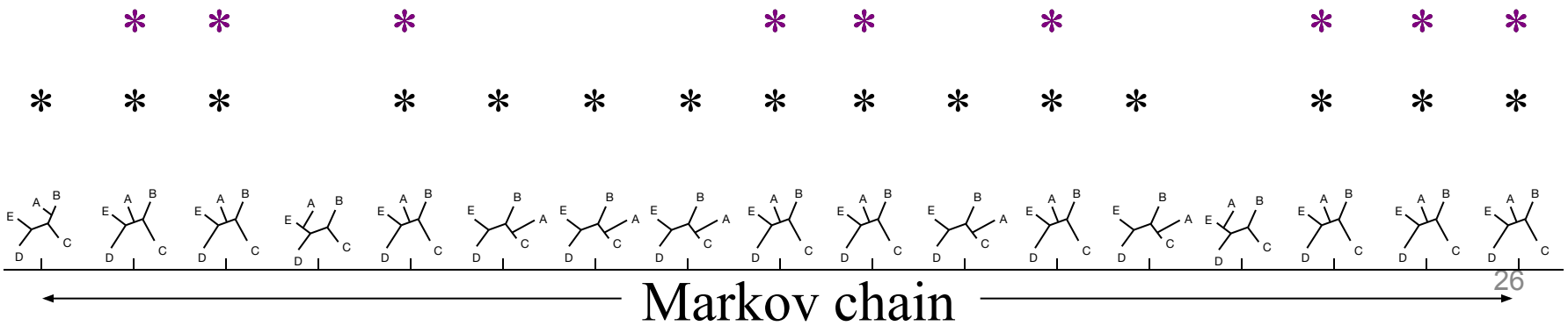


Posterior Probability

- If no *a priori* preference is given to specific trees, the posterior probability of trees and bipartitions is equal to their frequency in the Markov chain

Posterior () = 9 / 17 ~ 0.53

Posterior (ABC | DE) = 15 / 17 ~ 0.88



Posteriors on TREES

- Simply the frequency of the tree (integrated over all branch lengths) in the Markov chain

- Addresses all phylogenetic hypotheses at once

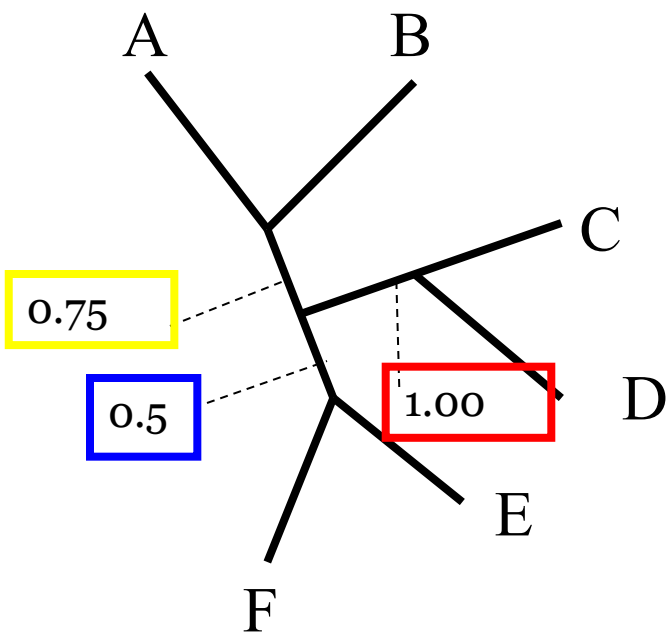
Posteriors on TREES

- Can be very unstable!
- Practical example:
  - 30-sequence alignment
  - 3,000,000 iteration chain
  - 30,000 trees saved in chain (1 / 100 thinning)
  - >25,000 different trees!

- Most-frequent tree sampled twice, so posterior = (2/30,000)

Posteriors on SPLITS

- Far more stable (independent evaluation of tree features)
- Lose information about dependencies within tree



Interpreting Posteriors

- ‘Confidence intervals’ of models
  - Rank the models in decreasing order of PP, and take the set that corresponds to the top x% (e.g., the top 95%)
  - May include multiple trees or splits, but will certainly EXCLUDE a lot more

Interpreting Posteriors

- Bayes factors
  - The ratio of posterior probabilities for two hypotheses (models) H_1 and H_2

$$\frac{P(H_1)}{P(H_2)} = B(x)$$

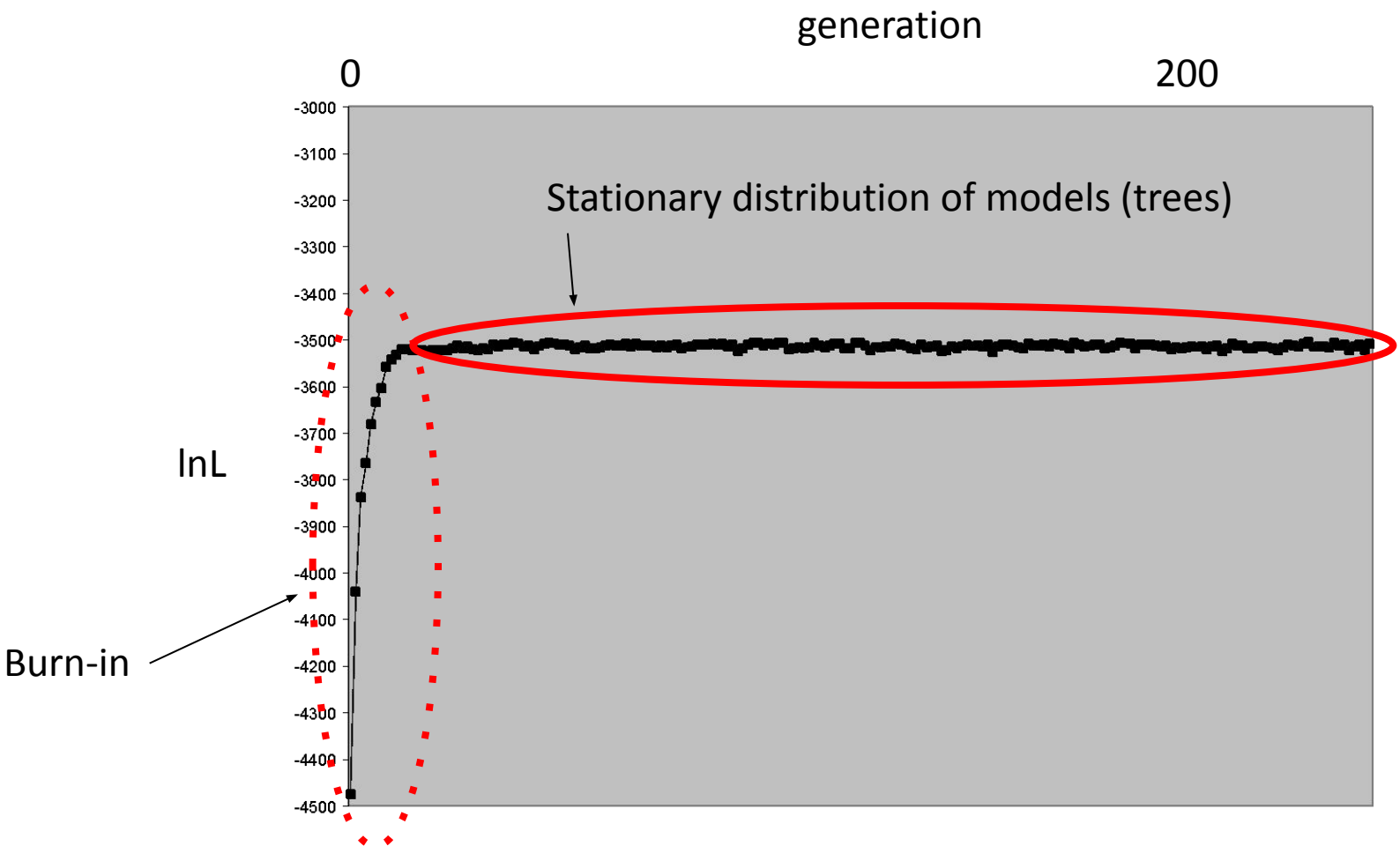
Different rules of thumb for evaluating Bayes factors (see e.g.
Jeffreys, H. (1961). Theory of Probability. Oxford: Clarendon Press.)

For instance:

B(x)	Interpretation
1-3	Barely worth mentioning
3-10	Moderate preference
10-100	Strong preference
100+	Overwhelming(!) preference

Markov chains in action!

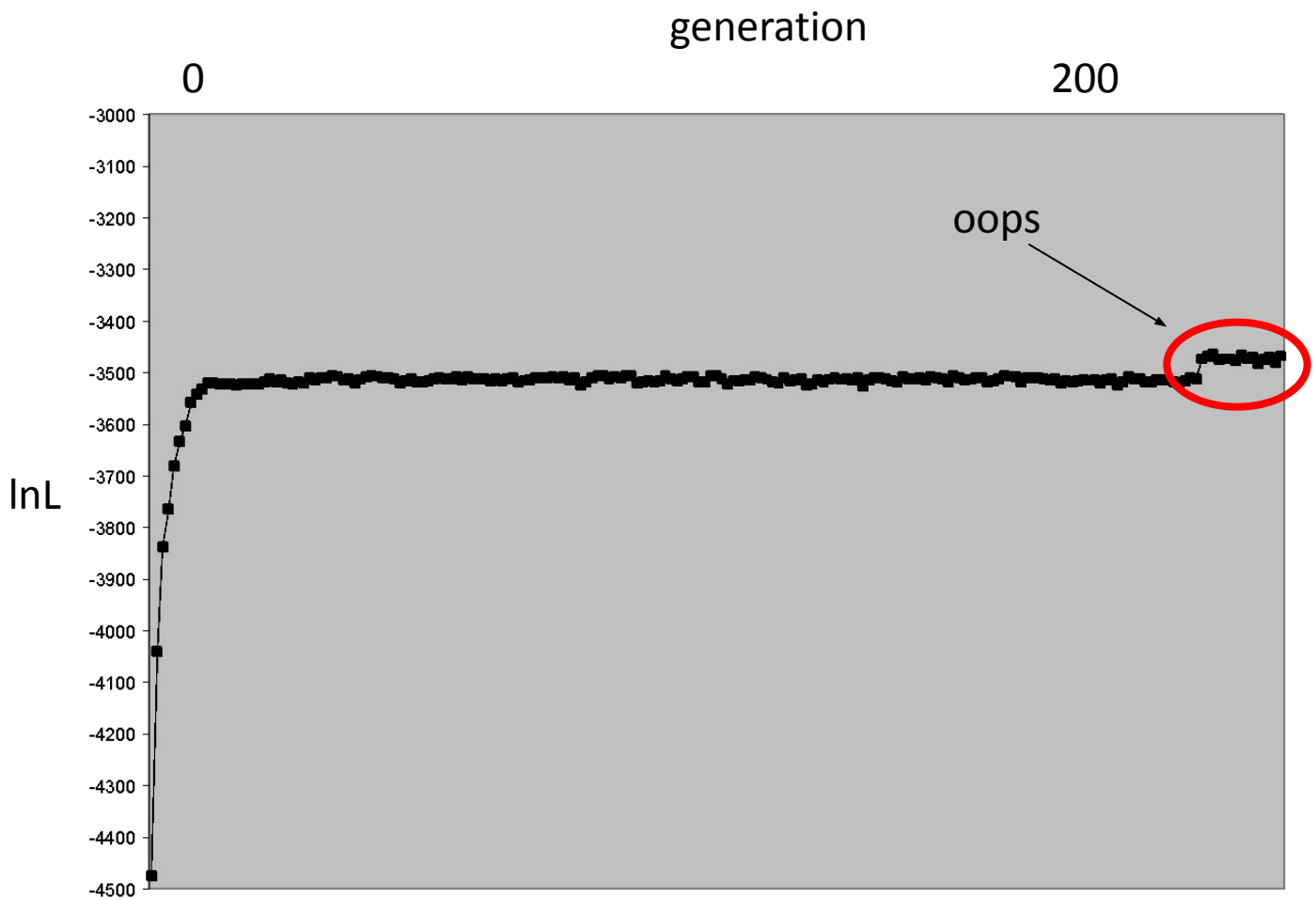
- Evaluate progress using e.g. a log-likelihood plot



Markov chains in action!

- However, problems can arise

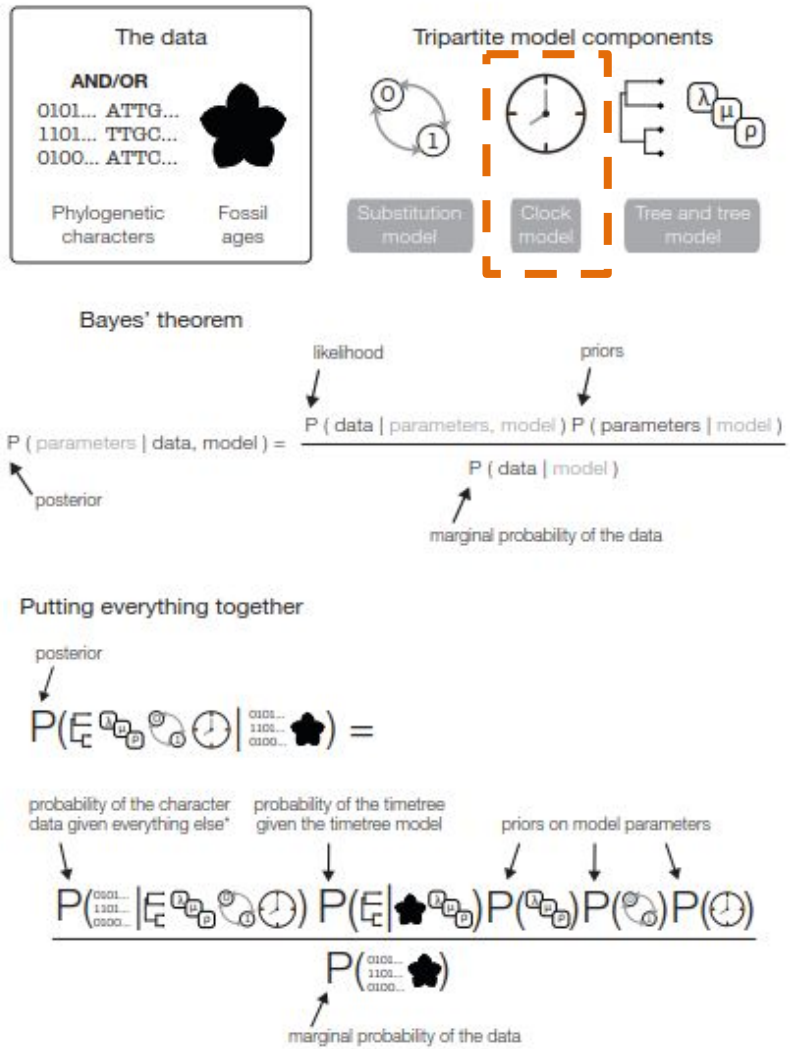
The chance of this happening increases with increasing model complexity (More parameters to worry about!)



A couple of solutions

- Metropolis-coupled MCMC: heated chains
  - Cold chain: collects samples
  - Heated chains are more likely to accept bad moves
  - Chains can SWAP

Estimating divergence times



Warnock and Wright, *EcoarXiv*

OPEN ACCESS Freely available online PLOS BIOLOGY

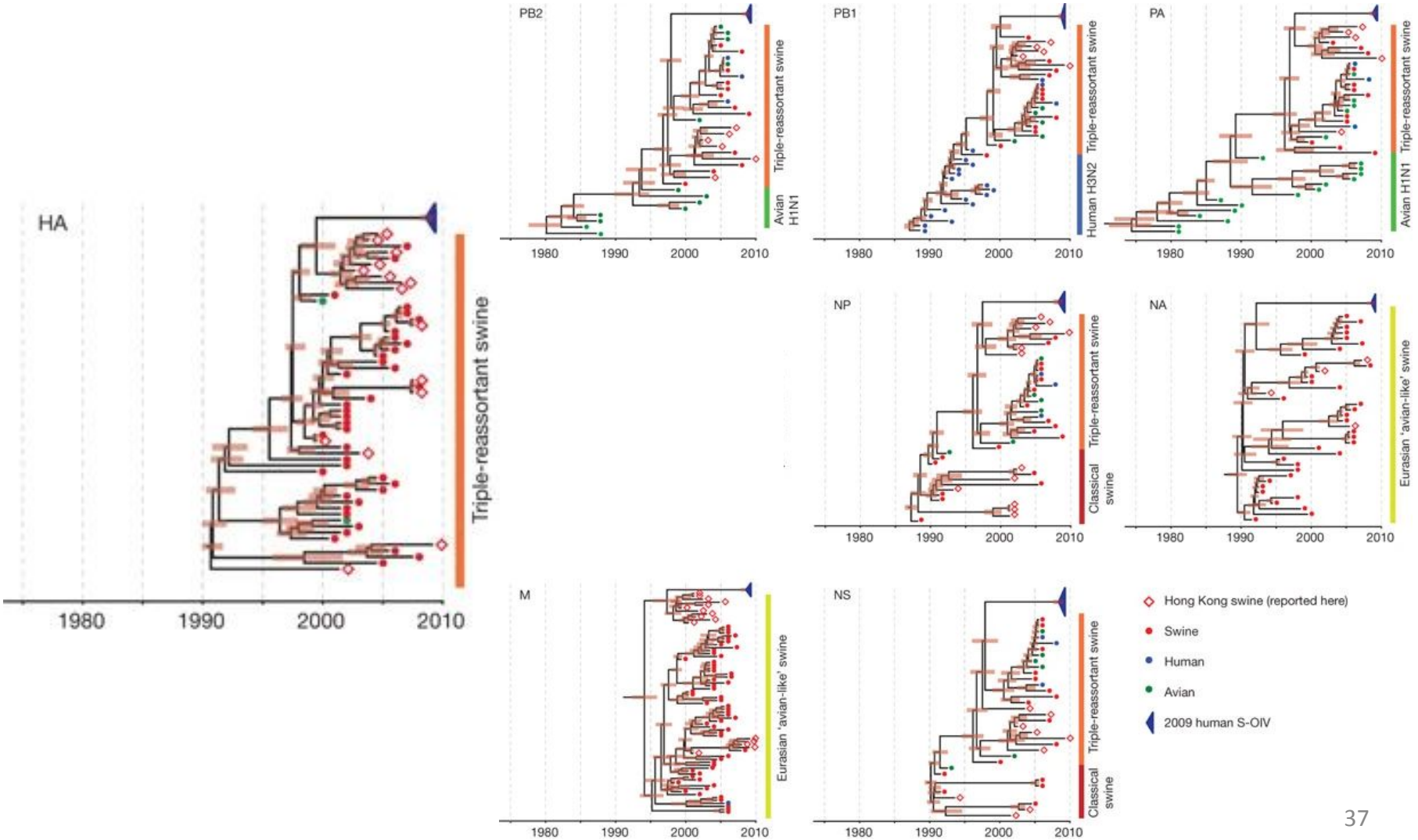
Relaxed Phylogenetics and Dating with Confidence

Alexei J. Drummond, Simon Y. W. Ho, Matthew J. Phillips, Andrew Rambaut*

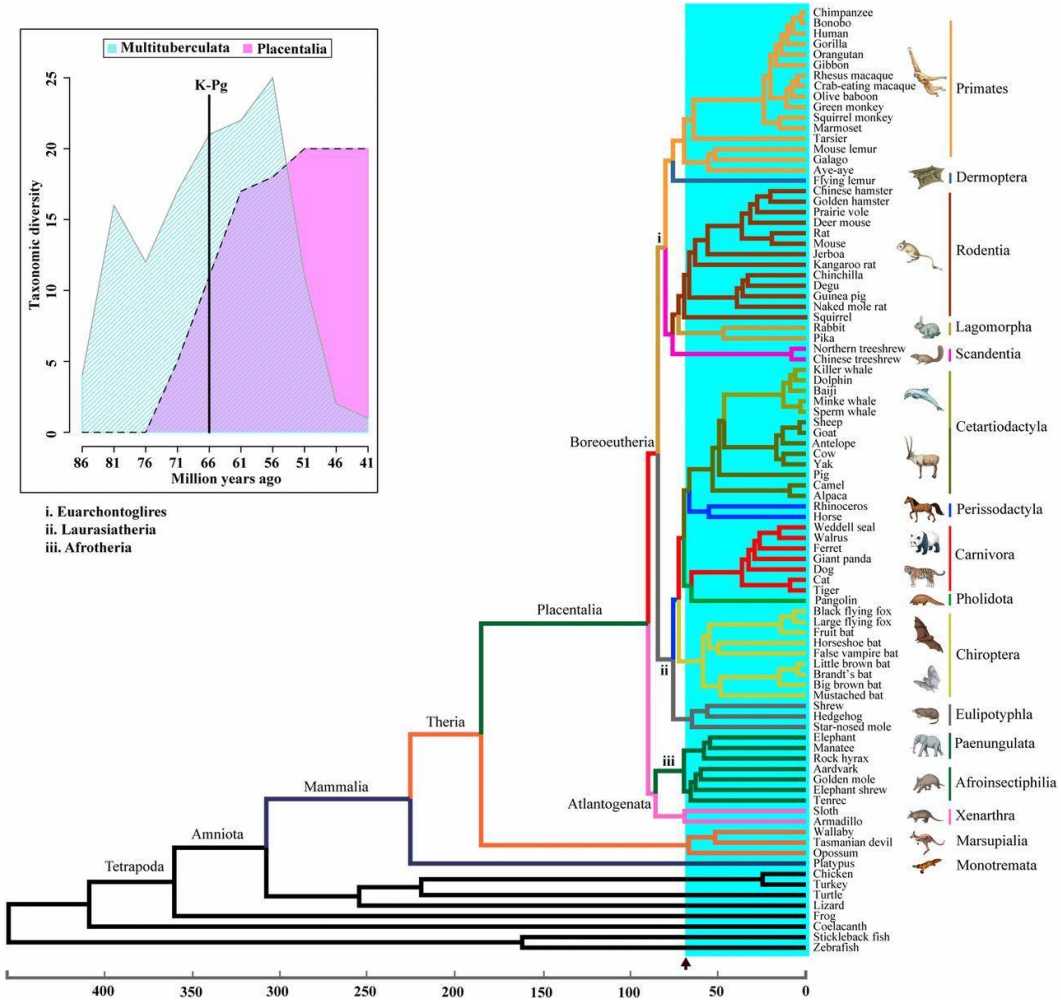
Department of Zoology, University of Oxford, Oxford, United Kingdom

In phylogenetics, the unrooted model of phylogeny and the strict molecular clock model are two extremes of a continuum. Despite their dominance in phylogenetic inference, it is evident that both are biologically unrealistic and that the real evolutionary process lies between these two extremes. Fortunately, intermediate models employing

Estimating divergence times (recent) using samples at or near internal nodes



Estimating divergence times (ancient) using the fossil record for calibration



Liang Liu et al. PNAS 2017;114:35:E7282-E7290

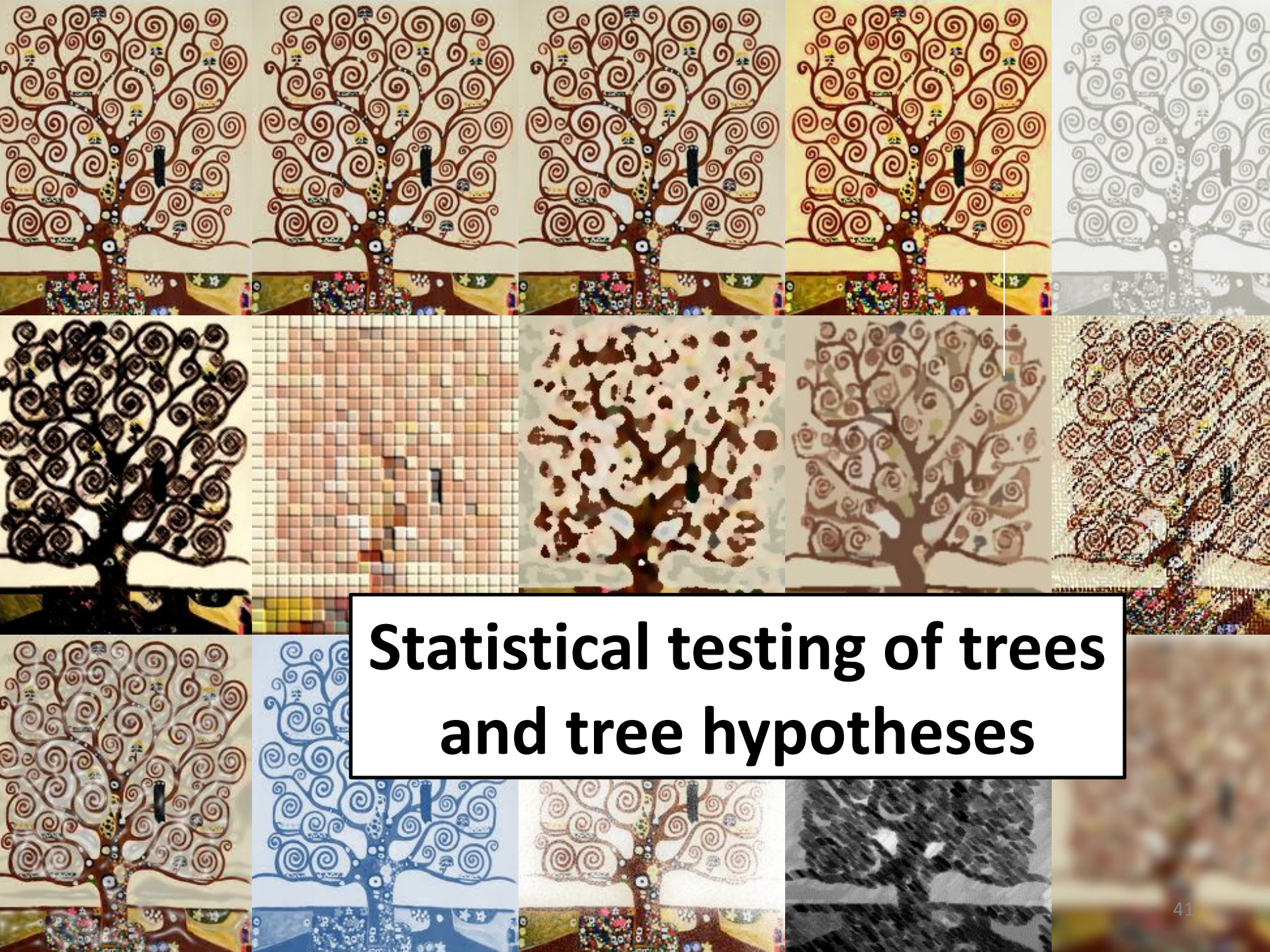
See

- RevBayes https://revbayes.github.io/tutorials/
- BEAST https://beast.community/

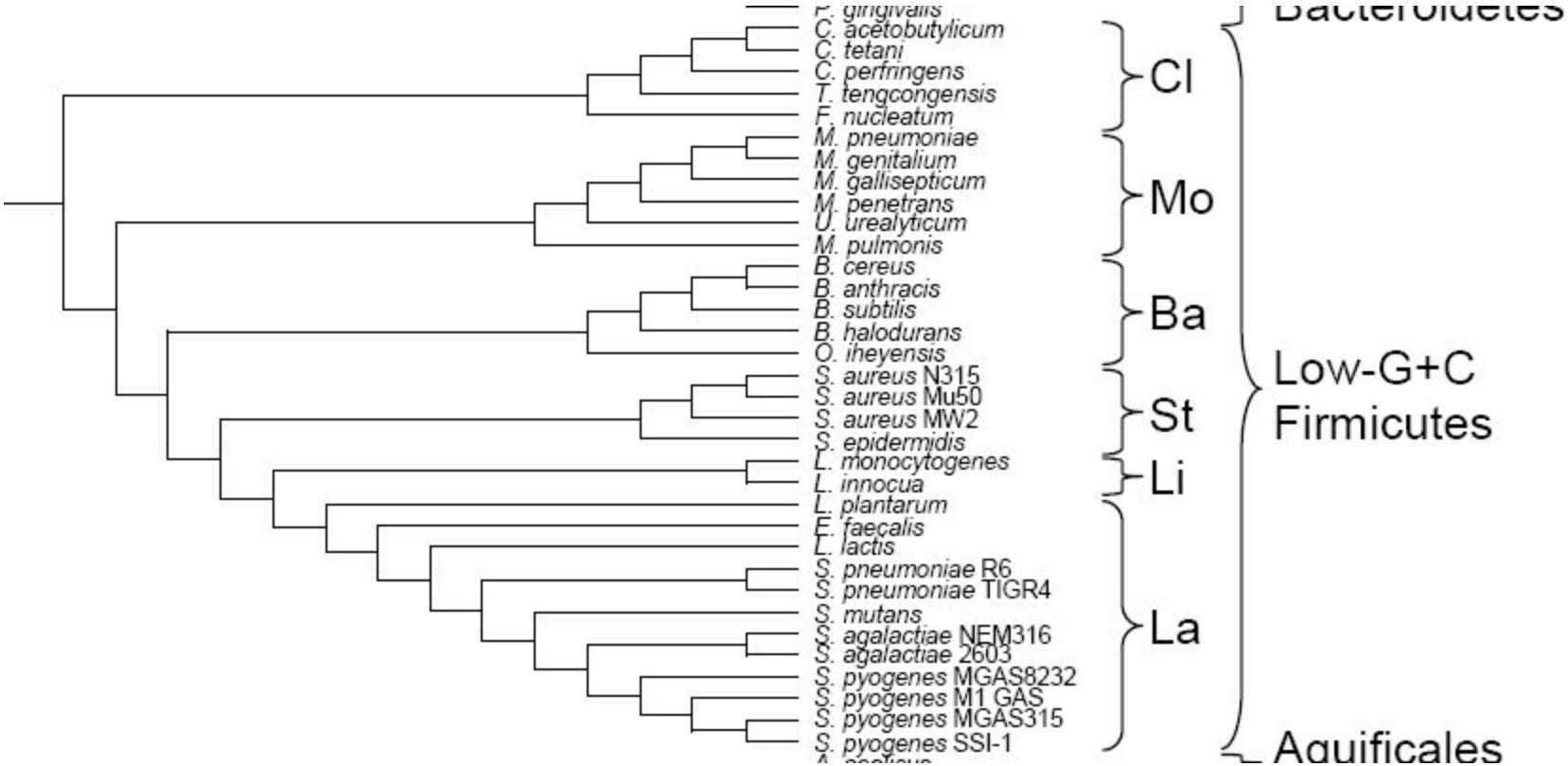
Conclusion

- ML is currently the most widely-used method for phylogenetic inference
  - It is computationally expensive

- Bayesian methods can take a long time but give you a probability distribution across trees, rather than simply the best* tree
  - If you can parameterize it, you can sample from it!



Statistical testing of trees and tree hypotheses



Is a *hypothesis*

…but what is the strength of support for this hypothesis?

Significant Significance Questions

1. Do the data (that's usually the alignment) strongly support the relationships in the tree?

2. Is the recovered tree statistically better than all other possible trees?

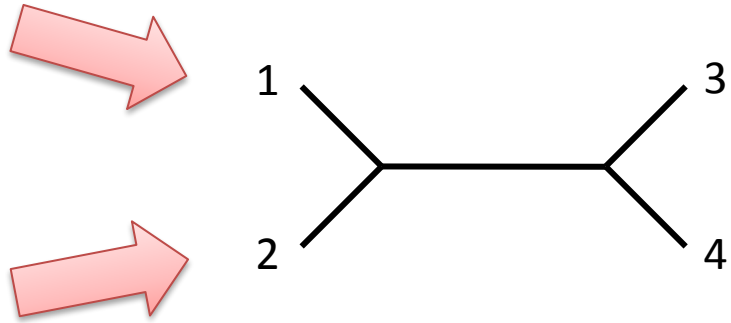
3. Is a *tree* really the best explanation of the data?

Why ask these awkward questions?

Ask for a tree, get a tree

```
1 ACCGAGCAA
2 ACCGAGCAA
3 ACCGAGCAA
4 ACCGAGCAA
```

```
1 ACCGAATGA
2 ACCGAGCAG
3 GTTAGGCAG
4 GTTAGATGA
```



Problems with datasets

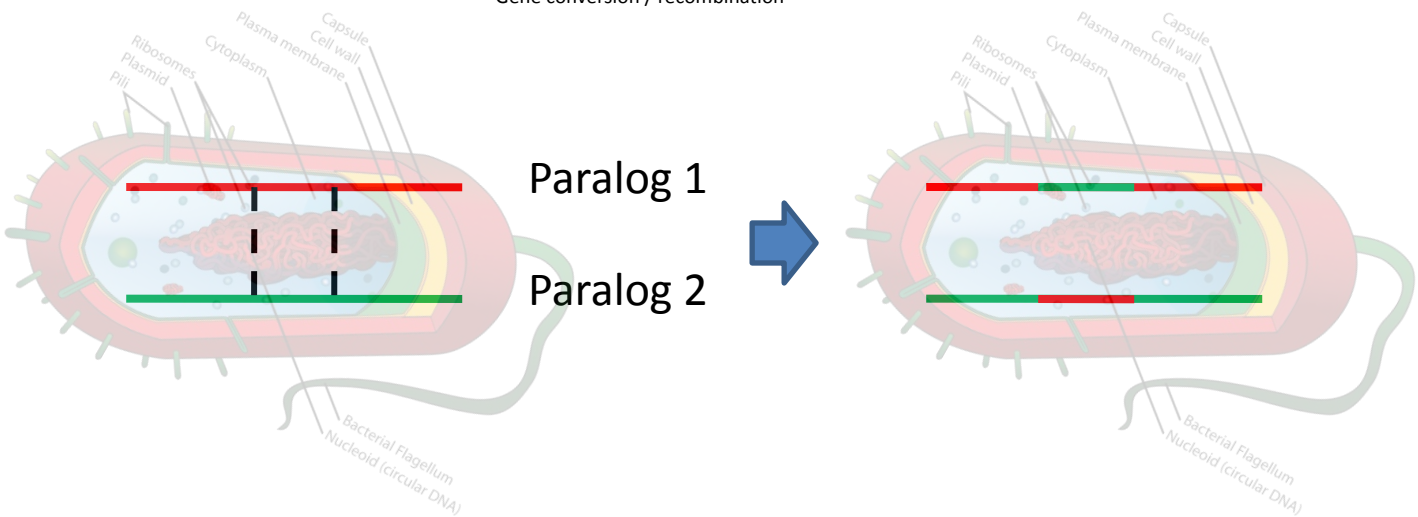
- Signal saturation – too many substitutions (and multiple substitutions!) between sequences

- Lack of signal – some short branches in the tree may lack supporting data or be sufficiently ancient to have been erased

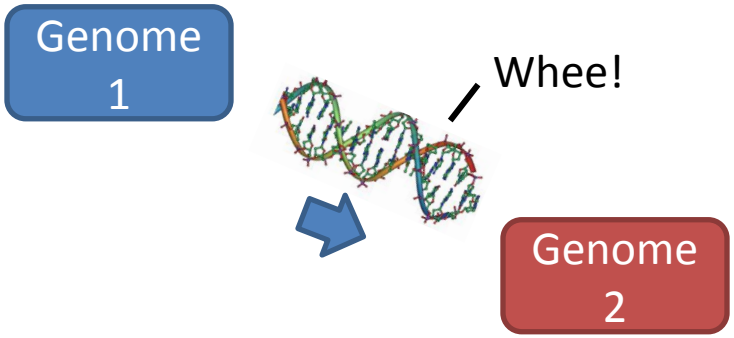
- Misleading signals may be relatively strong

Reticulate evolution

Gene conversion / recombination



Lateral gene transfer (one or more genes)



Addressing significance questions

1. Strength of support – resampling, subsampling, and simulation

2. Better than alternatives – Bayesian, paired-site comparisons

3. Treelike signal – phylogenetic networks


Tree support

Basic Principles

- Resample from the distribution of data points (alignment columns) and see whether we get the same answer
- Do this a bunch of times (100-ish)
- Map the results onto the **original** tree

The nonparametric bootstrap test

- Resample **with replacement** from the original population

- Original alignment: *n* columns
- Bootstrapped alignment: still *n* columns

- **But** some columns will be missing, and some will be present more than once

Alignment

```
   123456
A  catcga
B  ccgggt
C  gcggga
D  gaacgt
```

Resample the alignment

Bootstrapped alignments

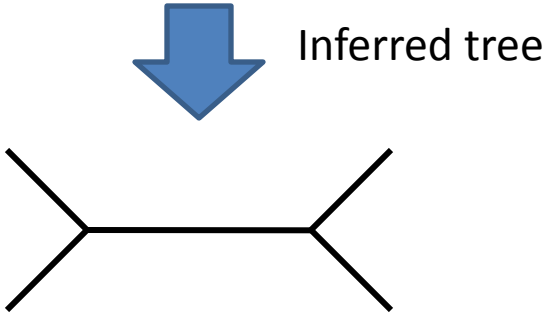
Trees

515621

364122

615343

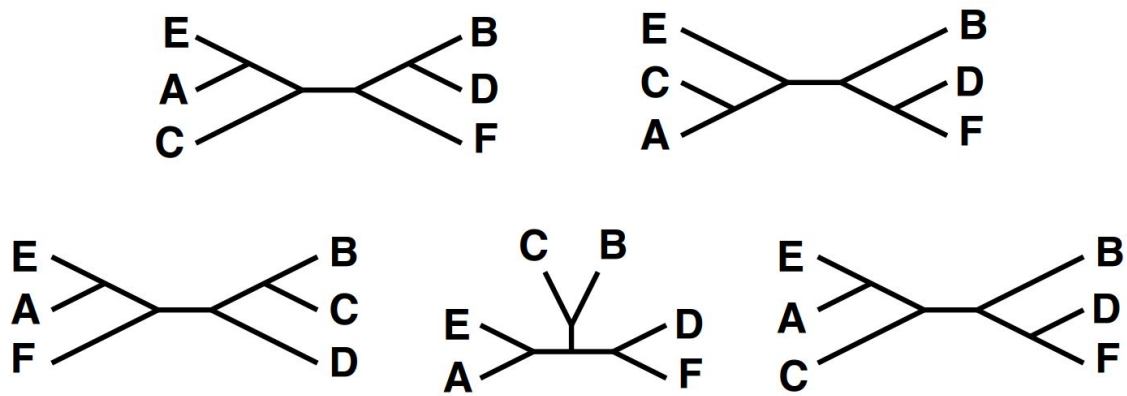
414436



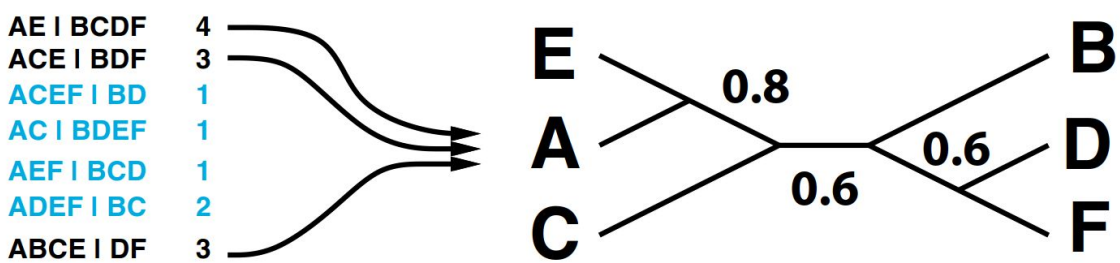
Inferred tree

The majority-rule consensus tree

Trees:

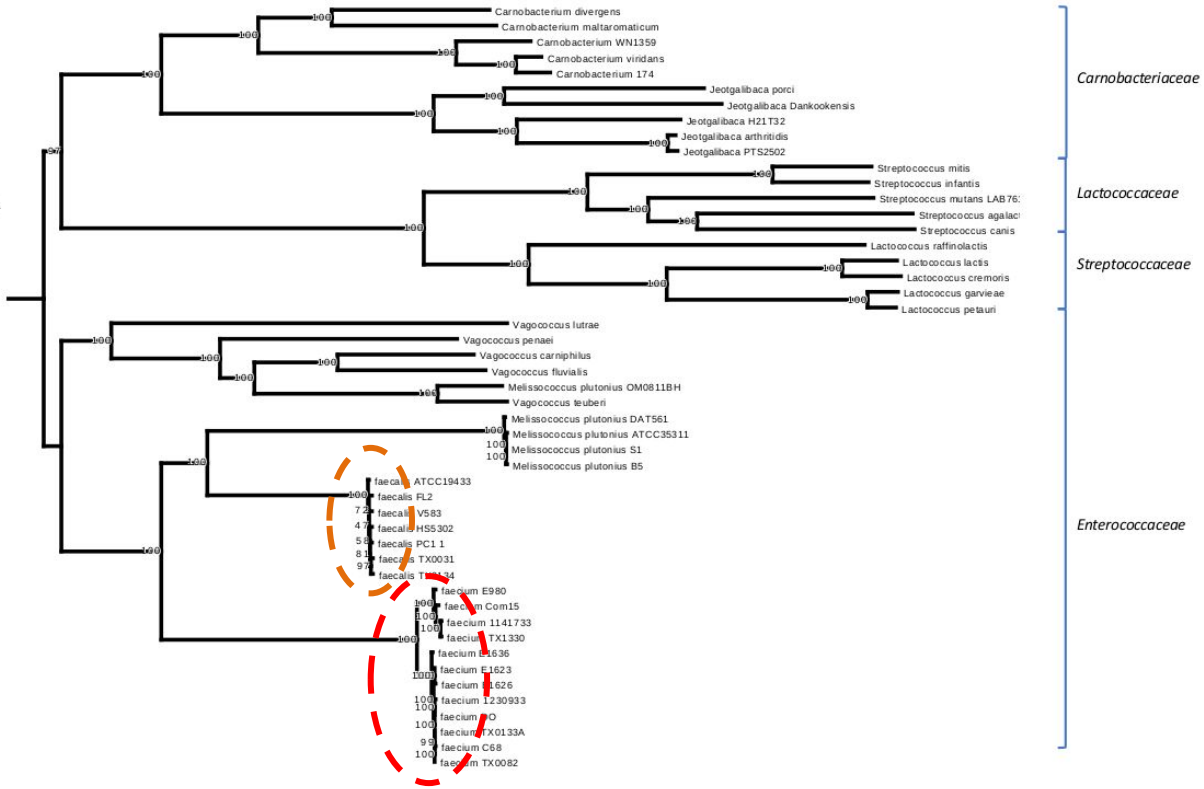


How many times each partition of species is found:



Slide from Joe Felsenstein

Support for tree features



Map bootstrap values onto the original tree

The bootstrap for a given grouping of taxa in the tree (supported by an edge) is equal to the frequency that grouping is observed among the bootstrap replicates

70% is often used as a support criterion (based on simulation)

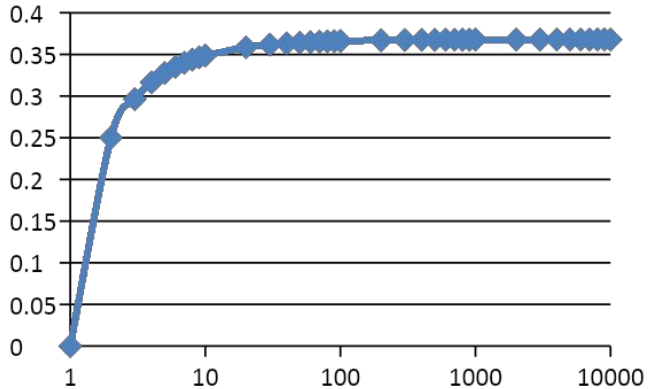
~ 100% (complete support)

~ 50% (much weaker support)

What is the bootstrap doing?

- The bootstrap is randomly reweighting characters in the alignment, and assessing the impact on the phylogeny

- The probability of a given character being excluded (weight = 0) is equal to $(1 - 1/N)^N$



Asymptote ≅ 0.36788

What is the bootstrap doing?

- The goal of the bootstrap is to simulate an infinite population (number of alignment columns) by considering a range of reweightings on the existing data

Limitation of nonparametric methods

- The (nonparametric) bootstrap method you have just seen are <u>limited</u> by the availability of reliable data

- This resampling procedure may therefore not cover the range of alternatives

- The **parametric bootstrap** <u>simulates</u> data on the proposed tree, and determines how often that tree can be recovered (**COMPLEX…**)

The nonparametric bootstrap is *slooooooow*

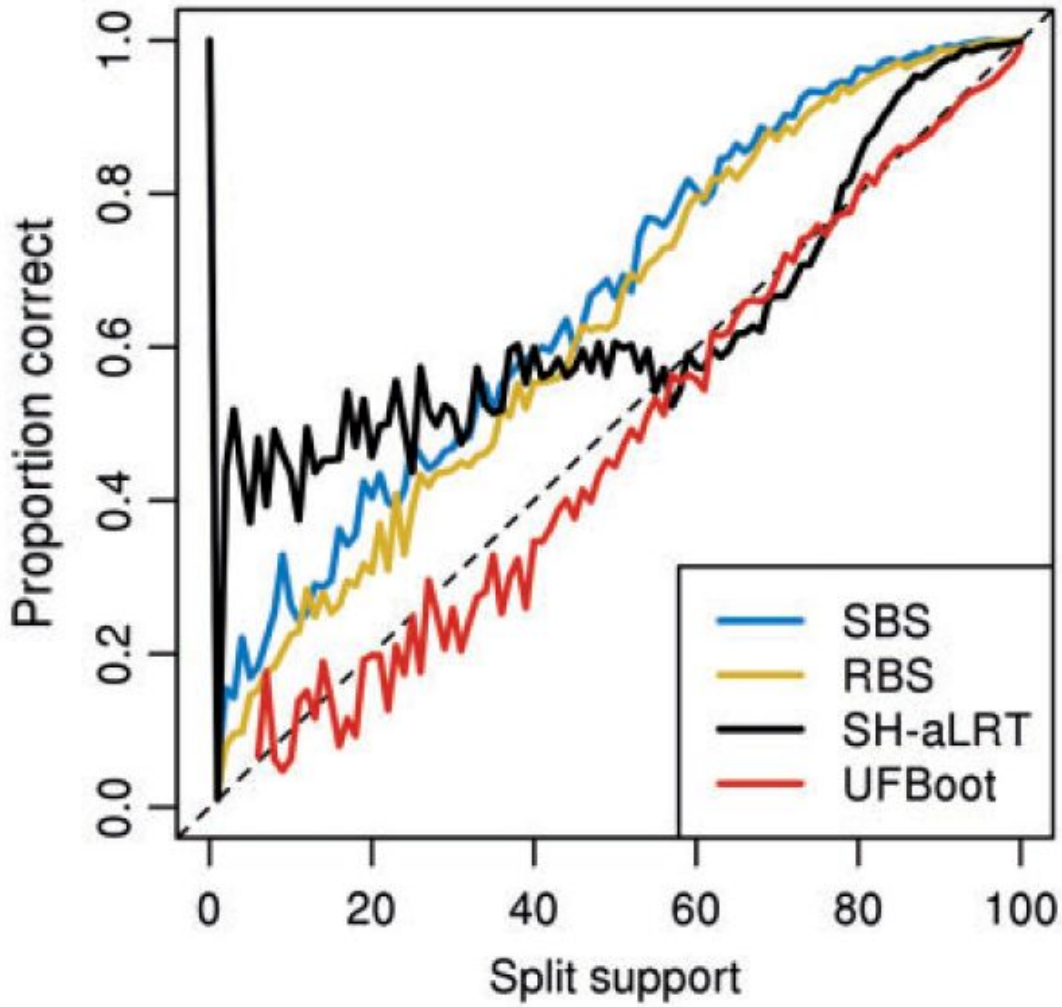
https://en.wikipedia.org/wiki/Snail#/media/File:Snail.jpg

Alternatives to doing the likelihood search 100 times

- aLRT: Estimate *local* support using e.g. NNI and re-use likelihoods (since the bootstrap replicates are just the same columns re-weighted)
- SH-aLRT: use simulations to generate a realistic distribution of likelihoods
- Ultrafast bootstrap: Perform the search for all bootstrap replicates *simultaneously*. Keep a record of the best tree for each bootstrap replicate, and update as better trees are found

Minh et al. (2013) *Mol Biol Evol*

Accuracy matters



Minh et al. (2013) *Mol Biol Evol*

Problems with resampling in general

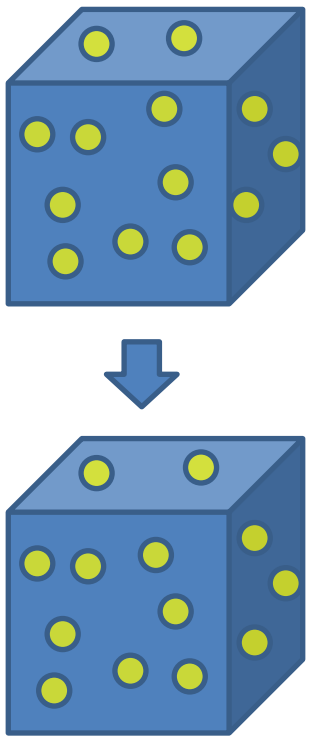
- Limited to asking the question, “to what extent do the data support the tree”?

- Do not directly address issues of:
  - Second-best trees
  - Bias in methods including model misspecification
  - Non-tree-like signal

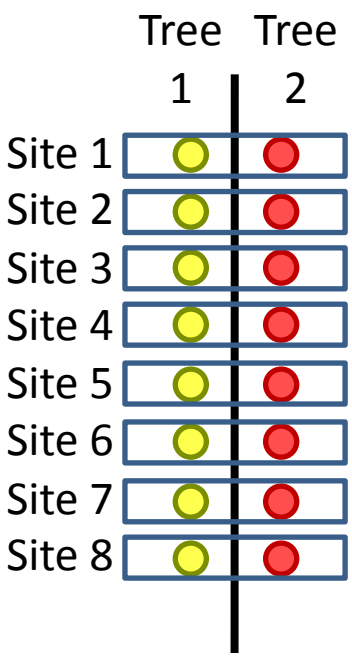
Best tree?

Is the best tree better than some other tree?

- We need to approach the data somewhat differently



So far – reshuffle data, but only infer results from complete data sets

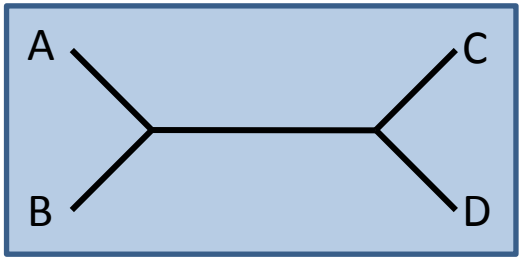


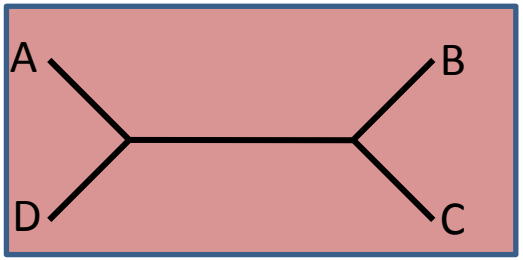
Now – compare individual sites to come up with an overall conclusion of significance

Basic principles

- For two trees, compare the fit at each alignment site either **quantitatively** or **qualitatively**

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2





The winning sites test: “An up-or-down vote”

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						

4 sites favour the red tree
2 favour the blue tree

Use the **binomal distribution** to assess the significance of this difference

$$\binom{n}{k} p^k (1-p)^{n-k}$$

What is the probability that 4 or greater coin tosses will come up with the same result?

Need to evaluate the above formula for n=6, k=0,1,2,4,5,6 (two-tailed)

4 out of 6: $p = 0.6875$ (not significant)

40 out of 60: $p = 0.0124$
(significant at threshold of 0.05)

400 out of 600: $p = 2.3 \times 10^{-16}$

Paired *t* test

	1	2	3	4	5	6
A	c	a	t	c	c	t
B	c	c	g	g	g	t
C	g	c	g	g	g	a
D	g	a	a	c	g	t
Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

Here we consider the mean and variance of differences across all sites

Favours?						
By how much?	5.2	3.1	0.9	6.6	0.3	0.2

Mean of differences: (-5.2 + 3.1 + 0.9 + 6.6 + 0.3 – 0.2) / 6 = 0.916

Variance: 15.22

We compute a ***t statistic*** using the following formula:

$$t = \frac{\bar{x}}{\mathrm{var}}\sqrt{N} \qquad = 0.148$$

Compare to the *t* distribution for 5 degrees of freedom
p = 0.888

Paired sites vs t-test

the influence of small differences

These tests are very biased

- Statistical tests generally assume a *random sample*
- The (distributions of) trees we want to test are most definitely not!
- Less-biased tests often depend on more-sophisticated comparisons and (again) simulation

Summary

- Your trees may look great but be unsupported by the data
- Bootstrap tests: *How strong is the support for my tree?*
- Statistical comparisons of trees: *How much better is this tree than that tree?*