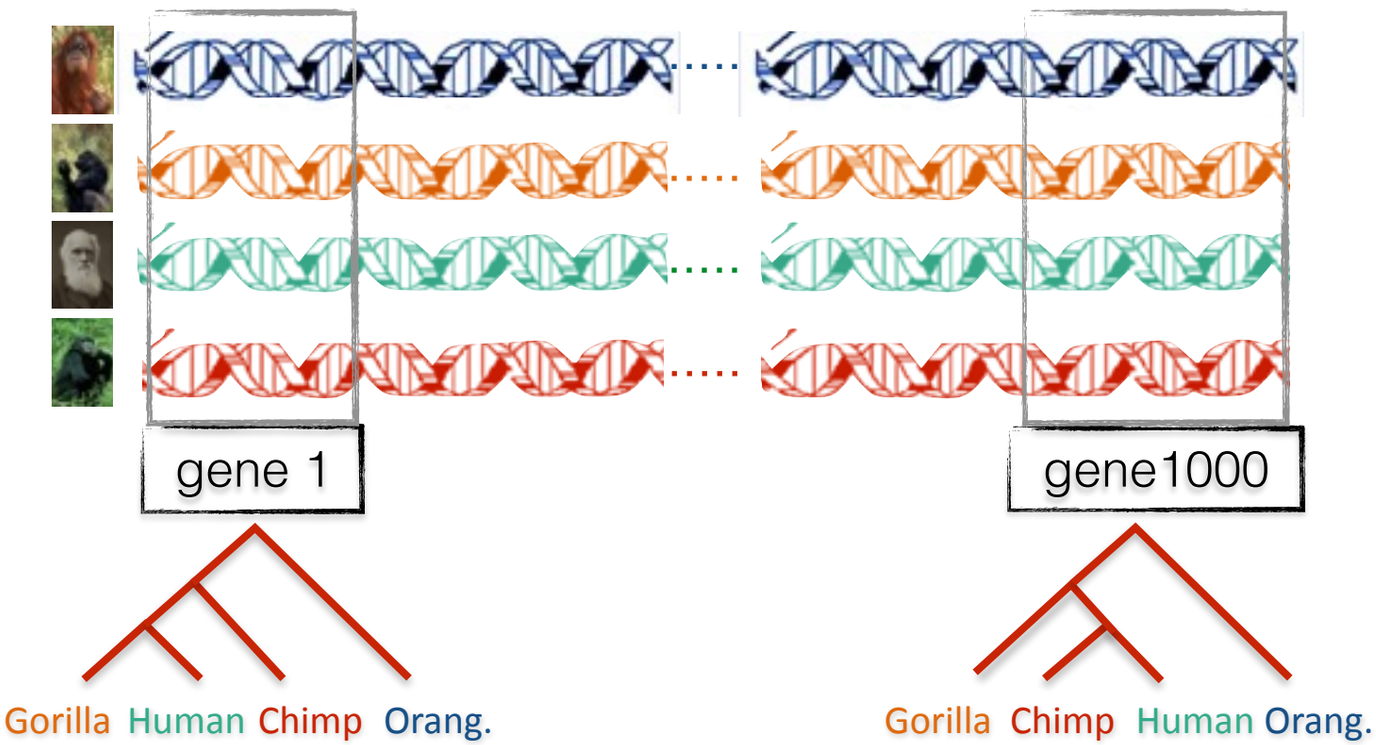
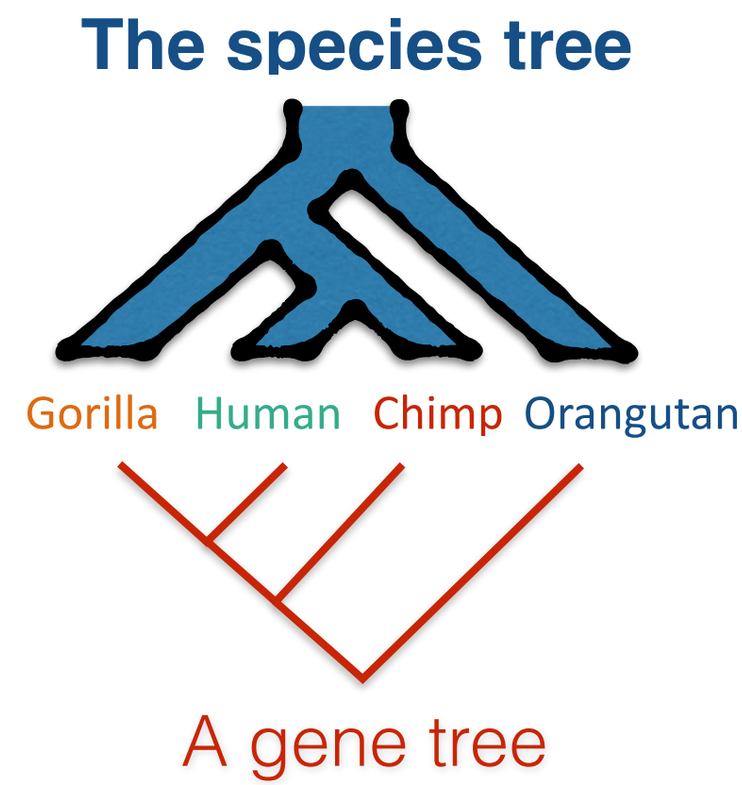
Upcoming challenges in phylogenomics

Siavash Mirarab
University of California, San Diego

Gene tree discordance

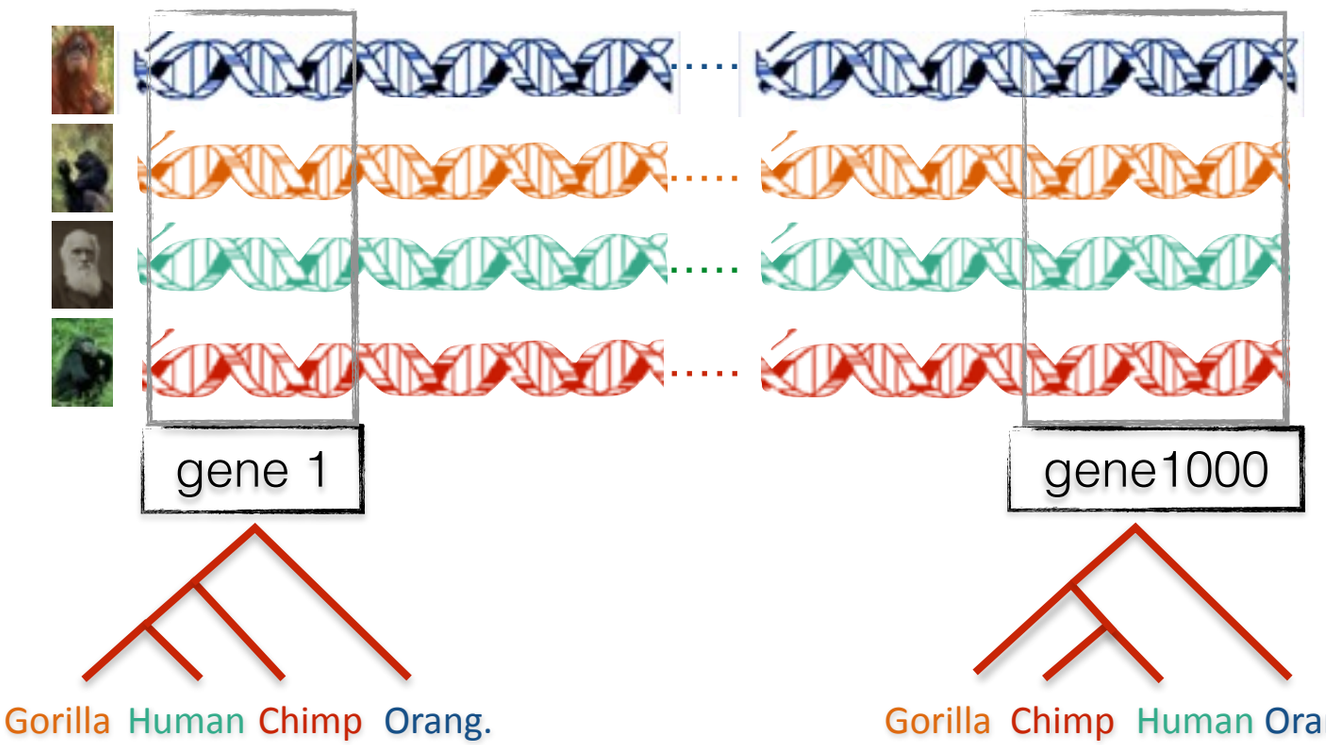


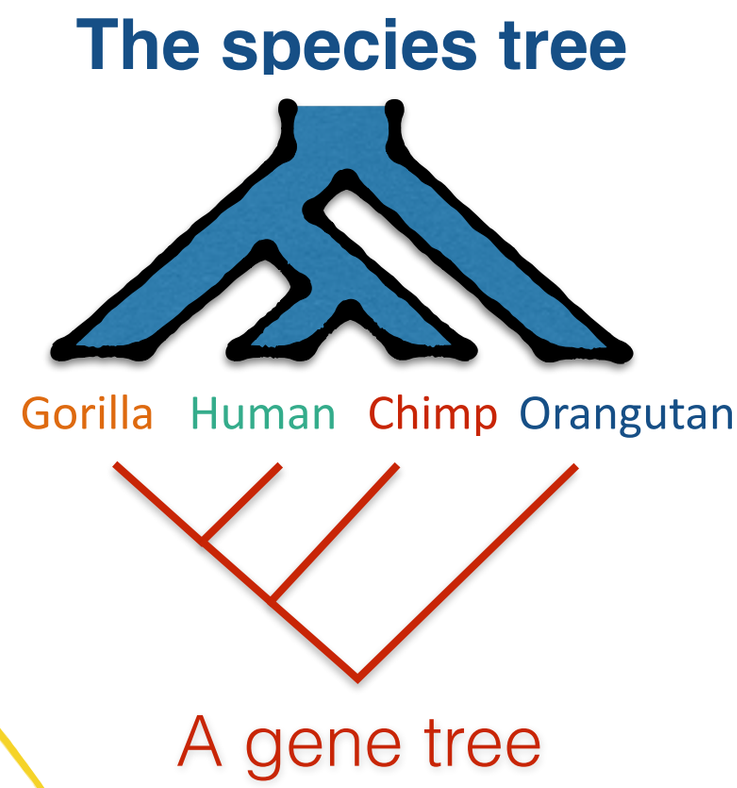


Causes of gene tree discordance include:

- Incomplete Lineage Sorting (ILS)
- Duplication and loss
- Horizontal Gene Transfer (HGT)
- Hybridization

Gene tree discordance

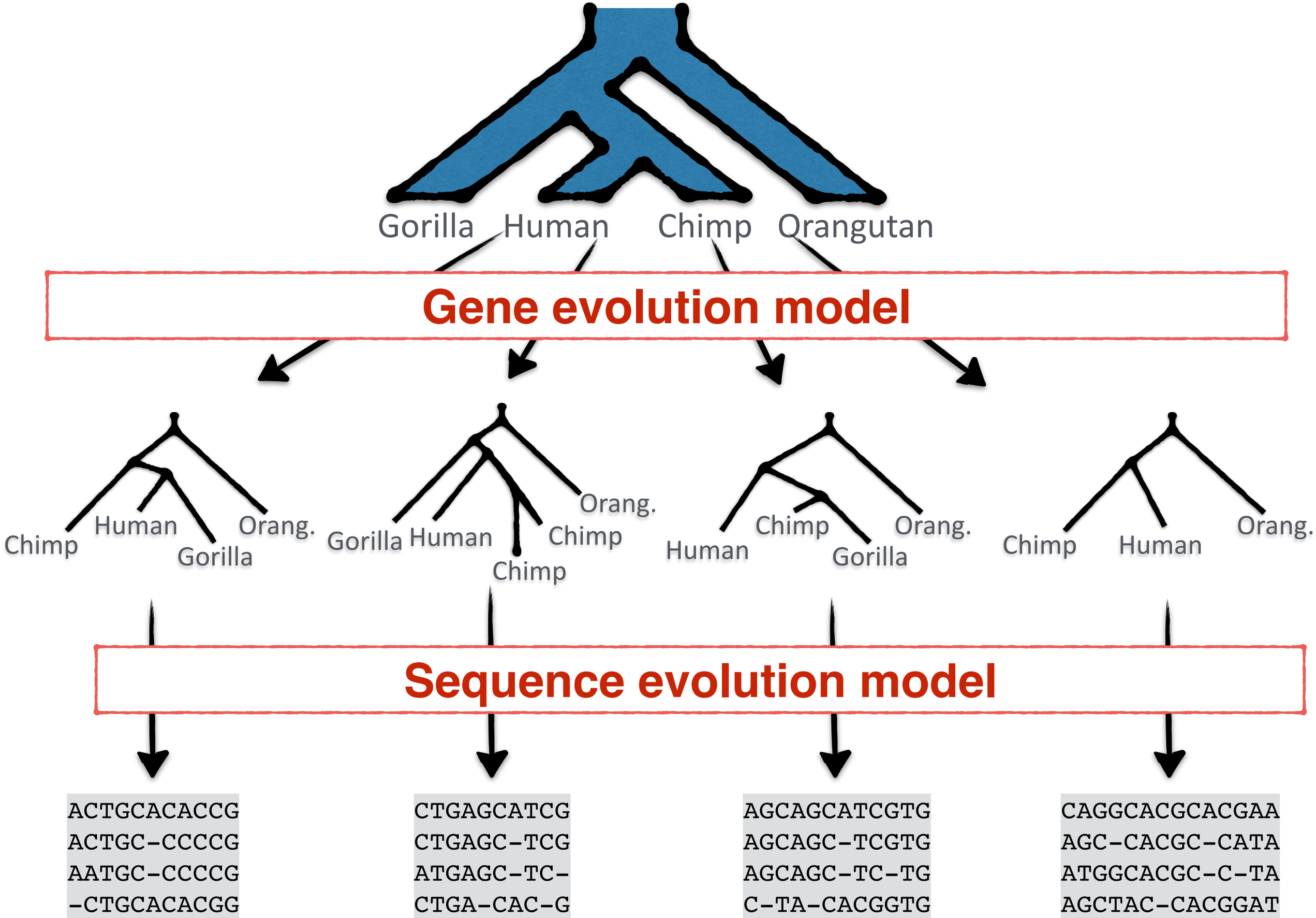




Causes of gene tree discordance include:

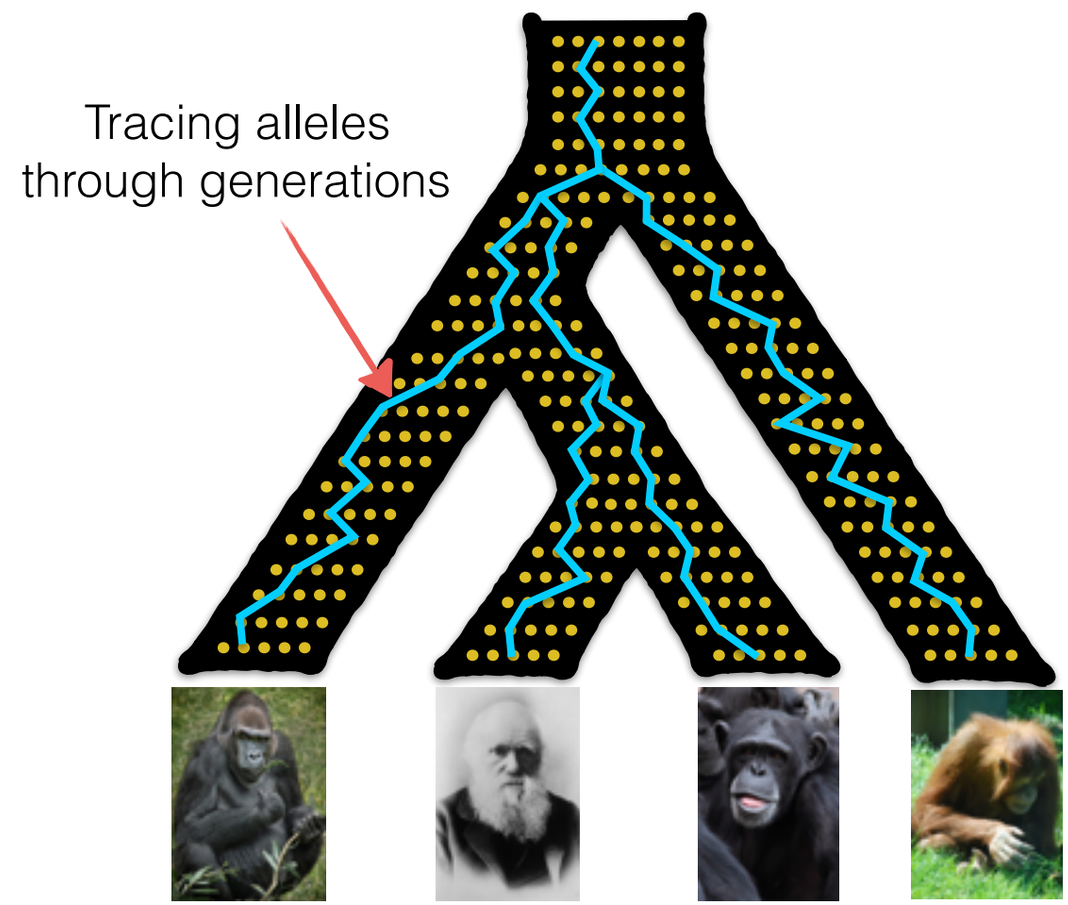
- Incomplete Lineage Sorting (ILS)
- Duplication and loss
- Horizontal Gene Transfer (HGT)
- Hybridization

"c-gene":
<u>recombination-free</u> <u>orthologous</u>
stretches of the genome


Gorilla
Human
Chimp
Orangutan
Gene evolution model
Chimp
Human
Orang.
Gorilla
Gorilla
Human
Orang.
Chimp
Chimp
Human
Chimp
Orang.
Gorilla
Chimp
Human
Orang.
Sequence evolution model
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

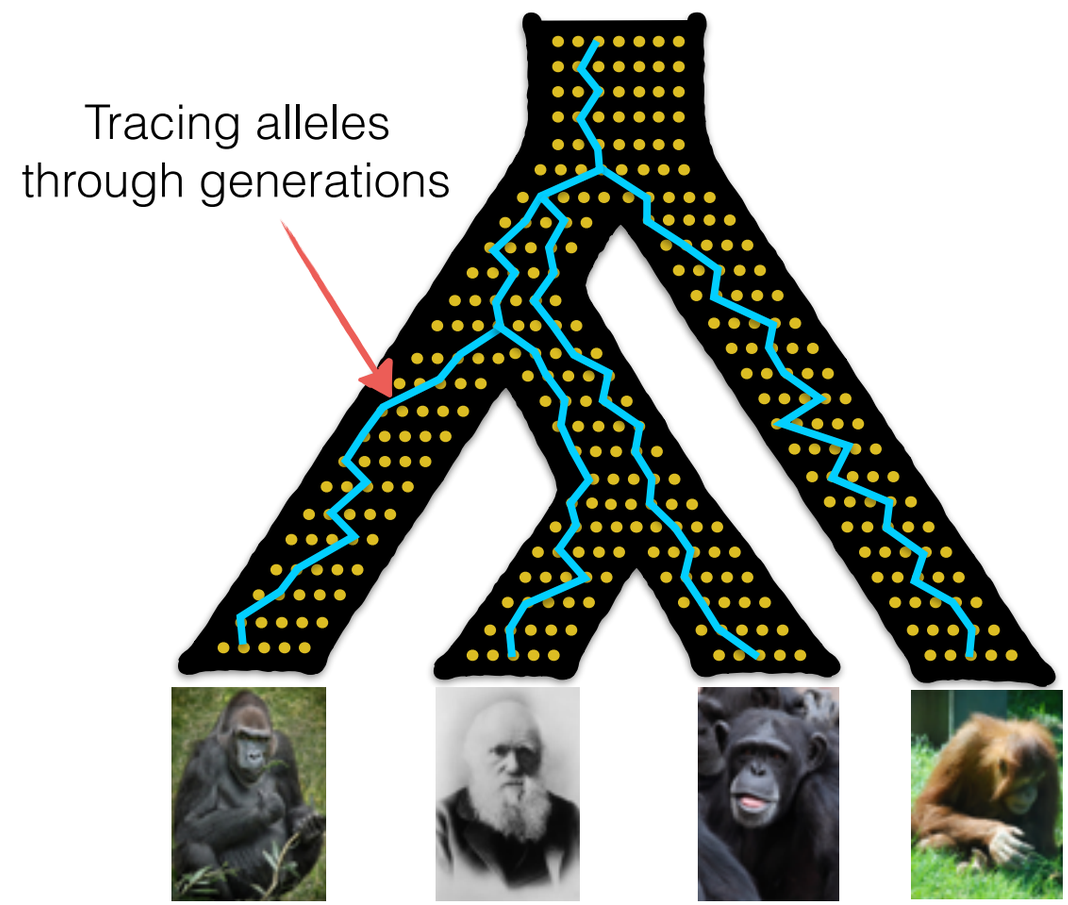
Incomplete Lineage Sorting (ILS)

- The coalescent process extended to multiple species
  - Omnipresent; most likely for short branches or large population sizes



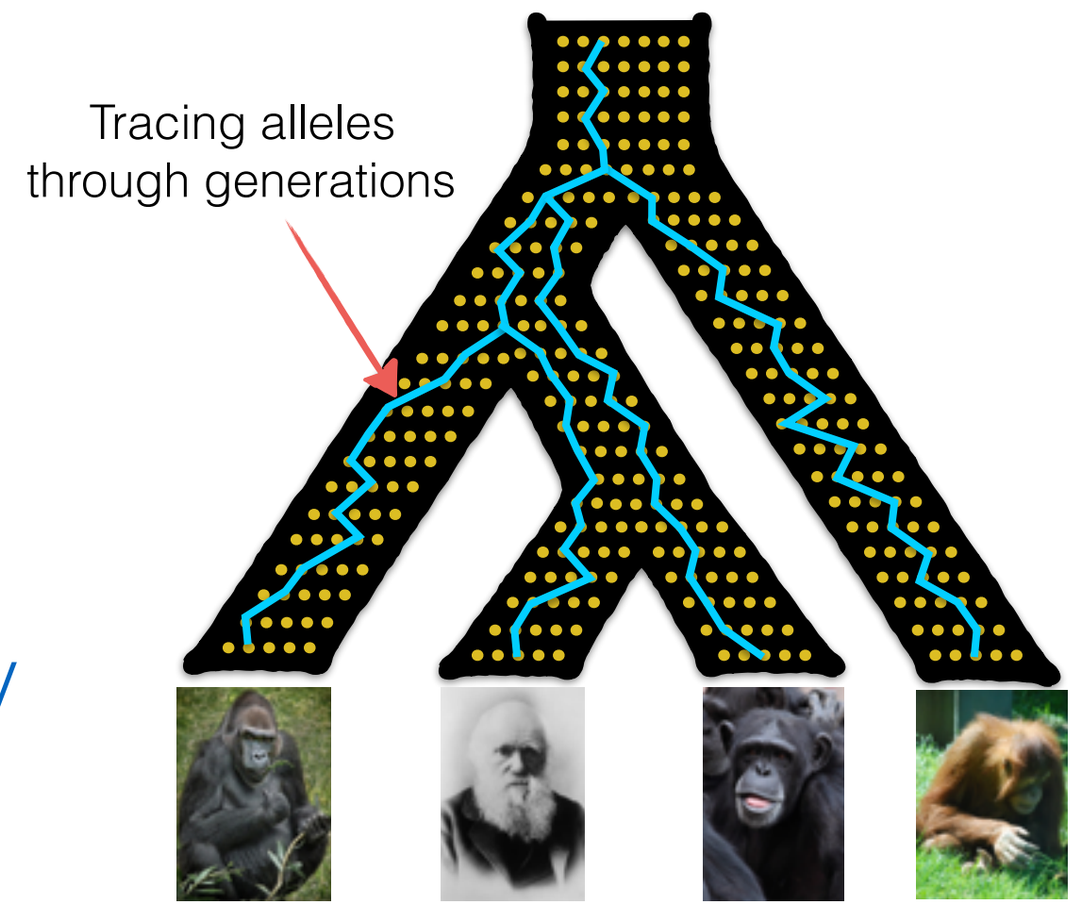
Incomplete Lineage Sorting (ILS)

- The coalescent process extended to multiple species
  - Omnipresent; most likely for short branches or large population sizes

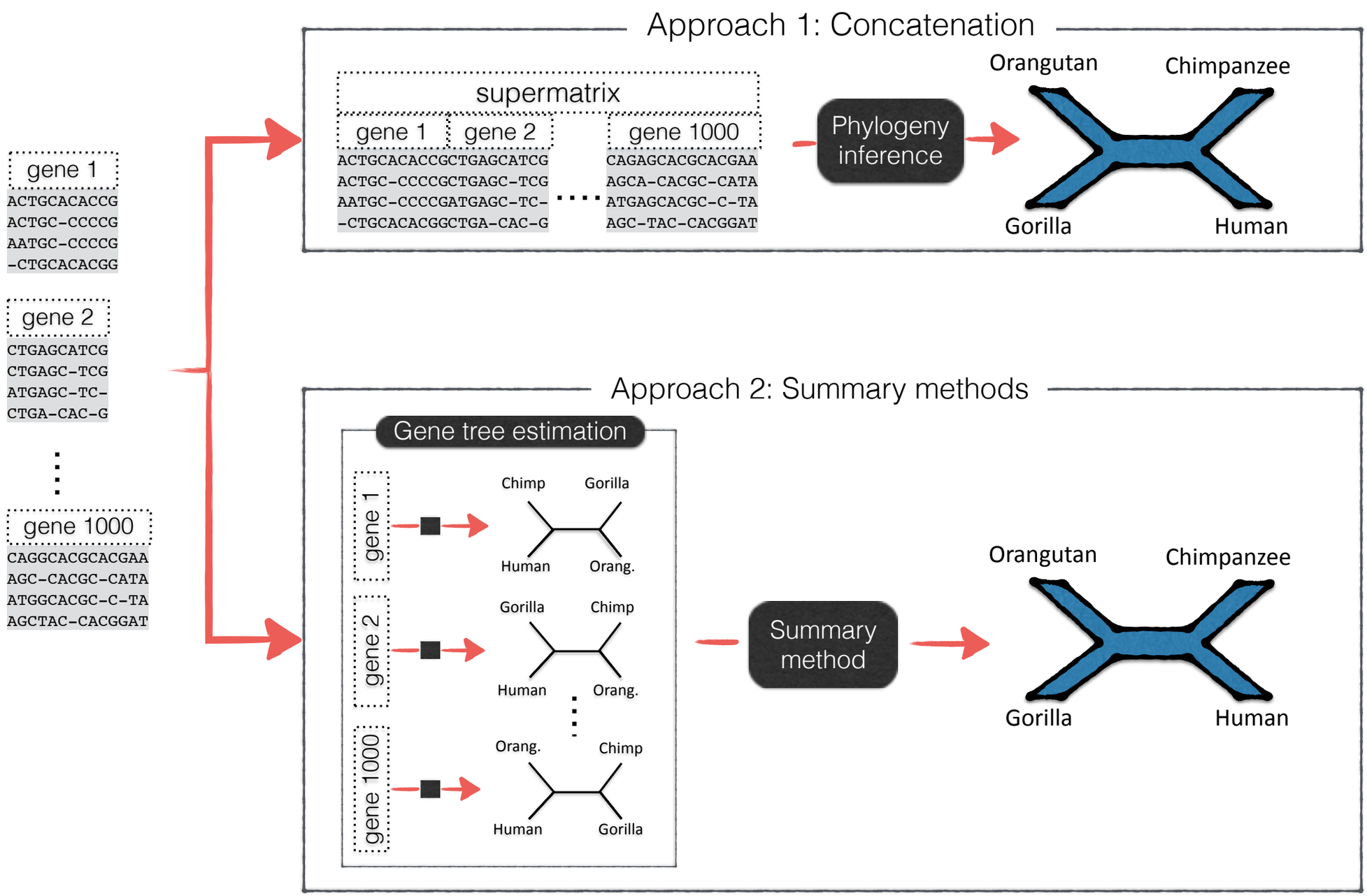


Incomplete Lineage Sorting (ILS)

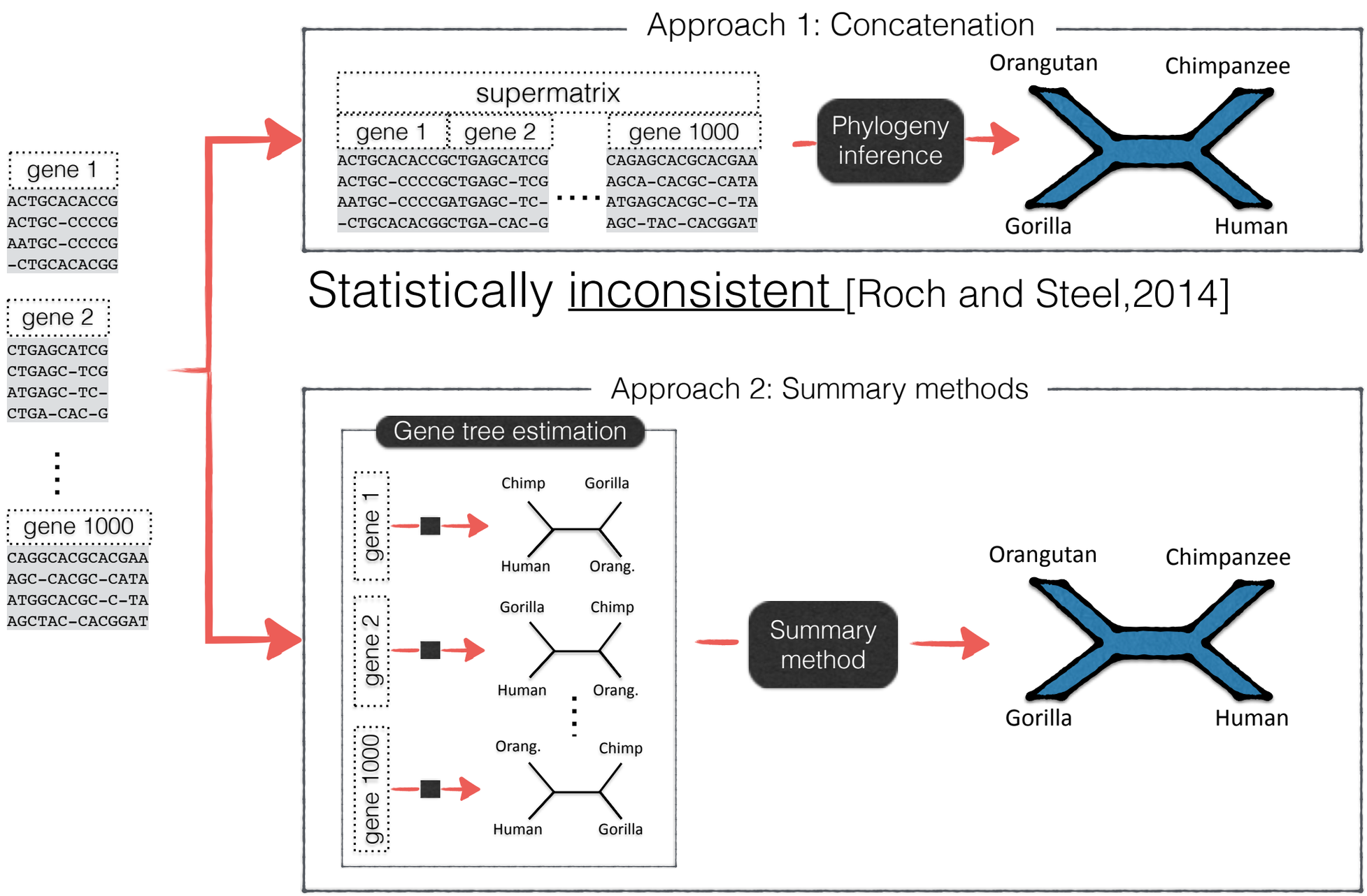
- The coalescent process extended to multiple species
  - Omnipresent; most likely for short branches or large population sizes
- Multi-species coalescent. The species tree defines the probability distribution on gene trees, and is identifiable from the distribution on gene tree topologies
  [Degnan and Salter, Int. J. Org. Evolution, 2005]



Multi-gene species tree estimation



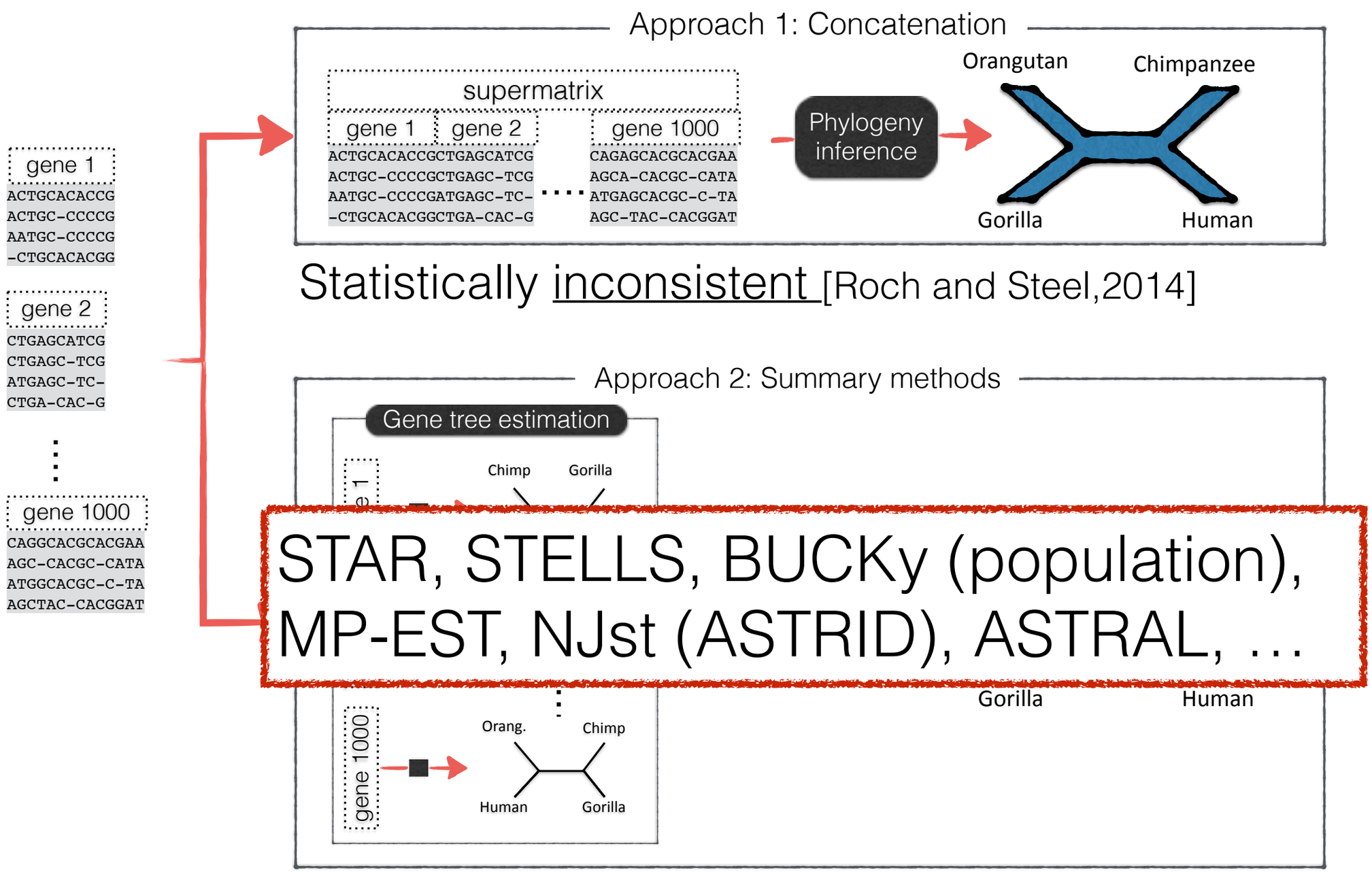
Multi-gene species tree estimation



Statistically <u>inconsistent</u> [Roch and Steel,2014]

Can be statistically consistent given <u>true gene trees</u>

Multi-gene species tree estimation

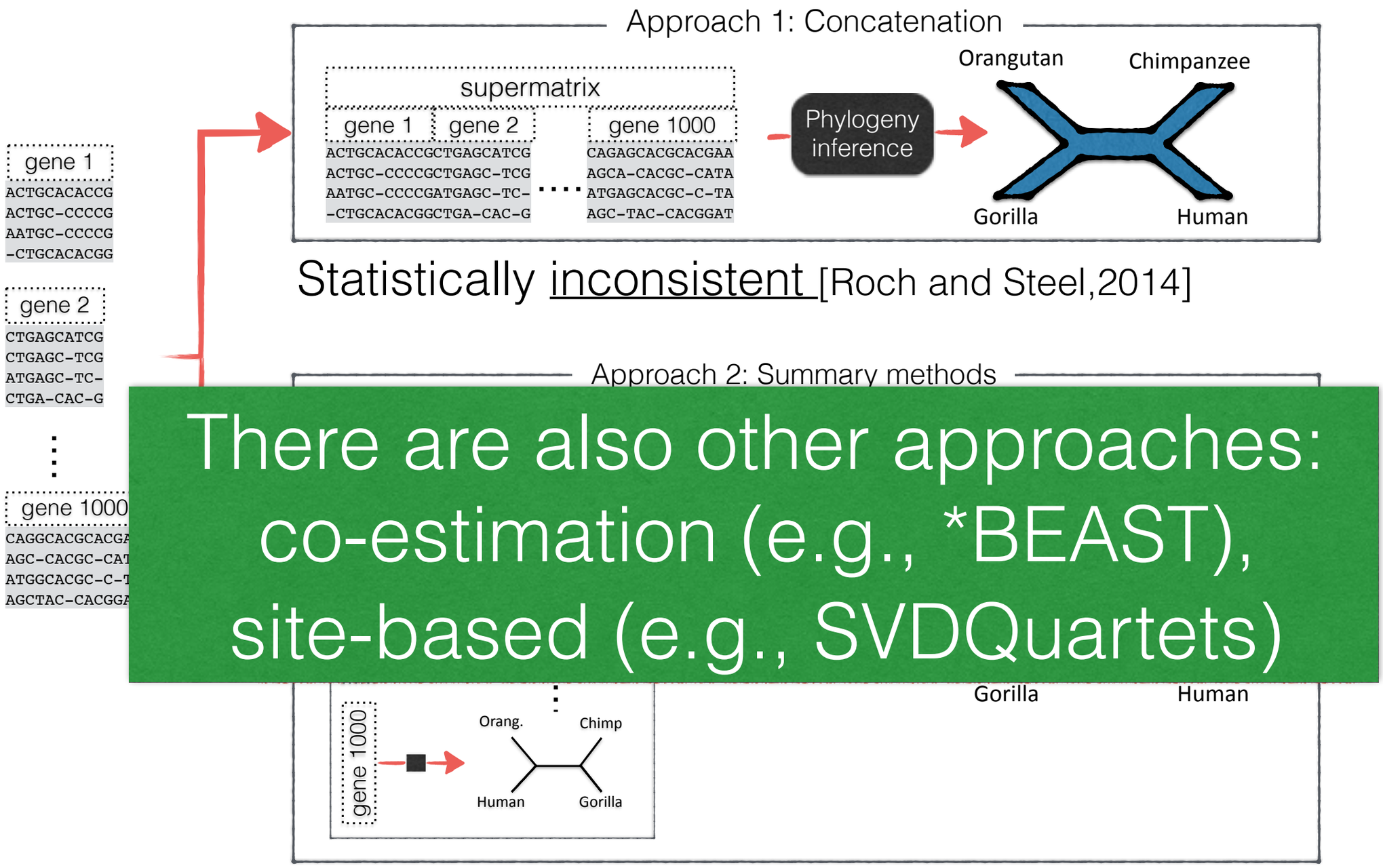


Statistically <u>inconsistent</u> [Roch and Steel,2014]

STAR, STELLS, BUCKy (population), MP-EST, NJst (ASTRID), ASTRAL, …

Can be statistically consistent given <u>true gene trees</u>

Multi-gene species tree estimation



Statistically <u>inconsistent</u> [Roch and Steel,2014]

There are also other approaches: co-estimation (e.g., *BEAST), site-based (e.g., SVDQuartets)

Can be statistically consistent given <u>true gene trees</u>

Challenges

- What is a gene or a species and how do we find them?
- Modeling: multiple evolutionary processes operate together, sometimes creating patterns that are hard to distinguish. How do we untangle them?
- Inference: phylogenetics is hard. Dealing with multi-locus datasets and complex evolutionary processes is often intractable.
- Reliability and interpretation
- Catching up with new data acquisition technologies

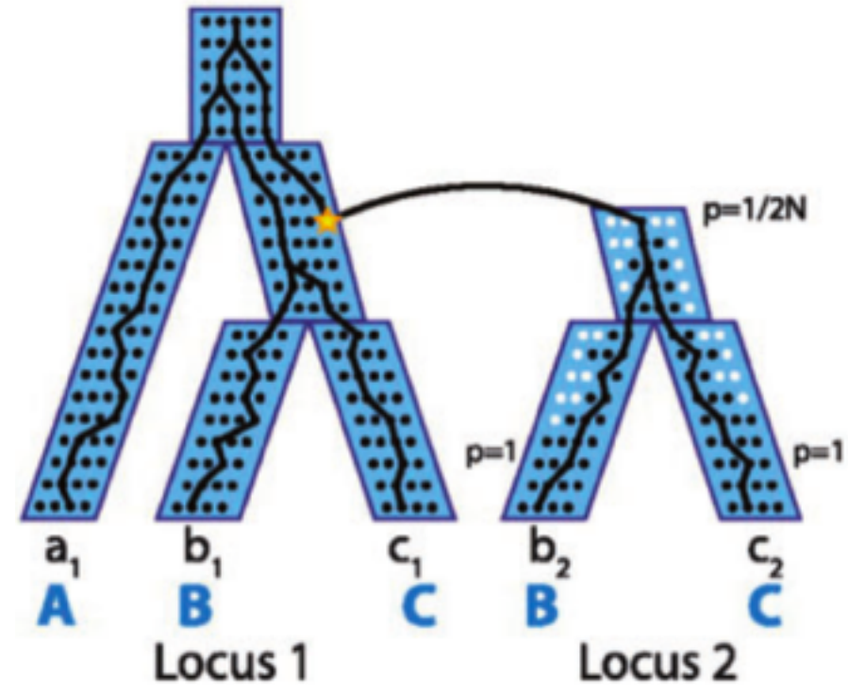
Recombination and gene boundaries

- For the coalescence theory to work, we need (c-)genes to be **recombination-free** regions.
  - Should we try to find recombination free regions? How? Is the signal preserved through millions of years of evolution?
  - Long enough to permit accurately reconstructing gene-specific trees?
- How robust or sensitive are various phylogenetic methods to presence of *some* recombination?

Species tree

- The definition of species and the delineation of boundaries between them is not trivial
- Trees are not always good models. Networks needed in the presence of hybridization, HGT and gene flow (migration)
- Are species trees the most useful “entity” to infer?
    - Maybe gene trees are more useful for “downstream” analyses

Models of discordance



- Single-cause statistical models:
  - ILS: multi-species coalescent
  - Duplication+Loss (duploss): birth+death models
  - Reticulation (HGT): (random models; Roch and Snir)
- ILS+Duploss (Rasmussen & Kellis)
- ILS+Hybridization (Yun et al., Luay's lab)
- Duploss+HGT (Tofigh et al., Szöllósi et al.)

Complex models

- Combining multiple causes of discordance results in complex (parameter-rich) models
  - Inference is hard
  - There are often identifiability issues
- See Szöllõsi et al., 2015 for a recent review

Inference

- Ideally, we combine sequence and gene evolution into a single hierarchical model and **co-estimate** gene and species trees
  - Combining all processes is computationally intractable
- **Pipeline**:
  assemble reads —>
  find orthologous genes (gene families) or genomic regions —>
  multiple sequence alignment per gene —>
  infer gene trees —> infer species trees/networks
  - Error propagates from one step to the next

Progress

- Co-estimation methods exist for
  — substitutions+ILS (e.g., *BEAST)
  — substitutions+Duploss (e.g., PHYLDOG)
  — ILS+Hybridization (e.g., PhyloNet).

- Scalability limited to
  — small numbers of genes (scalability is gradually improving)
  — small numbers of species (e.g., tens)

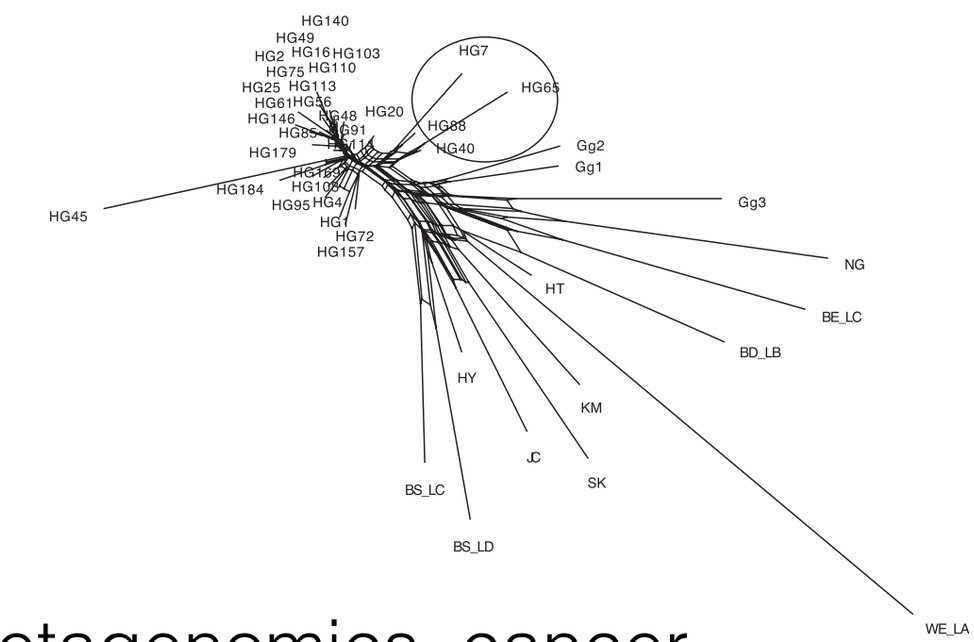
- Sequence-based (gene tree free) methods of species tree estimation (e.g., SNAPP, SVDQuartets, etc.)

- Heuristic methods of improving gene trees (e.g., gene binning)

- HMM-based methods of scanning genomes (e.g., CoalHMM)

Interpretability, Data, …

- Interpretation: truth is knowable in phylogenetics.
  - How do we evaluate models, methods, results?
    — Need good generative models (ideally more complex than inference models)
  - Best ways to estimate support?
  - How to interpret networks?
    — Data visualization
    — Biological events



- Evolution in new types of data (e.g., metagenomics, cancer, immunogenetics, HIV, etc.); data generation models.

Where to go?

- Can inference under existing models become scalable to hundreds of species and thousands of genes?
- Can we combine even more processes into a single model? For example, a model of ILS+Duploss+Transfer+Substitutions+Indel?
- Can smart scalable heuristic approaches be designed to side-step some of the scalability challenges?
- What are the fundamental limits of a full inference of past evolutionary processes?
- Are there “magic markers” out there? Are large-scale processes such as rearrangements full of signal, waiting to be discovered?

Challenges

- What is a gene or a species and how do we find them?
- Modeling: multiple evolutionary processes operate together, sometimes creating patterns that are hard to distinguish. How do we untangle them?
- Inference: phylogenetics is hard. Dealing with multi-locus datasets and complex evolutionary processes is often intractable.
- Reliability and interpretation
- Catching up with new data acquisition technologies