Review

Cell PRESS

Gene tree discordance, phylogenetic inference and the multispecies coalescent

James H. Degnan[1,2] and Noah A. Rosenberg[1,3,4]

Presented by

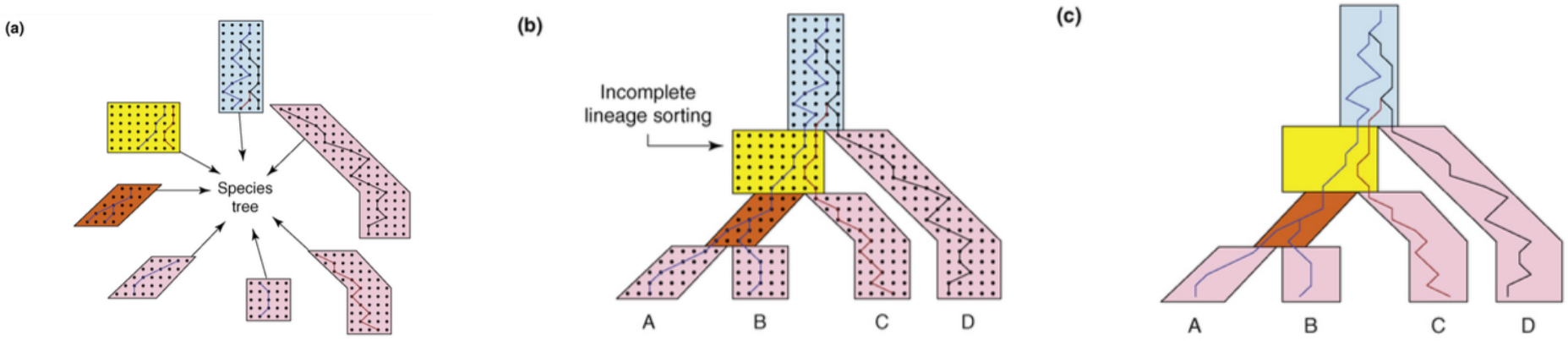
Farzaneh Khajouei

The Multispecies Coalescent

Generalizes the Wright-Fisher model of genetic drift

- Constant effective population size (Ne)
- Non-overlapping generations
- Neutral evolution for the loci modeled
- No structure within populations
- Random joining of lineages backward in time

The coalescent model approximates the process of choosing random parents backward in time when the population size is large relative to the number of sampled lineages.



Incomplete Lineage Sorting (ILS)

Failure of two or more lineages in a population to coalesce, leading to the possibility that at least one of the lineages first coalesces with a lineage from a less closely related population.

- Typical with shallow species trees, where taxa are closely related and the root of the tree is recent
- In deep phylogenies, for some combinations of branching patterns and branch lengths, lineages are likely to sort in a way that violates monophyly of lineages for a species deep in the tree

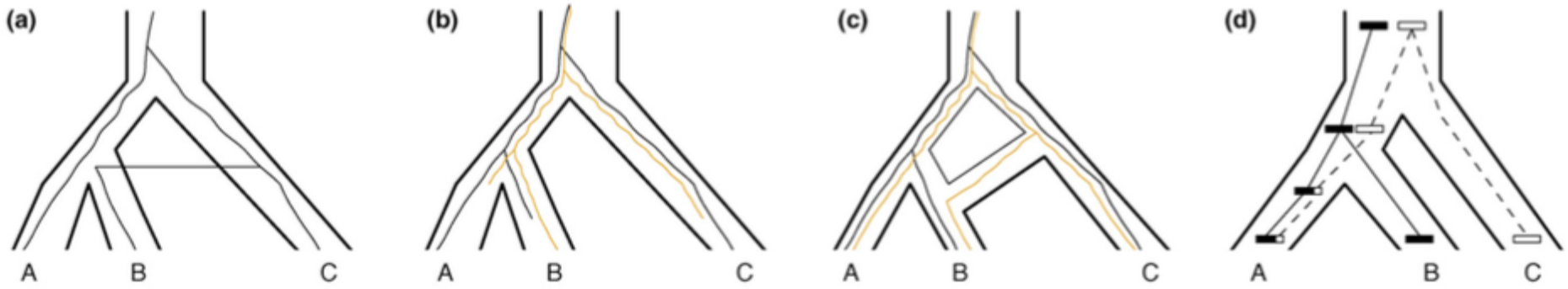
Different usage of the term ILS

- Particular types of genealogical pattern
- A process that explains the gene tree discordance detected in genetic data
- When polymorphisms exist at a a locus in descendant population

Hemiplasy: The gene tree incongruence specifically caused by incomplete lineage sorting when ancestral polymorphism is retained through speciation events.

Gene Tree and Species Tree Discordance

- ✓ Incomplete Lineage Sorting.
- Horizontal Gene Transfer
- Gene Duplication and Loss:
- Hybridization
  - Hybridization affects whole genomes, whereas HGT typically affects only small DNA segments.
- Recombination



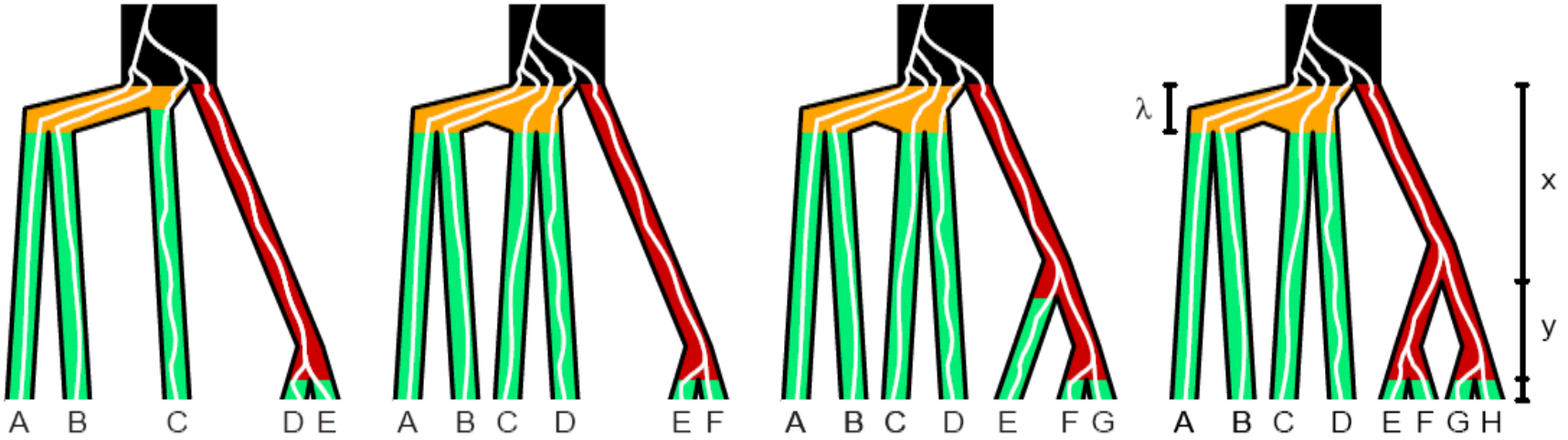
Anomalous Gene Tree (AGT)

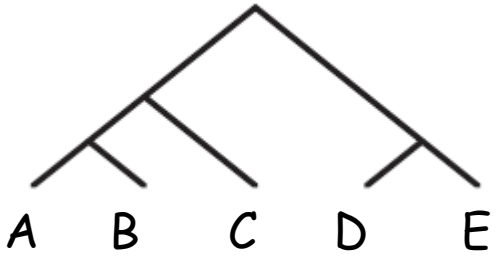
A gene tree topology that is more probable than the gene tree topology that matches the species tree topology

AGT arise with an assignment of species tree branch length for species tree topology with at least five taxa, and also for asymmetric four-taxon tree

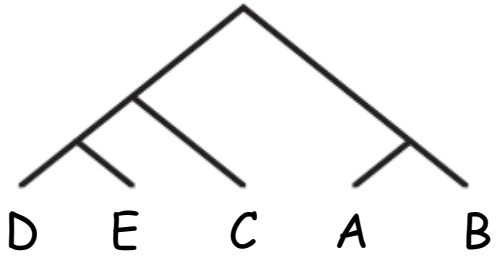
- Long branches, lineages are likely to have coalesced within each population (5Ne)
- Shorter branches, multiple gene lineages tend to persist into deeper portions of the species tree

With 5 or more species, any species tree topology produces at least one anomalous gene tree.





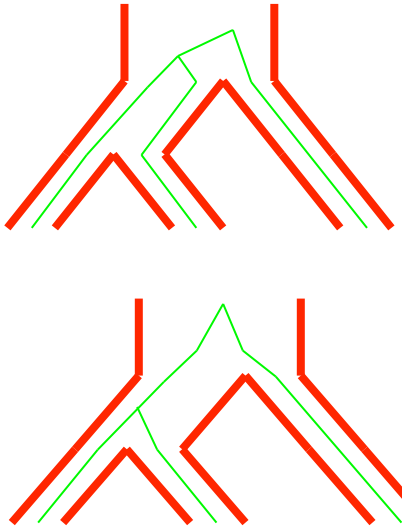
Species Tree



Gene Tree

Gene tree probabilities under the multispecies coalescent model

The probability that *i* lineages have *j* ancestors at *T* coalescent time units ($T = t / N$) in the past is



$$g_{ij}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!\,(k-j)!\,i_{(k)}}$$

$$g_{11}(T) = 1$$
$$g_{21}(T) = 1 - e^{-T}$$
$$g_{22}(T) = e^{-T}$$

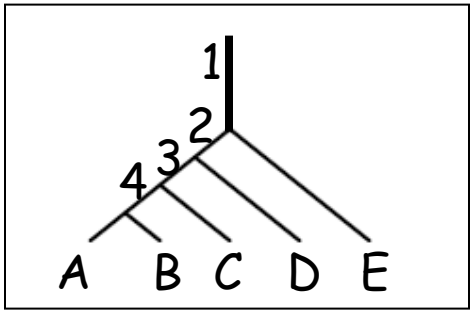
$$g_{31}(T) = 1 - \tfrac{3}{2} e^{-T} + \tfrac{1}{2} e^{-3T}$$
$$g_{32}(T) = \tfrac{3}{2} e^{-T} - \tfrac{3}{2} e^{-3T}$$
$$g_{33}(T) = e^{-3T}$$

$a_{[k]} = a(a-1)\ldots(a-k+1)$

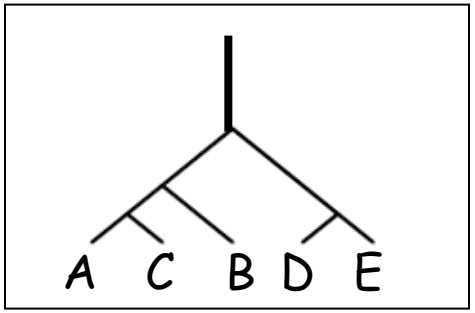
$a_{(k)} = a(a+1)\ldots(a+k-1)$

Coalescent Histories for a five taxa Tree

Species tree

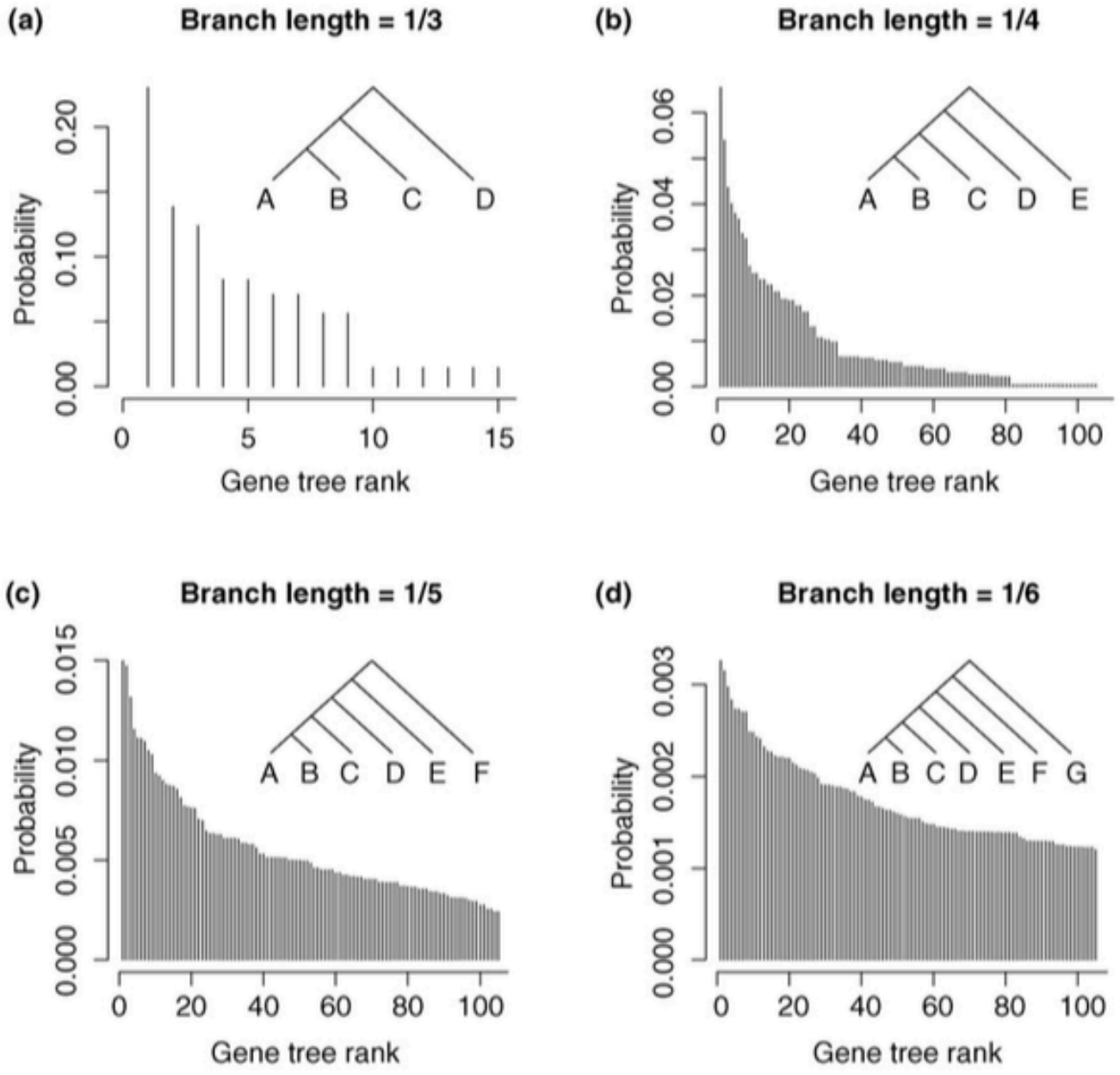


Gene tree



(A,C)	((AC),B)	(D,E)	(((AC)B,(DE))	Probability
1	1	1	1	$\frac{1}{60} g_{22}(T_4) g_{33}(T_3) g_{44}(T_2)$
2	1	1	1	$\frac{1}{54} g_{22}(T_4) g_{33}(T_3) g_{43}(T_2)$
2	2	1	1	$\frac{1}{54} g_{22}(T_4) g_{33}(T_3) g_{42}(T_2)$
3	1	1	1	$\frac{1}{27} g_{22}(T_4) g_{32}(T_3) g_{33}(T_2)$
3	2	1	1	$\frac{1}{27} g_{22}(T_4) g_{32}(T_3) g_{32}(T_2)$
3	3	1	1	$\frac{1}{9} g_{22}(T_4) g_{31}(T_3) g_{22}(T_2)$

$g_{ij}(T)$ is the probability that i lineages coalesce to j lineages during time T


(a)
Branch length = 1/3
Probability
0.00
0.10
0.20
A B C D
0 5 10 15
Gene tree rank
(b)
Branch length = 1/4
Probability
0.00
0.02
0.04
0.06
A B C D E
0 20 40 60 80 100
Gene tree rank
(c)
Branch length = 1/5
Probability
0.000
0.005
0.010
0.015
A B C D E F
0 20 40 60 80 100
Gene tree rank
(d)
Branch length = 1/6
Probability
0.000
0.001
0.002
0.003
A B C D E F G
0 20 40 60 80 100
Gene tree rank

Species Tree Inference

- Democratic vote: the most commonly occurring gene tree topology is used as the estimate of the species tree.
  - Converges on an incorrect estimate when four or more taxa are present and an AGT exists
  - sensitive to sampling variation for small numbers of loci
- Consensus: construct a tree that summarizes input trees defined on the same set of taxa
- Concatenation: all sampled genes are concatenated for each taxon and are then analyzed
- Maximum Likelihood

Species Tree Inference
New Approaches

- Minimizing the number of deep coalescent
- Maximum likelihood (ML): a species tree likelihood is obtained by conditioning on the gene trees at each locus and summing over all possible sets of gene trees
- Bayesian approach

Summary

- A species tree can disagree with the gene tree that it is most likely to produce
- Conflicts in gene tree and species tree can give information about how the species is evolved.
- Conflicting gene genealogies can be used to infer ancestral population parameters
    - population size
    - divergence times

- The number of coalescent histories increases quickly
- This severe discordance only gets worse with more taxa
- Some algorithms can infer the correct species tree even when gene tree discordance is extreme

Outstanding Questions(1)

i. Which species tree estimators from multilocus data are statistically consistent, even when there are AGTs? Among consistent algorithms, which offer the fastest convergence to the species tree?

ii. Do computationally tractable ML algorithms exist that consistently infer the species tree while accounting for variation among gene trees?

iii. What are the effects of taxon sampling for methods of inferring species trees? Do improvements in gene tree estimation owing to increased taxon sampling lead to improvements in species tree estimation?

iv. What is the computational complexity of the evaluation of gene tree probabilities? For a given number of taxa, which gene tree-species tree combination maximizes the number of coalescent histories, and what is this maximum? If the gene tree matches the species tree, which topologies minimize and maximize the number of coalescent histories?

v. Is there a way of computing gene tree probabilities that does not depend linearly on the number of coalescent histories?

Outstanding Questions(2)

vi. For data sets with high levels of gene tree conflict, how can researchers determine whether an AGT is likely? How often do AGTs arise in real data sets?

vii. How sensitive are predictions under the multispecies coalescent to violations of assumptions? What outcomes are expected in cases with ancestral population structure or high levels of intragenic recombination?

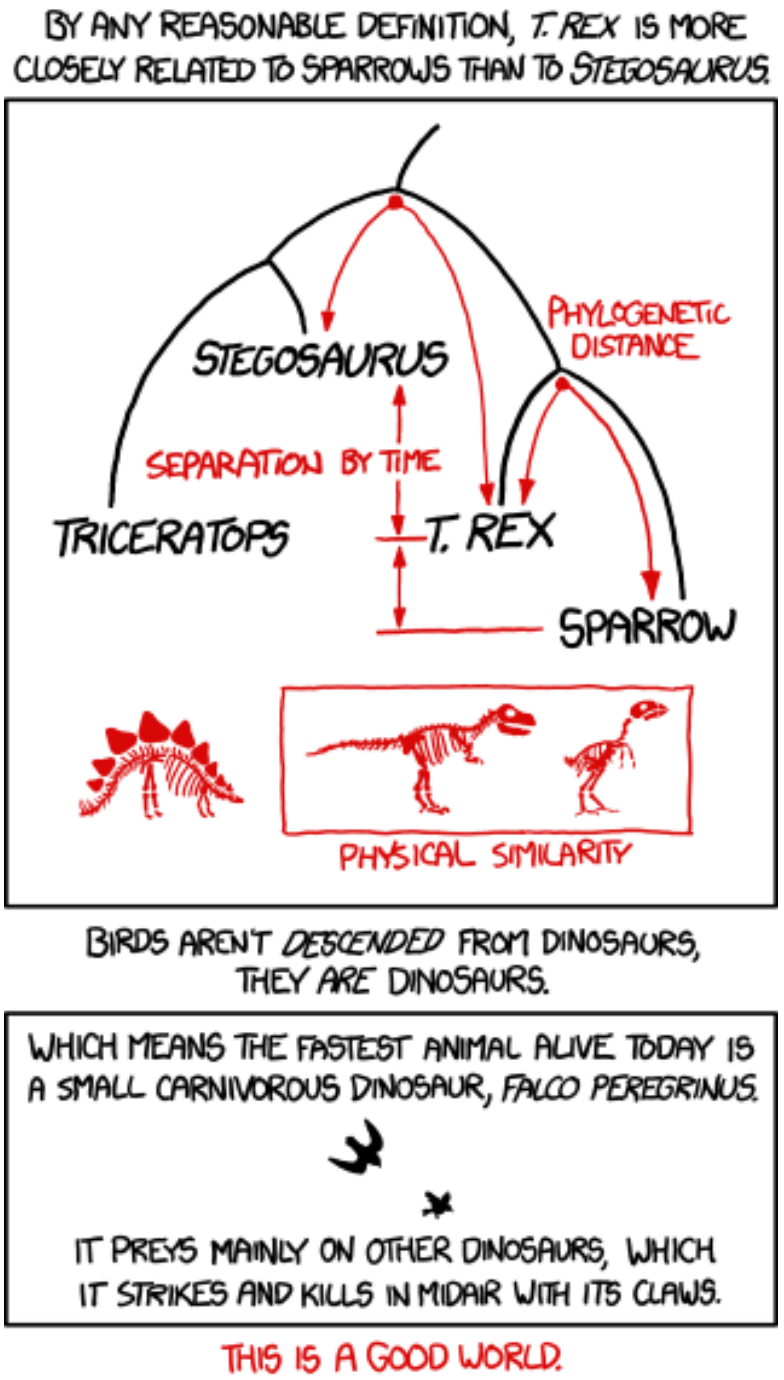
viii. How much discordance in real data sets can be attributed to incomplete lineage sorting, hybridization, gene duplication, HGT, natural selection, recombination and sampling error? What are the best ways of distinguishing sources of discordance?

ix. How does heterogeneity in evolutionary processes interact with gene tree discordance in phylogenetic inference? To what extent do difficulties such as heterogeneity in sequence evolution compound the problems of gene tree discordance?

x. How should tradeoffs among sampling longer sequences, more genes and more individuals per species affect the design of multilocus phylogenetic studies?

Questions?

https://xkcd.com/1211/



Thank You