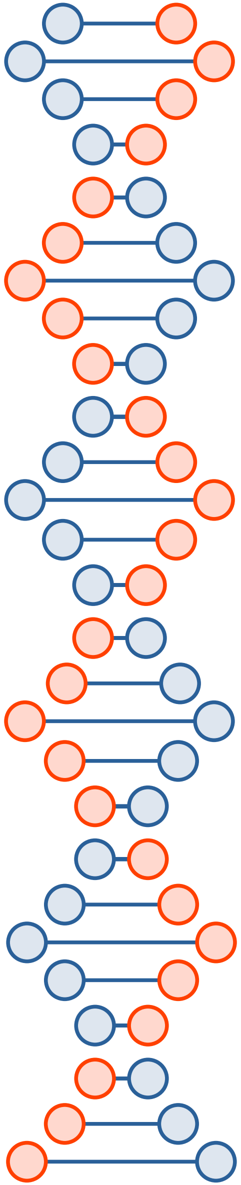


Sample Complexity of Species Tree Estimation when Genes Evolve at Random Rates

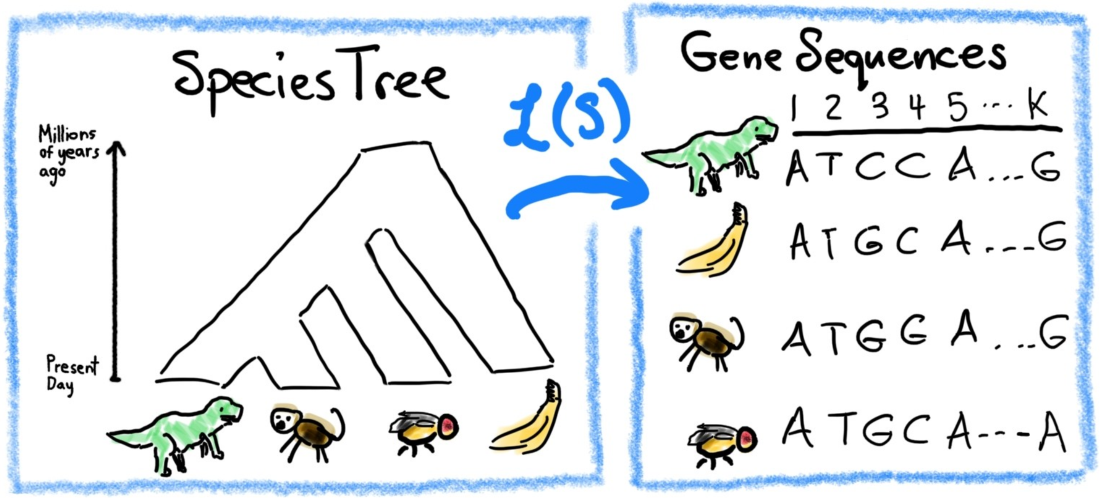
Max Hill (University of Hawaii, Manoa)

Based on joint work with:
Sebastien Roch (University of Wisconsin-Madison)

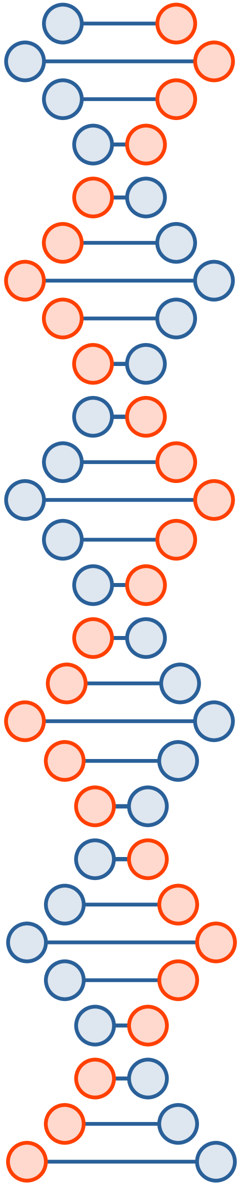
Seminar on Stochastic Processes (March 2025)

Phylogenetics in a Nutshell

- The evolutionary history of a set of taxa is represented by an edge-weighted tree, called a **species tree.**



- We model evolution via a stochastic process parameterized by the species tree.
- **Phylogenetic reconstruction problem:** recover the topology of the species tree from DNA sequence data.

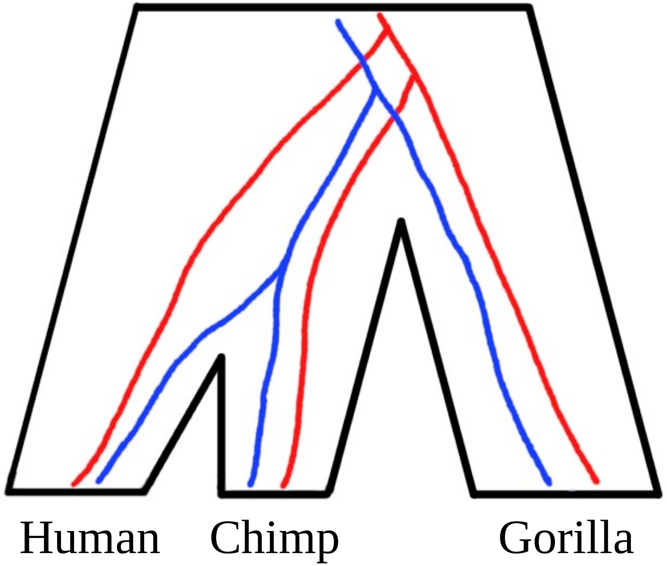


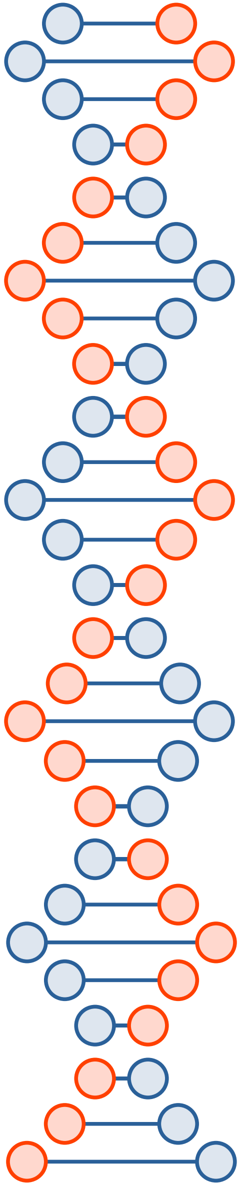
Model of Evolution

- We consider a standard model of gene evolution, the

  Multi-species Coalescent

  which relates the overall evolutionary history of the species with the histories of individual genes, called **gene trees.**
- This is a backwards-in-time stochastic process which gives a distribution of gene trees. This is then combined with a model of how nucleotides (A,T,C,G) mutate.
- In this work, we modify this model to allow for genes to evolve at variable and unknown rates.





Our Results

- **Sample complexity**: how much data is required to achieve high probability of correct inference – in particular when mutation rates aren’t known.

  Our main result: An information-theoretic lower bound which demonstrates that the amount of data required to correctly estimate the evolutionary history is substantially increased when one does not have *a priori* information about the rates of gene evolution.

  - This takes form of a bound on the total variation distance between the sample distributions from two species trees which differ in a minor way: by a short edge of length f>0
  - We show that as f approaches 0, the number of gene samples must grow at least as fast as $(1/f)^2$

- We situate this result in the context of a well-studied information-theoretic tradeoff between different types of data:

  number vs quality of gene tree estimates

  and show that when genes exhibit random variation in mutation rates, this tradeoff collapses.