

# **A Statistical Approach to the Estimation of Phylogeny from Genome Arrangements**

**Bret Larget**

Departments of Statistics and of Botany

*University of Wisconsin - Madison*

November 25, 2003

# Background

---

- **Problem.**— We have [genome arrangement data](#) in the form of linear or circular gene orders for several [taxa](#) and we want to analyze it to learn something about their evolutionary relationships.
- **Data representation.**— Genome arrangements can be represented by [signed permutations](#).
- Assume that each genome has exactly one copy of each gene under study.
- Assume that [gene inversion](#) is the sole mechanism of genome rearrangement.

# Gene Inversions and Signed Permutations

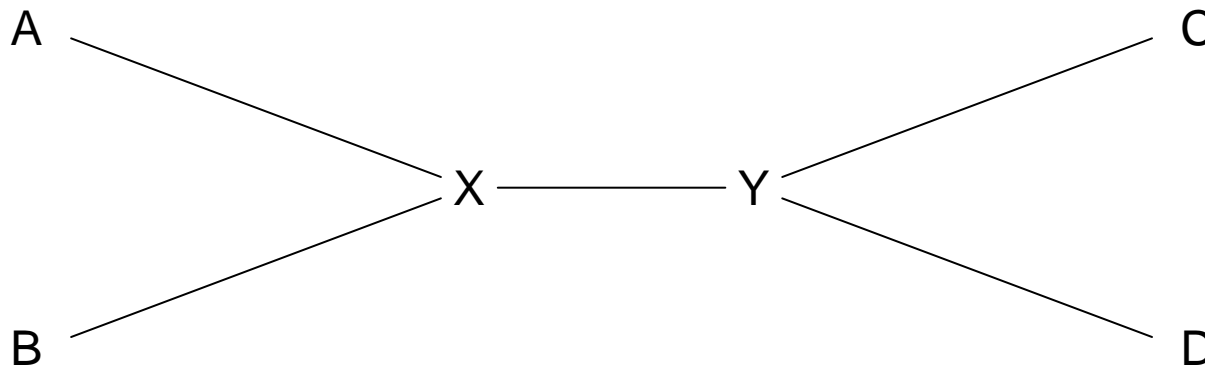
---

In the signed permutation representation, a single gene inversion corresponds to **reversing the order** and **changing the sign** of a contiguous part of the permutation.

1	2	-4	5	-3	-6
			↓		
1	2	-4	3	-5	-6

# The Multiple Genome Rearrangement Problem

- The **multiple genome rearrangement problem** is a **maximum parsimony** method — find the simplest explanation of the data.
- In this setting, the objective is to search for the **unrooted tree topology** and **ancestral arrangements** at internal nodes so that the sum of the **inversion distances** over branches is minimized.



# Example Data

---

The following table shows the non-tRNA mitochondrial genome arrangements for four species labeled relative to the human genome.

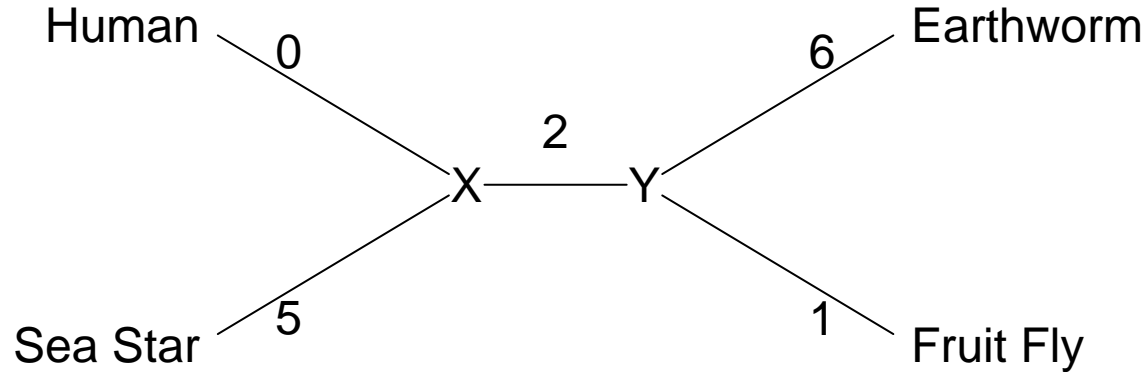
Taxon	Arrangement
Human	(1–14)
Sea star	(6) (1–5) (7–11) ( <u>12</u> ) ( <u>14–13</u> )
Earthworm	(1–2) (4) ( <u>9</u> ) (10) (3) (8) (6–7) (11–13) (5) (14)
Fruit fly	(1–5) ( <u>8–6</u> ) ( <u>9</u> ) (10) ( <u>13–11</u> ) (14)

	H	SS	E	F
Human	0	5	7	3
Sea star	5	0	11	7
Earthworm	7	11	0	7
Fruit fly	3	7	7	0

# Maximum Parsimony Solution

---

The most parsimonious reconstruction requires 14 gene inversions.



# What is the Standard Error?

---

- The most parsimonious solution is a **point estimate**, but the solution does not provide an estimate of uncertainty.
- **How much evidence** is there that the most parsimonious tree is the true tree?
- **How confident should we be** in the reconstructed ancestral sequences?
- **How sure are we** that the true evolutionary history had 14 gene inversions?

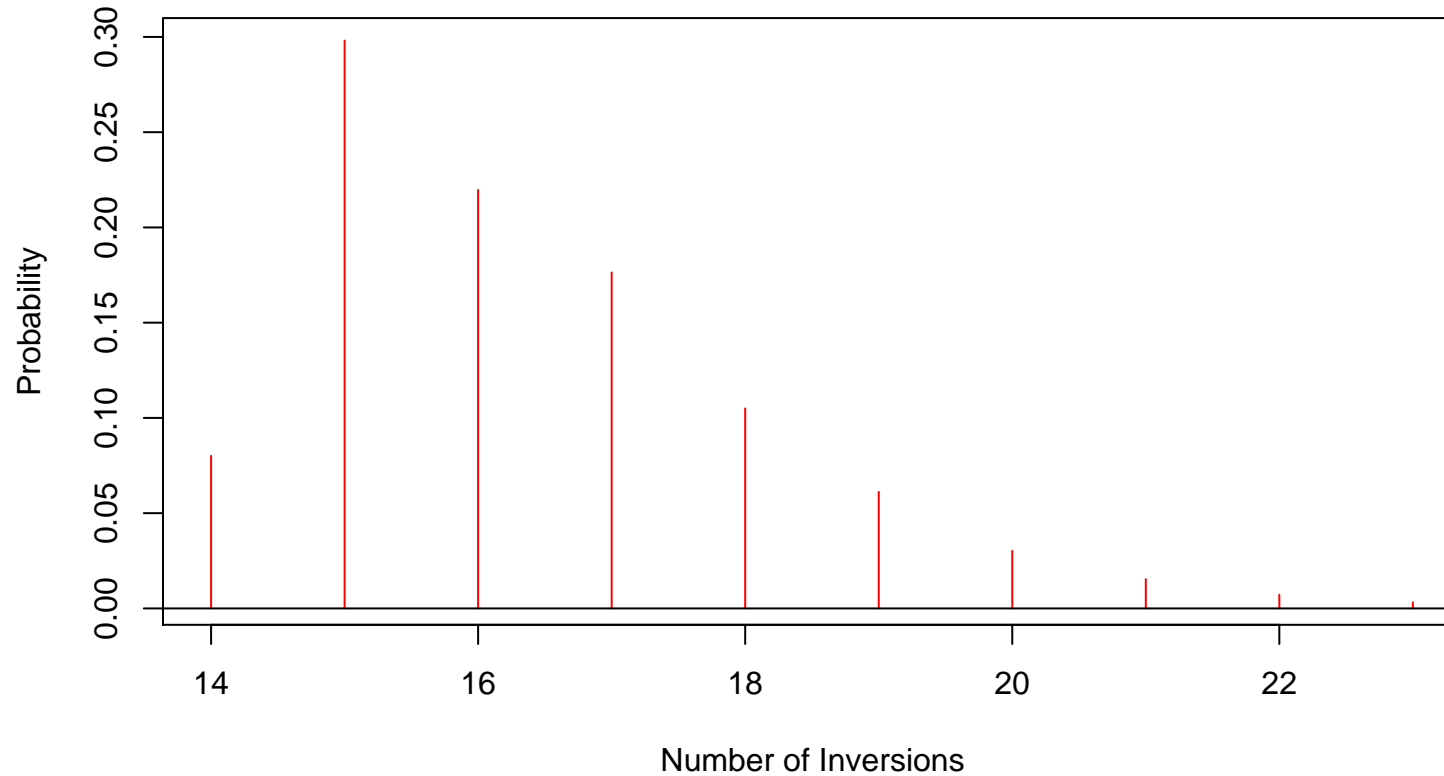
# A Statistical Approach

---

- A Bayesian approach to the problem is based on a posterior probability distribution on the space of all possible reconstructions.
- We can say (with assumptions to be specified later) that:
  1. There is an 87% chance that the displayed tree is correct;
  2. There is only an 8% chance that there were 14 gene inversions and a 92% chance that there were more;
  3. There is a 50% chance that the ancestor of humans had the same genome arrangement as humans.

# Posterior Distribution of Number of Inversions

---



# A Single Branch Model for Genome Rearrangement

---

- Give the branch a random, Gamma distributed length.
- Choose a number of inversions from a Poisson distribution whose mean is the branch length.
- (The unconditional number of inversions on the branch has a negative binomial distribution and is over-dispersed relative to the Poisson distribution.)
- Place these inversions on the branch at uniformly at random.
- For each realized inversion, independently select the genes to be inverted uniformly at random from the set of all possible gene inversions.
- Apply the inversions in order to the arrangement on one end to determine the arrangement on the other end.

# The Model Extended to a Tree

---

- All unrooted tree topologies from a set are equally likely.  
 $\tau \sim \text{Uniform}(\text{Tree topologies with } n \text{ leaves})$   
( $n$  leaves implies  $m = 2n - 3$  branches.)
- Branch lengths are independent Gamma random variables.  
 $\mu_1, \dots, \mu_m \sim \text{i.i.d. Gamma}(\alpha, \lambda)$
- Given the tree topology and branch lengths, place inversions down with a Poisson point process on the tree.
- Conditional distribution of # of inversions is Poisson.  
 $X_i | \mu_i \sim \text{Poisson}(\mu_i), \quad \text{for } i = 1, \dots, m$
- Realized inversions are independent and equally likely, locations on branch are uniformly distributed.  
 $r_{ij} | X_i \sim \text{Uniform}(\text{all inversions}), \quad 1 \leq i \leq m, \quad 1 \leq j \leq X_i$   
 $t_{ij} | X_i, \mu_i \sim \text{Uniform}(0, \mu_i), \quad 1 \leq i \leq m, \quad 1 \leq j \leq X_i$

# Posterior Distribution on Tree

---

$$\begin{aligned}\mathbb{P}\{\text{Tree} \mid \text{Data}\} &= \\ &= \frac{\mathbb{P}\{\text{Data} \mid \text{Tree}\} \times \mathbb{P}\{\text{Tree}\}}{\mathbb{P}\{\text{Data}\}} \\ &= \frac{\sum_{\text{Histories}} \mathbb{P}\{\text{Data and History} \mid \text{Tree}\} \times \mathbb{P}\{\text{Tree}\}}{\mathbb{P}\{\text{Data}\}} \\ &= \frac{\sum_{\text{Histories}} \mathbb{P}\{\text{Data and History} \mid \text{Tree}\} \times \mathbb{P}\{\text{Tree}\}}{\sum_{\text{Trees}} \sum_{\text{Histories}} \mathbb{P}\{\text{Data and History} \mid \text{Tree}\} \times \mathbb{P}\{\text{Tree}\}}\end{aligned}$$

# Data Augmentation

---

- We do not know how to sum directly over the infinite number of rearrangement histories compatible with the observed data for each tree topology.
- We can integrate out branch lengths analytically.
- We **augment** the state space to include the ordered list of gene inversions on each branch.
- By conditioning on this information as well, we can calculate the probability easily.
- We use MCMC to sample from the augmented space of tree topologies and ordered lists of gene inversions.

# Posterior of Augmented Space

---

$$\mathbb{P}\{\text{Tree and History} \mid \text{Data}\} =$$

$$= \frac{\mathbb{P}\{\text{Data} \mid \text{Tree and History}\} \times \mathbb{P}\{\text{Tree and History}\}}{\mathbb{P}\{\text{Data}\}}$$

$$= \frac{\mathbb{P}\{\text{Data} \mid \text{Tree and History}\} \times \mathbb{P}\{\text{History} \mid \text{Tree}\} \times \mathbb{P}\{\text{Tree}\}}{\mathbb{P}\{\text{Data}\}}$$

- We use [Metropolis-Hastings MCMC](#) to sample from this distribution.
- To do so, we only need to be able to compute the ratio of the above equation evaluated at two different points — the uncomputable normalizing constant in the denominator cancels.

# Sample-based Inference

---

- By construction, the **stationary distribution** of the Markov chain is the desired posterior distribution.
- To account for the **strong dependence in the sample**, we typically need to take very large samples.
- Sample statistics **converge to the corresponding posterior expectations**.
- For example, the sample proportion of each tree topology and of each ancestral arrangement converge to the corresponding posterior probabilities.

# Specific Update Schemes

---

We use four different types of updates.

1. Replace the entire list of reversals on an edge.
2. Slide an internal node between two of its neighbors and update the list of reversals on the other adjacent edge.
3. Do a nearest neighbor interchange and update histories on two edges.
4. Update part of a history on an edge.

# Example — Campanulaceae Chloroplast Arrangements

---

- Several authors have analyzed a data set of 13 chloroplast genome arrangements including tobacco as the outgroup.
- There are 105 markers (but pairwise distances are much smaller).
- There are 13,749,310,575 possible unrooted tree topologies with 13 taxa.

# The Campanulaceae Data

Genera	Arrangement
Trachelium	(1–15) ( <del>76–56</del> ) ( <del>53–49</del> ) (37–40) ( <del>35–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36</del> ) ( <del>25–16</del> ) ( <del>90–84</del> ) (77–83) (91–96) ( <del>55–54</del> ) ( <del>105–97</del> )
Campanula	(1–15) ( <del>76–56</del> ) ( <del>53–49</del> ) ( <del>39–37</del> ) (40) ( <del>35–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36</del> ) ( <del>25–16</del> ) ( <del>90–84</del> ) (77–83) (91–96) ( <del>55–54</del> ) ( <del>105–97</del> )
Adenophora	(1–15) ( <del>76–56</del> ) ( <del>53–49</del> ) ( <del>39–37</del> ) (28–35) (40) (26–27) ( <del>44–41</del> ) (45–48) ( <del>36</del> ) ( <del>25–16</del> ) ( <del>90–84</del> ) (77–83) (91–96) ( <del>55–54</del> ) ( <del>105–97</del> )
Symphyandra	(1–15) ( <del>76–56</del> ) ( <del>39–37</del> ) (49–53) (40) ( <del>35–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36</del> ) ( <del>25–16</del> ) ( <del>90–84</del> ) (77–83) (91–96) ( <del>55–54</del> ) ( <del>105–97</del> )
Legousia	(1–4) (9–15) ( <del>76–56</del> ) ( <del>27–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36–35</del> ) ( <del>25–16</del> ) ( <del>90–84</del> ) (77–83) (91–96) (5–8) ( <del>55–53</del> ) ( <del>105–98</del> ) (28–34) ( <del>40–37</del> ) (49–52) ( <del>97</del> )
Asyneuma	(1–15) ( <del>76–61</del> ) ( <del>56–53</del> ) ( <del>60–57</del> ) ( <del>27–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36–35</del> ) ( <del>25–16</del> ) ( <del>89–84</del> ) (77–83) (90–96) ( <del>105–98</del> ) (28–34) ( <del>40–37</del> ) (49–52) ( <del>97</del> )
Triodanus	(1–15) ( <del>76–56</del> ) ( <del>27–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36–35</del> ) ( <del>25–16</del> ) ( <del>89–84</del> ) (77–83) (90–96) ( <del>55–53</del> ) ( <del>105–98</del> ) (28–34) ( <del>40–37</del> ) (49–52) ( <del>97</del> )
Wahlenbergia	(1–11) ( <del>60–56</del> ) ( <del>53–49</del> ) (37–40) ( <del>35–28</del> ) (12–15) ( <del>76–61</del> ) ( <del>27–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36</del> ) (54) ( <del>25–16</del> ) ( <del>90–84</del> ) (77–83) (91–96) ( <del>55</del> ) ( <del>105–97</del> )
Merciera	(1–10) (49–53) (28–35) ( <del>40–37</del> ) ( <del>60–56</del> ) (11–15) ( <del>76–61</del> ) ( <del>27–26</del> ) ( <del>44–41</del> ) (45–48) ( <del>36</del> ) (54) ( <del>25–16</del> ) ( <del>90–85</del> ) (77–84) (91–96) ( <del>55</del> ) ( <del>105–97</del> )
Codonopsis	(1–8) ( <del>36–18</del> ) ( <del>15–9</del> ) (40) (56–60) (37–39) ( <del>44–41</del> ) (45–53) (16–17) (54–55) (61–76) ( <del>96–77</del> ) ( <del>105–97</del> )
Cyananthus	(1–8) (28) ( <del>36–29</del> ) ( <del>27–26</del> ) (40) (56–60) (37–39) ( <del>25–9</del> ) ( <del>44–41</del> ) (45–48) ( <del>55–49</del> ) (61–96) ( <del>105–97</del> )
Platycodon	(1) (8) (2–5) (29–36) ( <del>56–50</del> ) ( <del>28–26</del> ) (9) ( <del>49–45</del> ) (41–44) (37–40) (16–25) (10–15) (57–59) (6–7) (60–96) ( <del>105–97</del> )
Tobacco	(1–105)

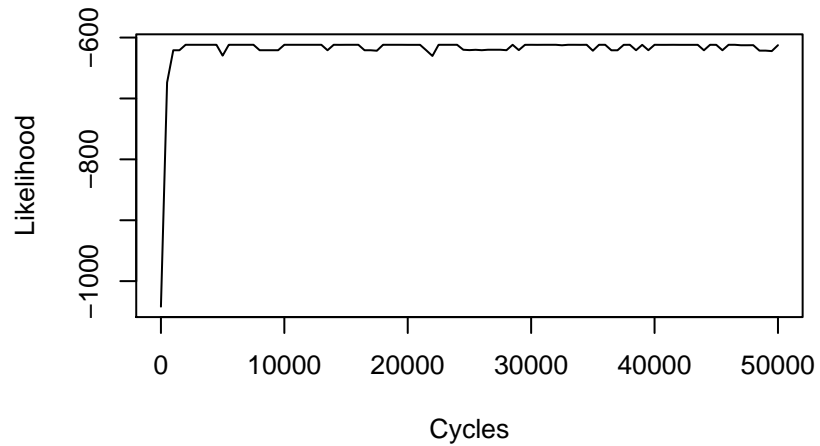
# Campanulaceae Results

---

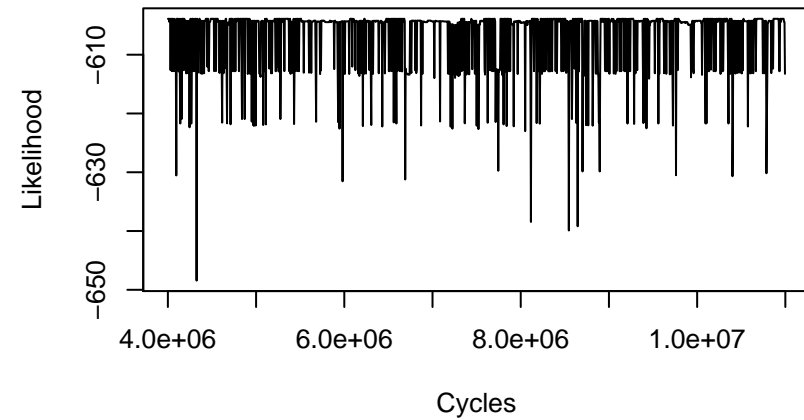
- Papers (by authors in the audience) as recently as 2001 and 2002 reported 67 and then 65 total inversions necessary to explain the data.
- Our MCMC approach finds 180 different trees that require only 64 gene inversions in a matter of minutes on a desktop PC.
- Ran several chains for 11,000,000 cycles (of using each update once).
- Sampled every 500th tree.
- On a P4 desktop computer, each run took about 2hr.

# Simulation Results

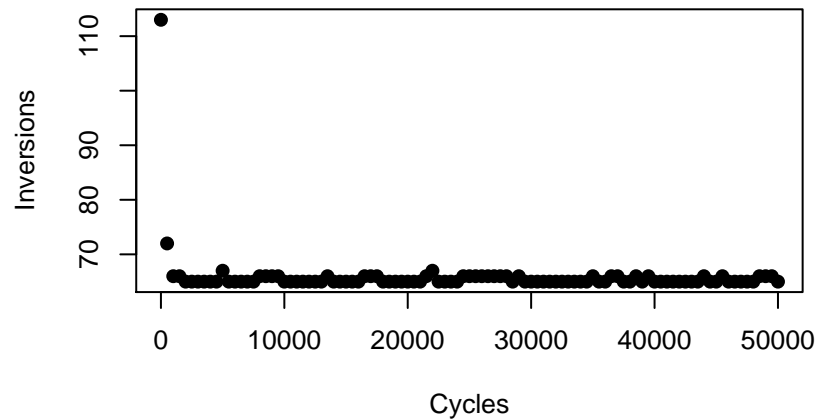
Early Part of Run



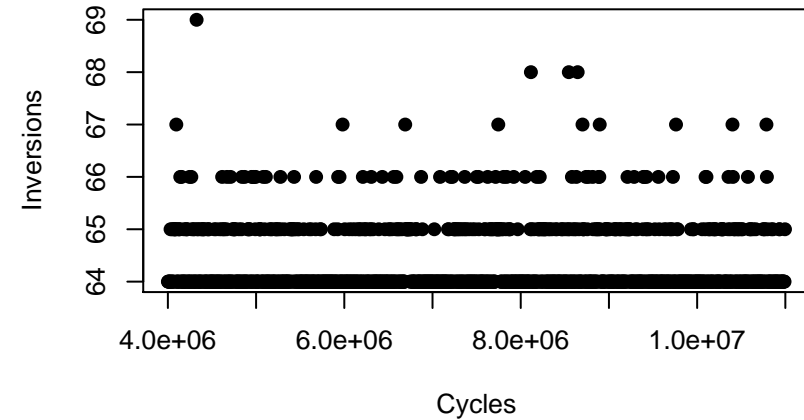
Post Burn-in



Early Part of Run



Post Burn-in



# Posterior Summary

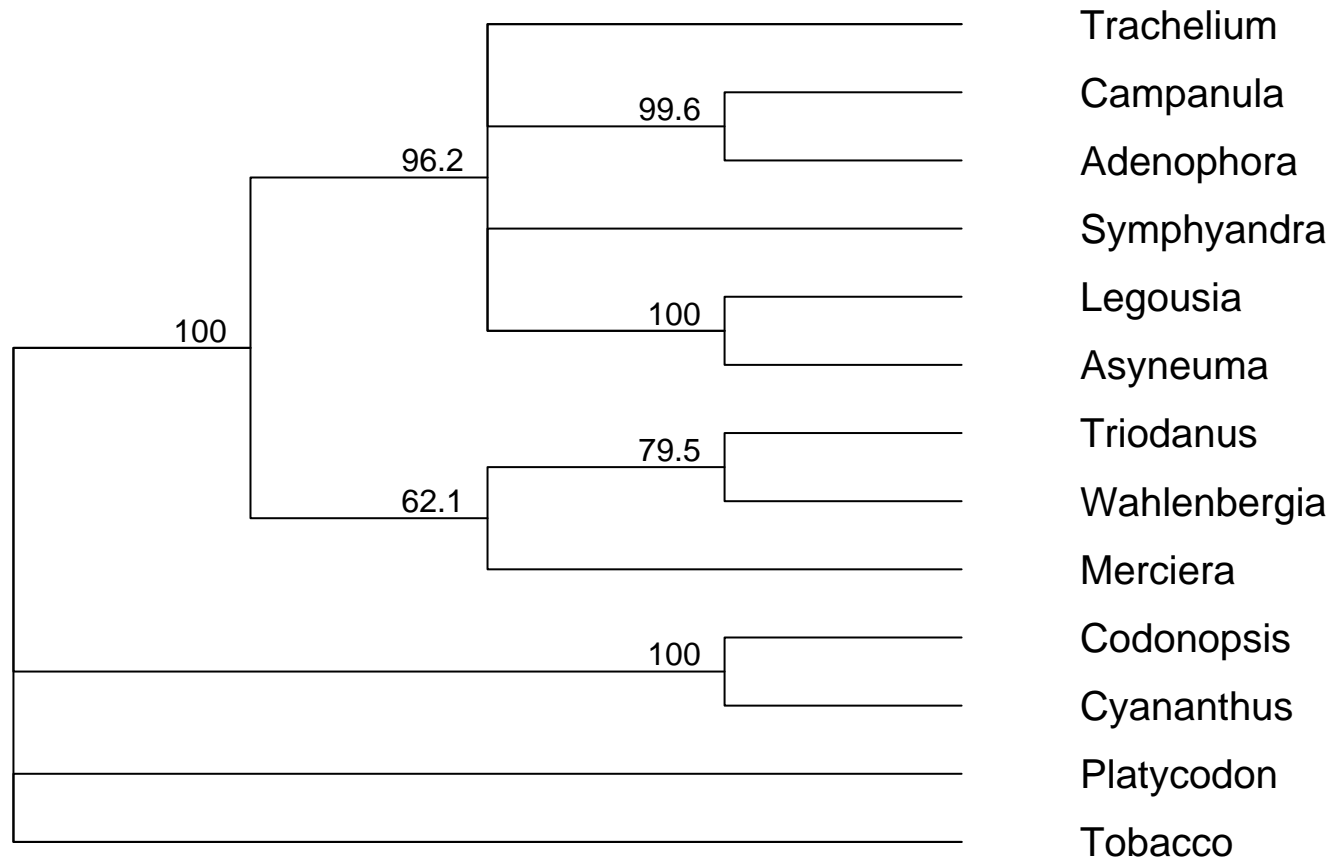
---

Recall that the prior had uniform probabilities on the 13,749,310,575 unrooted tree topologies.

	Credible Region			
	50%	90%	95%	99%
Number of trees	40	135	190	435

# Majority Rule Consensus Tree of Posterior Sample

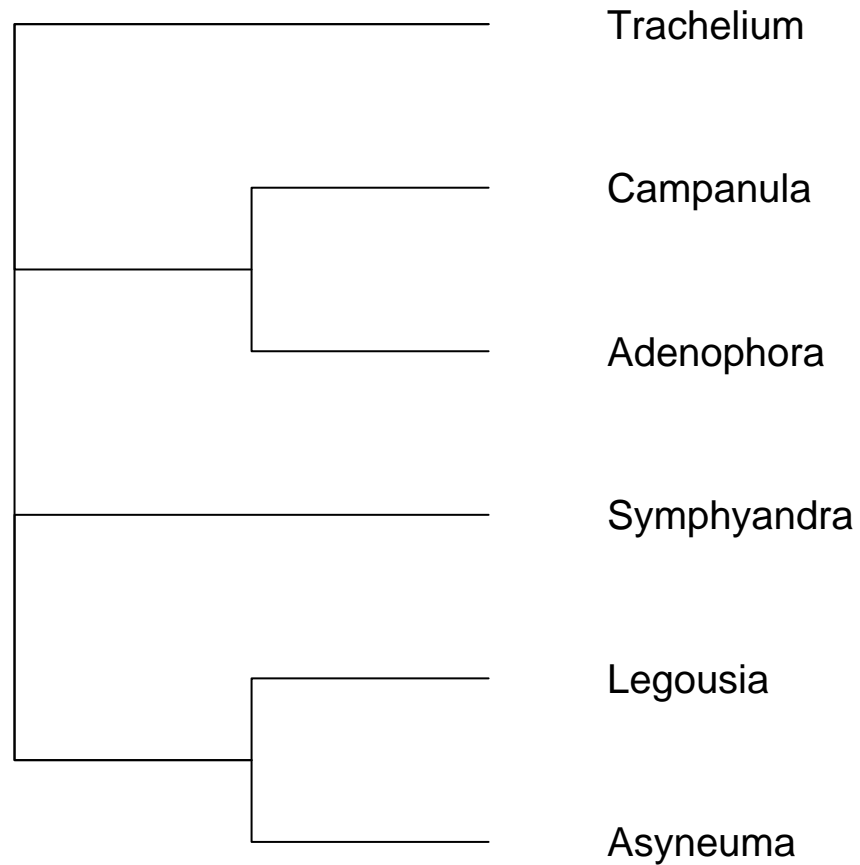
---



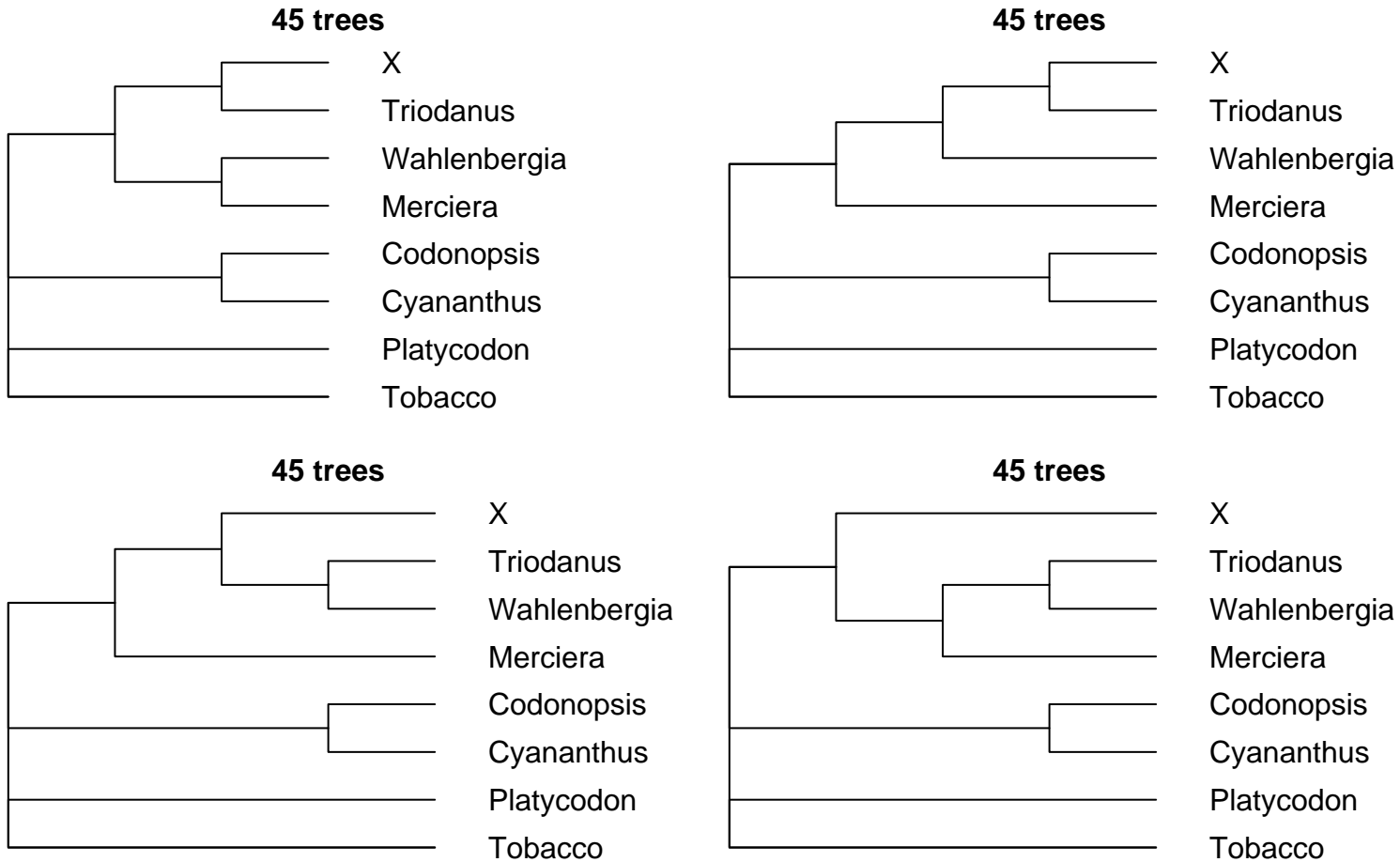
# Maximum Parsimony Trees

---

X (15 subtrees)



# Maximum Parsimony Trees



# Examples — Mitochondrial Genomes

---

- Most animals have the same 37 genes arranged in a ring of DNA in the mitochondria: 22 for tRNAs, 2 for ribosomal RNAs and 13 for proteins.
- Complete mitochondrial genomes are known for over 250 species.
- tRNAs rearrange by multiple mechanisms, so we ignore them for an inversion only analysis.

# The Cast of Characters

---

- There are several dozen metazoan [phyla](#).
- Here is a list of eight for which at least one individual mitochondrial genome is completely sequenced.
  - [Annelida](#) (segmented worms)
  - [Arthropoda](#) (spiders, crustaceans, insects)
  - [Brachiopoda](#) (lamp shells)
  - [Chordata](#) (vertebrates)
  - [Echinodermata](#) (sea stars, brittle stars, sand dollars, etc.)
  - [Hemichordata](#) (acorn worms)
  - [Mollusca](#) (clams, snails, squid, chitons)
  - [Nematoda](#) (round worms)

# The Data

Phylum	Species	Permutation
Chordata	Human	(1-14)
Chordata	Domestic Chicken	(1-8) (10) (9) (11-14)
Hemichordata	Acornworm	(1-8) (10) (9) (11-14)
Echinodermata	Sea star	(6) (1-5) (7-11) ( <u>12</u> ) ( <u>14-13</u> )
Echinodermata	Sea urchin	(6) (1-5) (7-11) ( <u>13-14</u> ) (12)
Echinodermata	Crinoid	(6) (1-5) (7-10) ( <u>11</u> ) ( <u>12</u> ) ( <u>14-13</u> )
Brachiopoda	<i>Laqueus rubellus</i>	(10) (3) (8) ( <u>9</u> ) (5-6) (4) (2) (14) (1) (12) (11) (13) (7)
Brachiopoda	<i>Terebratalia transversa</i>	(10) (2) (4) (3) (8) (11) ( <u>9</u> ) (12-13) (7) (6) (1) (5) (14)
Brachiopoda	<i>Terebratulina retusa</i>	(1-3) (11-13) ( <u>9</u> ) (10) (6-8) (4-5) (14)
Annelida	Common earthworm	(1-2) (4) ( <u>9</u> ) (10) (3) (8) (6-7) (11-13) (5) (14)
Arthropoda	Cattle tick	(1-5) ( <u>13-11</u> ) ( <u>8-6</u> ) ( <u>9</u> ) (10) (14)
Arthropoda	Fruit fly	(1-5) ( <u>8-6</u> ) ( <u>9</u> ) (10) ( <u>13-11</u> ) (14)
Arthropoda	Hermit crab	(1) (5) (14) (2-4) ( <u>8-6</u> ) ( <u>9</u> ) (10) ( <u>13-11</u> )
Arthropoda	Wallaby louse	(4) (13) (10) ( <u>7-6</u> ) (14) ( <u>8</u> ) (1) ( <u>5</u> ) ( <u>12-11</u> ) ( <u>3-2</u> ) ( <u>9</u> )
Mollusca	Squid	(1) ( <u>8-6</u> ) (2-5) ( <u>10</u> ) (9) ( <u>13-11</u> ) (14)
Mollusca	Black chiton	(1-3) ( <u>8-6</u> ) ( <u>10</u> ) (9) ( <u>13-11</u> ) (4-5) (14)
Mollusca	Land snail	(12) ( <u>9</u> ) (8) (13) (6) (10) (1) ( <u>2</u> ) ( <u>3</u> ) ( <u>11</u> ) ( <u>5-4</u> ) (7) (14)
Mollusca	Sea slug	(12) ( <u>9</u> ) (8) (13) (6) (10) (1) ( <u>2</u> ) ( <u>3</u> ) ( <u>11</u> ) ( <u>5</u> ) (7) ( <u>4</u> ) (14)
Nematoda	<i>Trichinella spirallis</i>	(1) (13) ( <u>14</u> ) ( <u>8-6</u> ) ( <u>9</u> ) (10-12) (3-4) (2) (5)

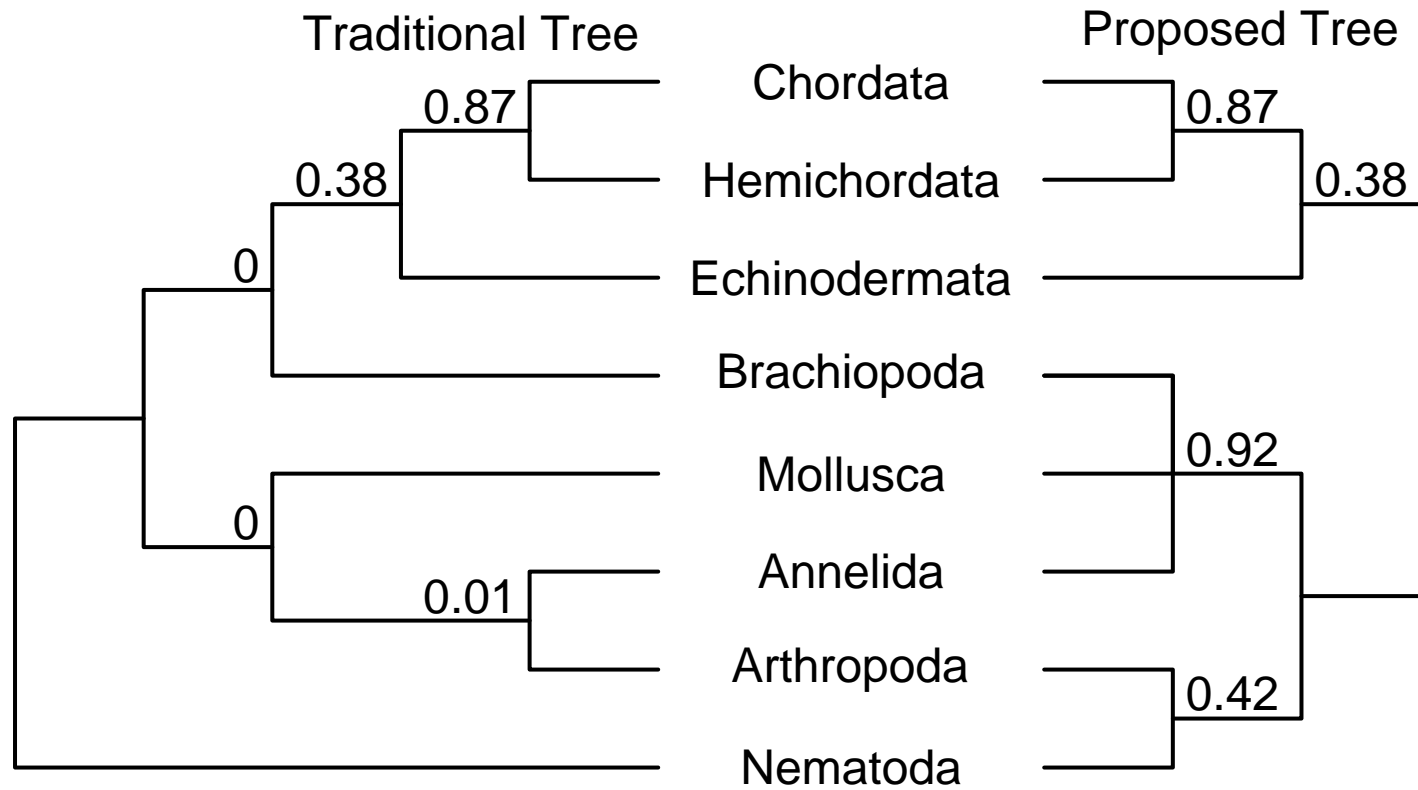
# Specification of a Prior

---

- There are about  $6.3 \times 10^{18}$  different unrooted trees with 19 taxa.
- Only 21,049,875 of these maintain the eight animal phyla.
- A uniform prior on all unrooted tree topologies with 19 taxa corresponds to **prior odds against correct classification of phyla of about 300 billion to one.**
- We believe that the phyla are well supported and put a uniform prior distribution on the smaller set of 21 million tree topologies.
- We select hyperparameters for the distribution of the number of rearrangements per branch so that there is a fairly high probability of no inversions but not too small a probability of ten or more inversions (mean = 2, variance 5 times over-dispersed relative to Poisson).

# Results

---



# Conclusions

---

- It is possible to assess uncertainty in phylogenetic analyses from genome arrangement data alone.
- Algorithmic improvements are possible to make the MCMC more efficient.
- While the statistical approach is not intended to solve the maximum parsimony problem, there are data sets for which it seems to work well for this alternative objective.

# Directions for Further Research

---

- Improve MCMC updates (sticky with long branches).
- Develop models to incorporate multiple genome rearrangement mechanisms.
- Develop methods to combine different data types.
- Incorporate additional genuine prior information.
- Develop better ways to summarize the posterior.
- Scale the approach to work on larger genomes and larger trees.

# Acknowledgments

---

Joseph B. Kadane (a statistician at Carnegie Mellon University) and Donald L. Simon (a computer scientist at Duquesne University) are collaborators on this project.

This work is supported by NIH grant R01 GM68950-01