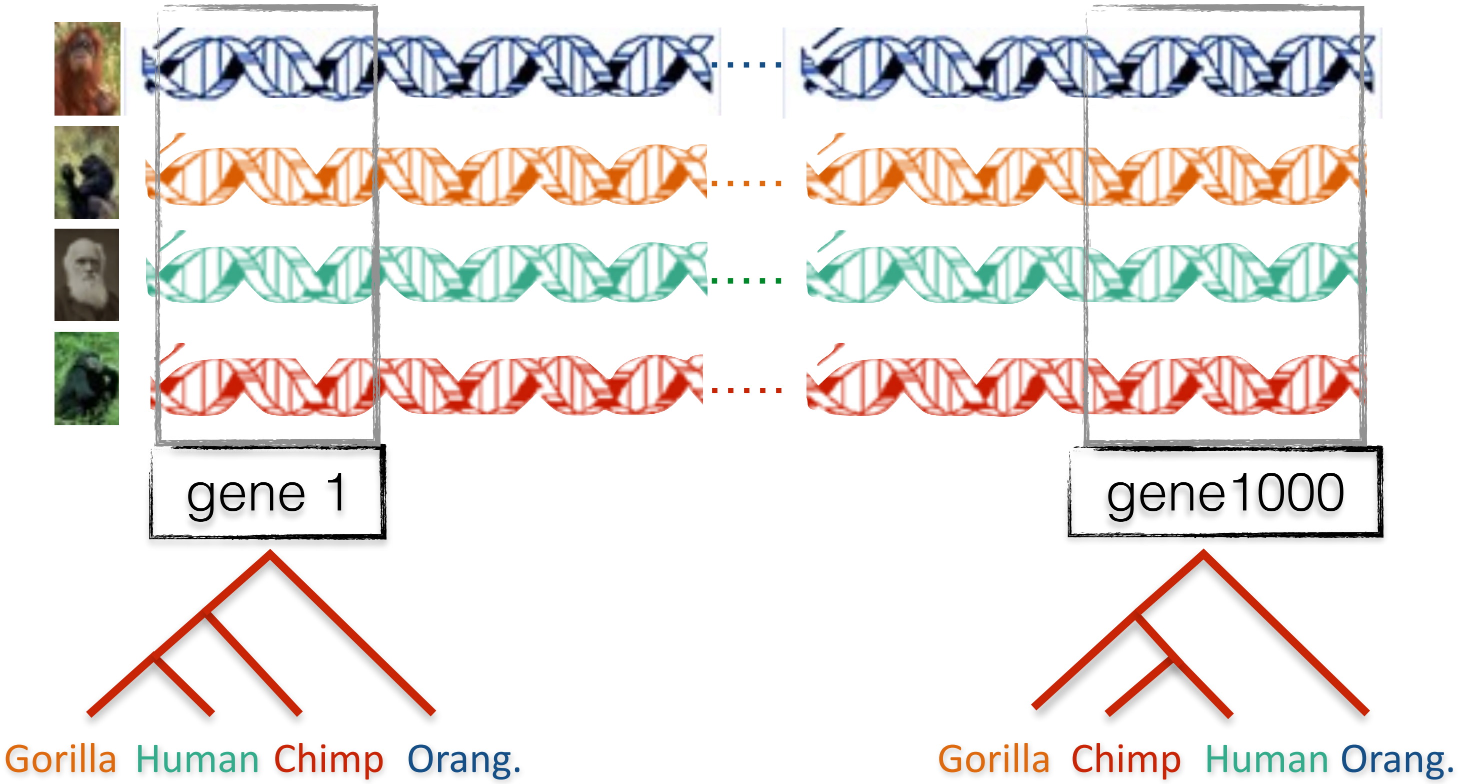
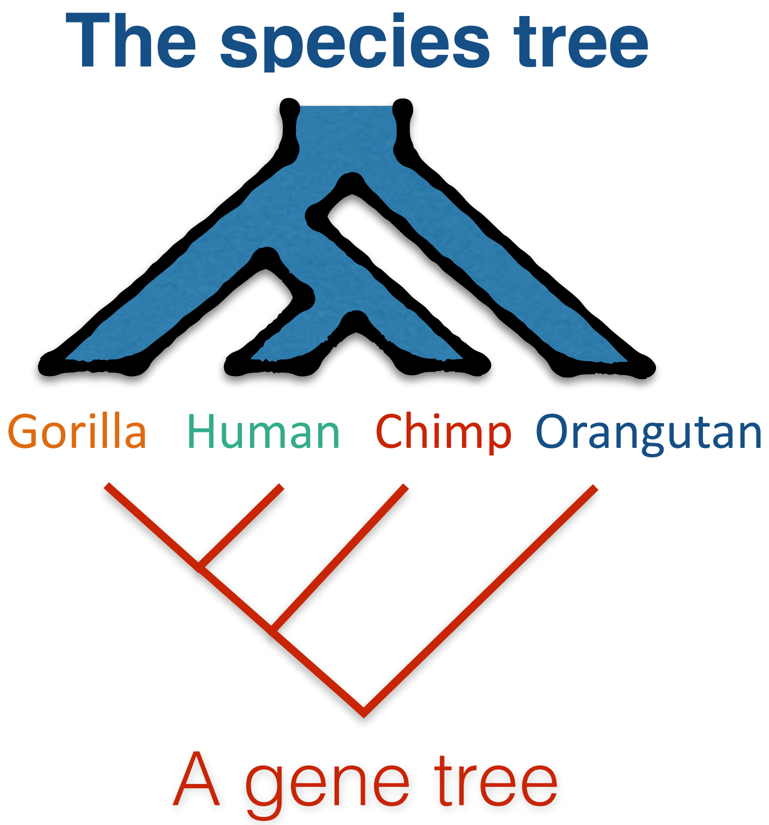
Challenges and advances in genome-wide species tree reconstruction

Siavash Mirarab
University of California, San Diego

Gene tree discordance

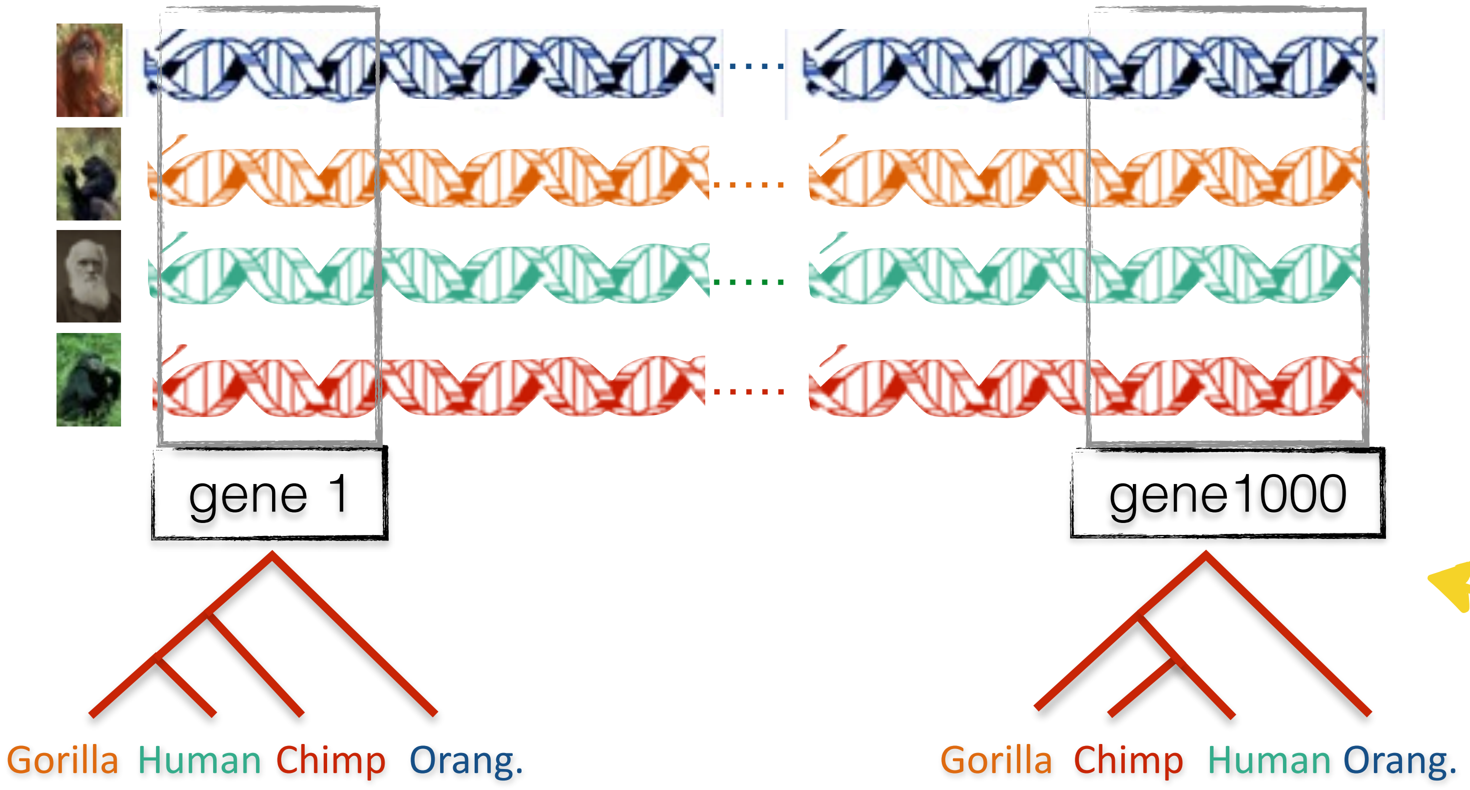




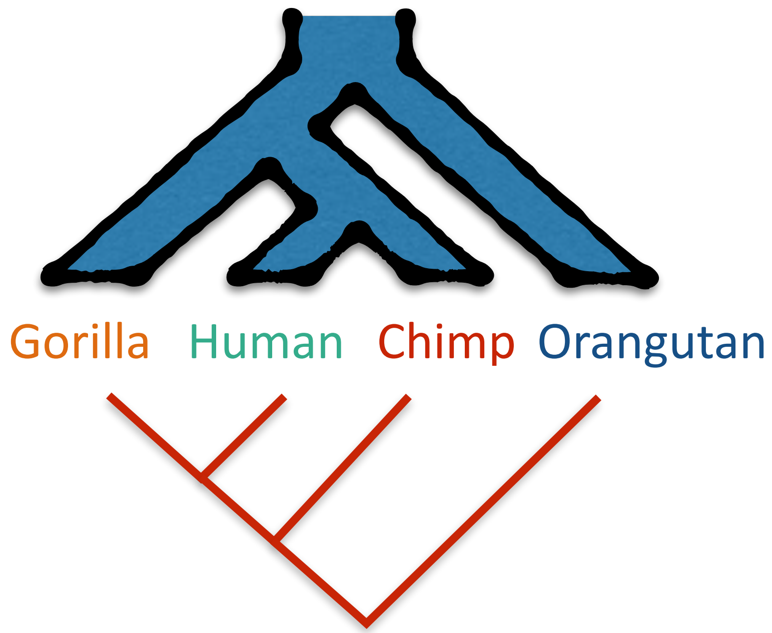
Causes of gene tree discordance include:

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

Gene tree discordance



The species tree

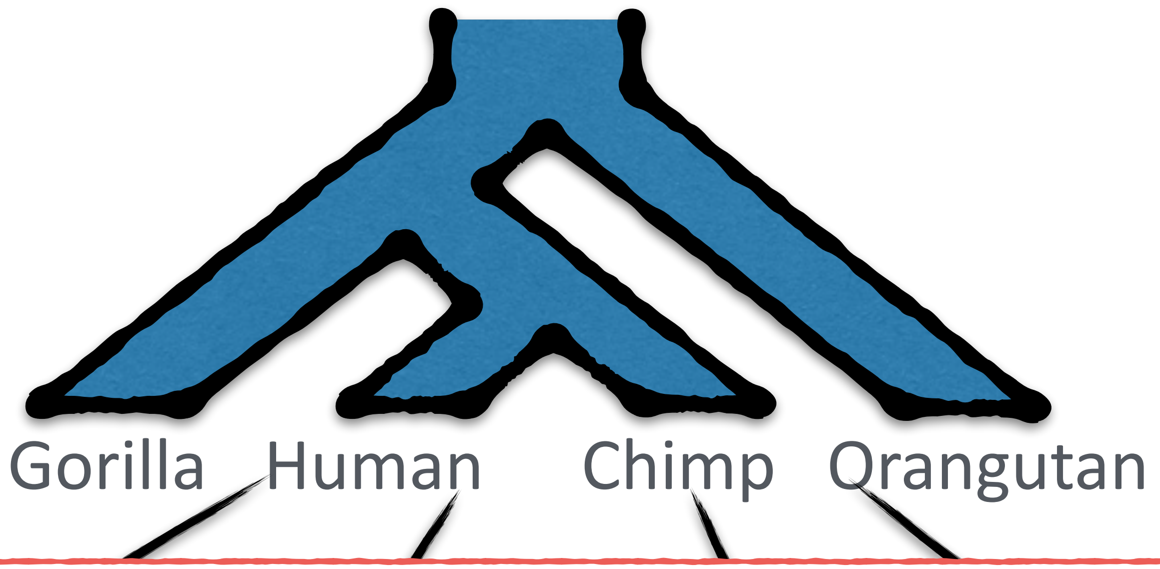


A gene tree

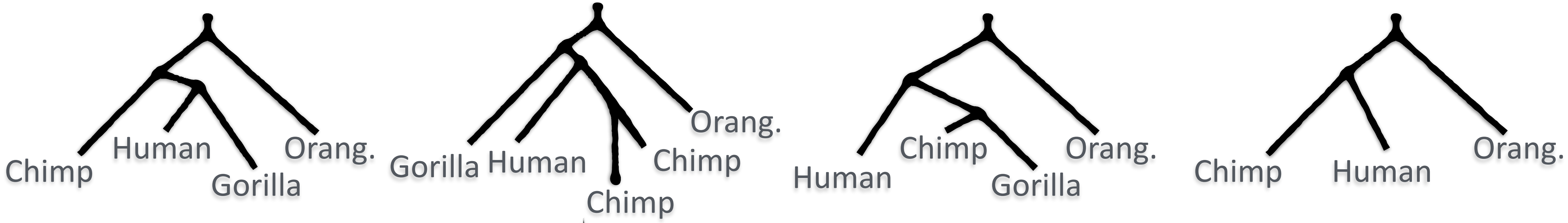
Causes of gene tree discordance include:

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

"c-gene":
recombination-free orthologous
stretches of the genome



Gene evolution model



Sequence evolution model

```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

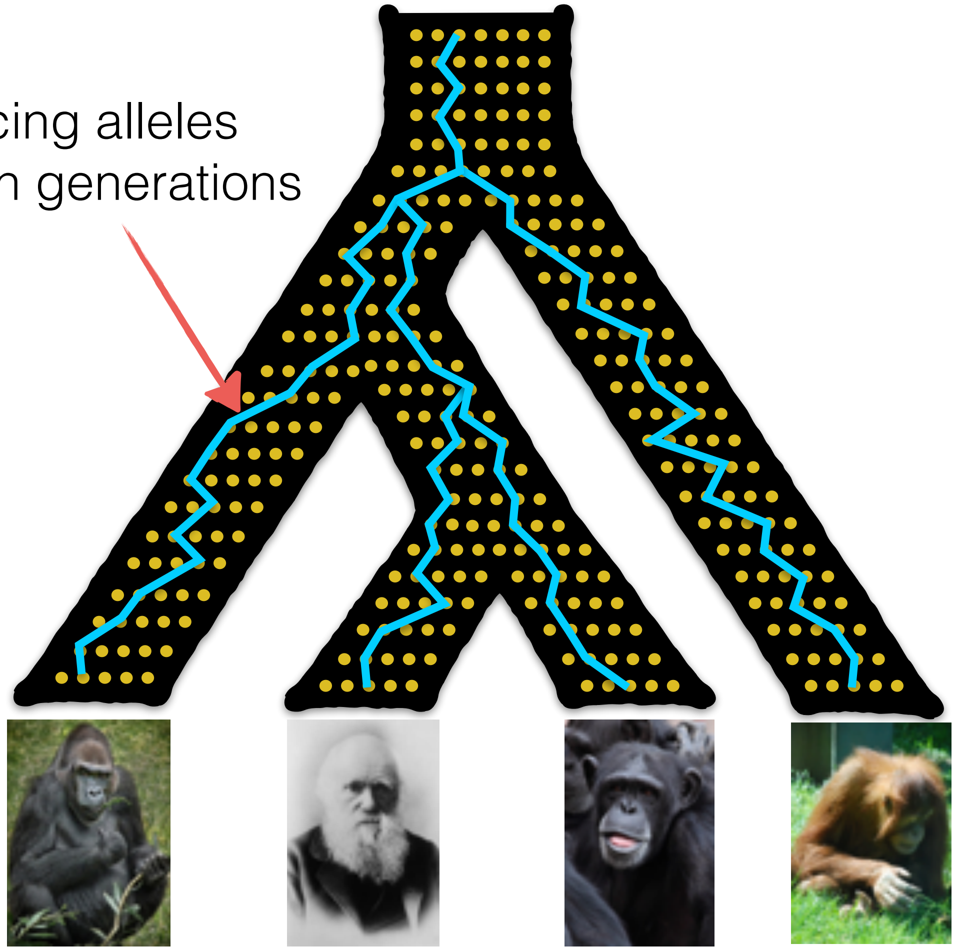
```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

Incomplete Lineage Sorting (ILS)

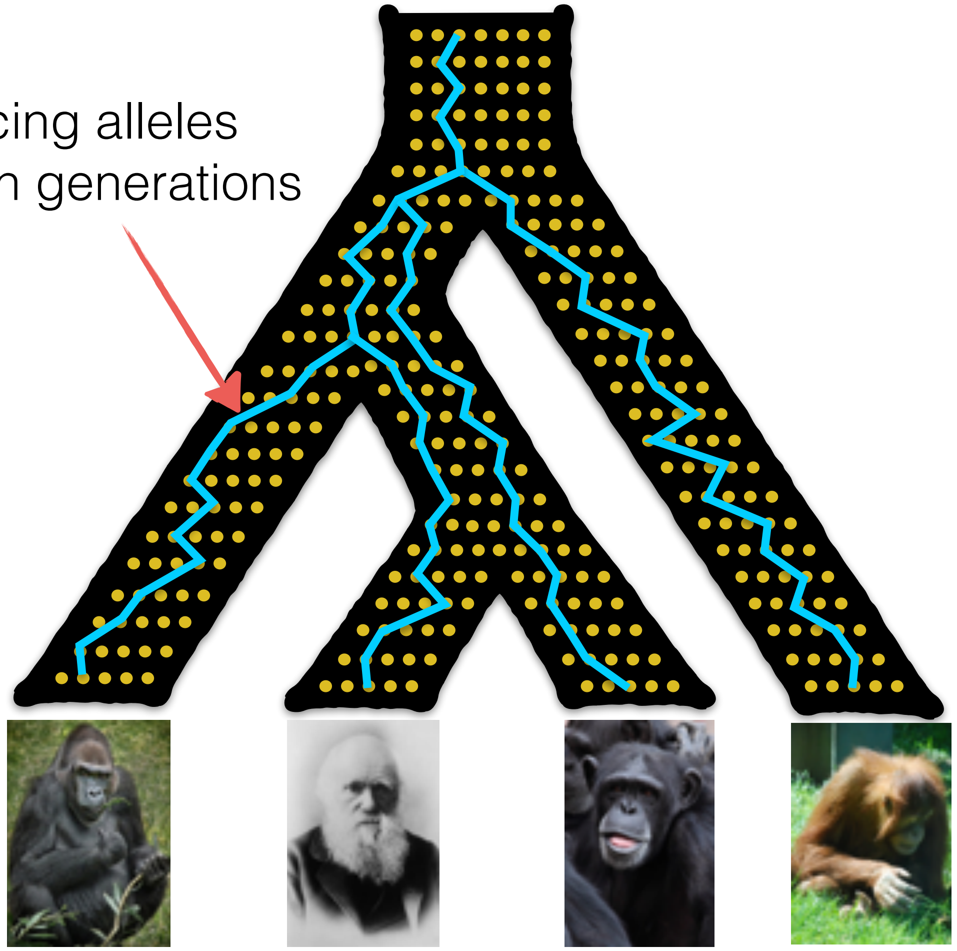
- The coalescent process extended to multiple species
  - Omnipresent; most likely for rapid radiations, like birds



Tracing alleles through generations

Incomplete Lineage Sorting (ILS)

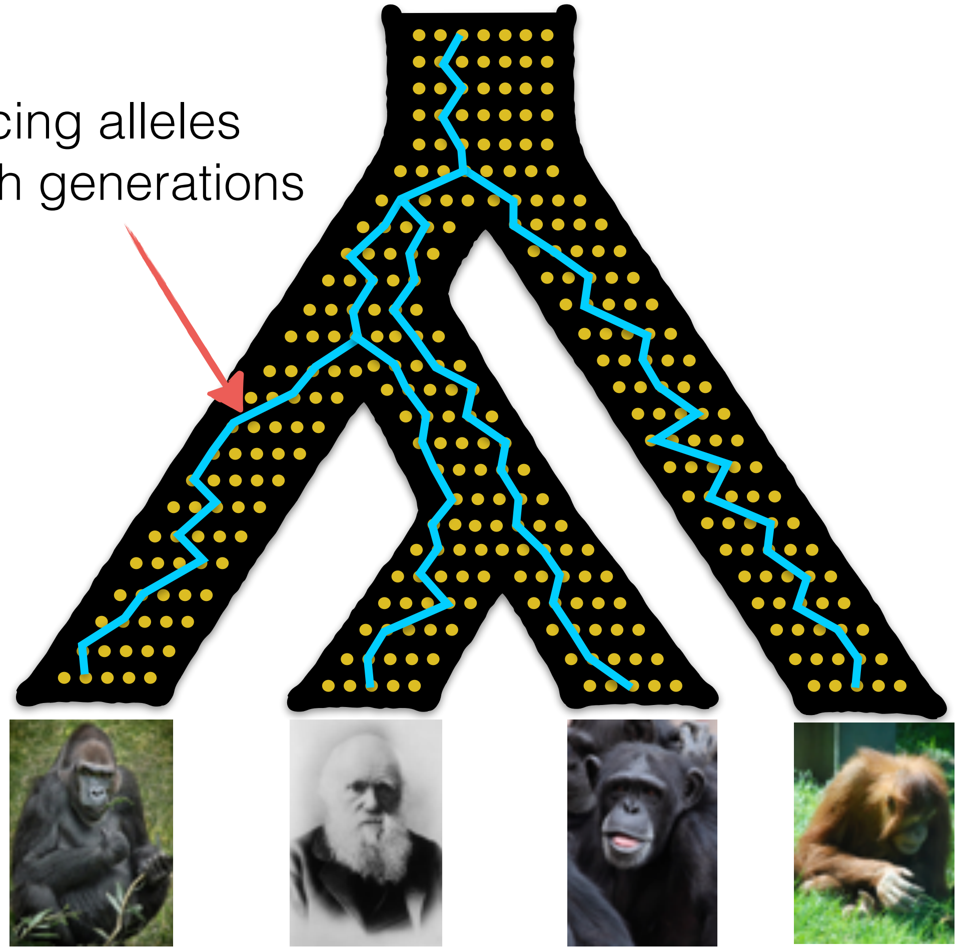
- The coalescent process extended to multiple species
  - Omnipresent; most likely for rapid radiations, like birds



Tracing alleles through generations

Incomplete Lineage Sorting (ILS)

- The coalescent process extended to multiple species
    - Omnipresent; most likely for rapid radiations, like birds
- 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 2005]



Tracing alleles through generations

Challenges

- What is a gene or a species and how do we find them?

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?

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: phylogenomics is hard. Dealing with multi-locus datasets and complex evolutionary processes often requires approximations.

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: phylogenomics is hard. Dealing with multi-locus datasets and complex evolutionary processes often requires approximations.

- Reliability and interpretation

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: phylogenomics is hard. Dealing with multi-locus datasets and complex evolutionary processes often requires approximations.
- Reliability and interpretation
- Catching up with new types of data

Challenges

- What is a gene or a species and how do we <u>find</u> them?

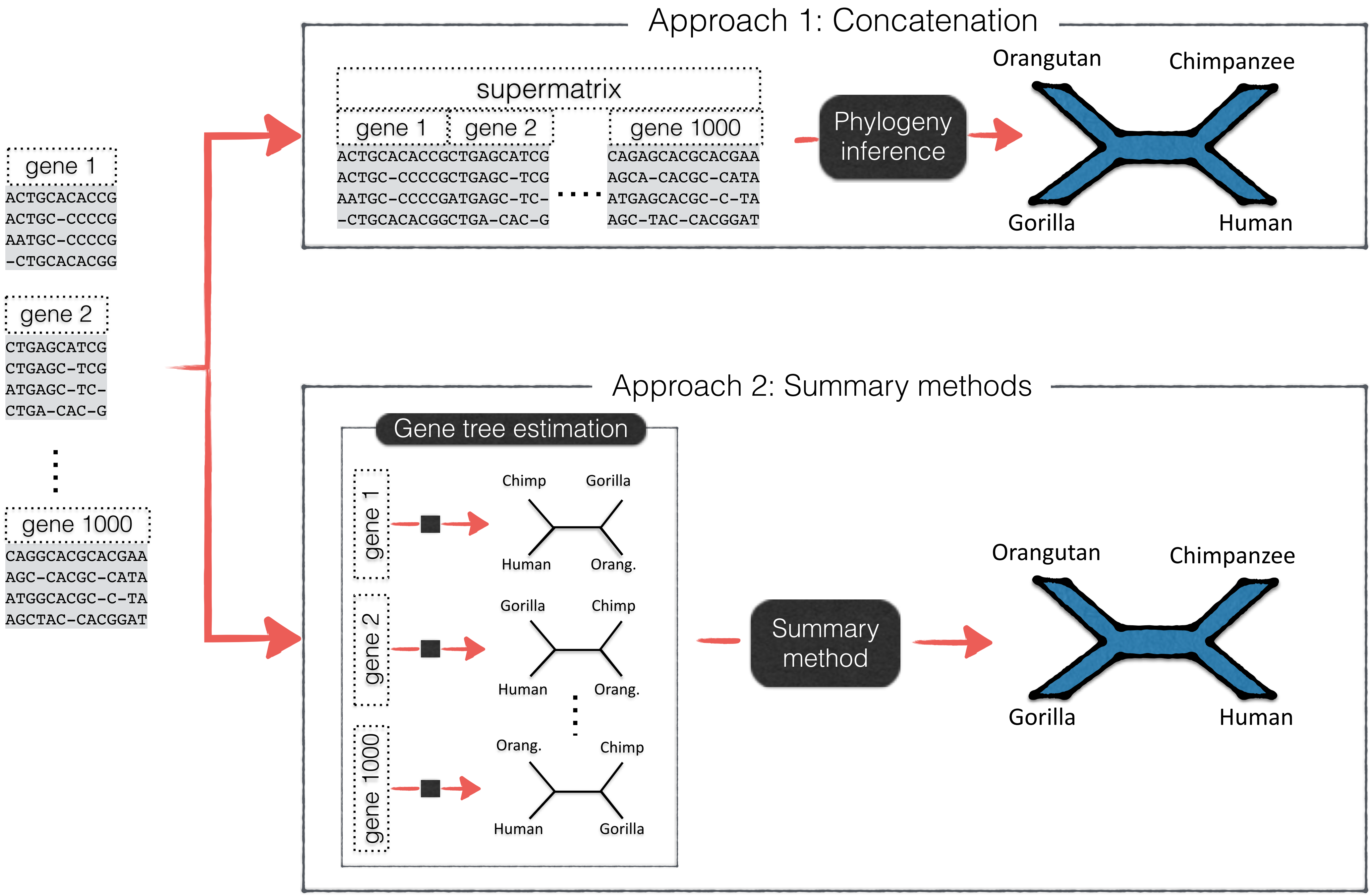
- Modeling: multiple evolutionary processes operate together, sometimes creating patterns that are hard to distinguish. How do we <u>untangle</u> them?

- **Inference: phylogenetics is hard. Dealing with multi-locus datasets and complex evolutionary processes is often requires approximations.**

- Reliability and interpretation

- Catching up with new types of data

Multi-gene species tree estimation



gene 1

ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG

gene 2

CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G

⋮

gene 1000

CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

Approach 1: Concatenation

supermatrix

gene 1 | gene 2 | gene 1000

ACTGCACACCGCTGAGCATCG … CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG … AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC- … ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G … AGC-TAC-CACGGAT

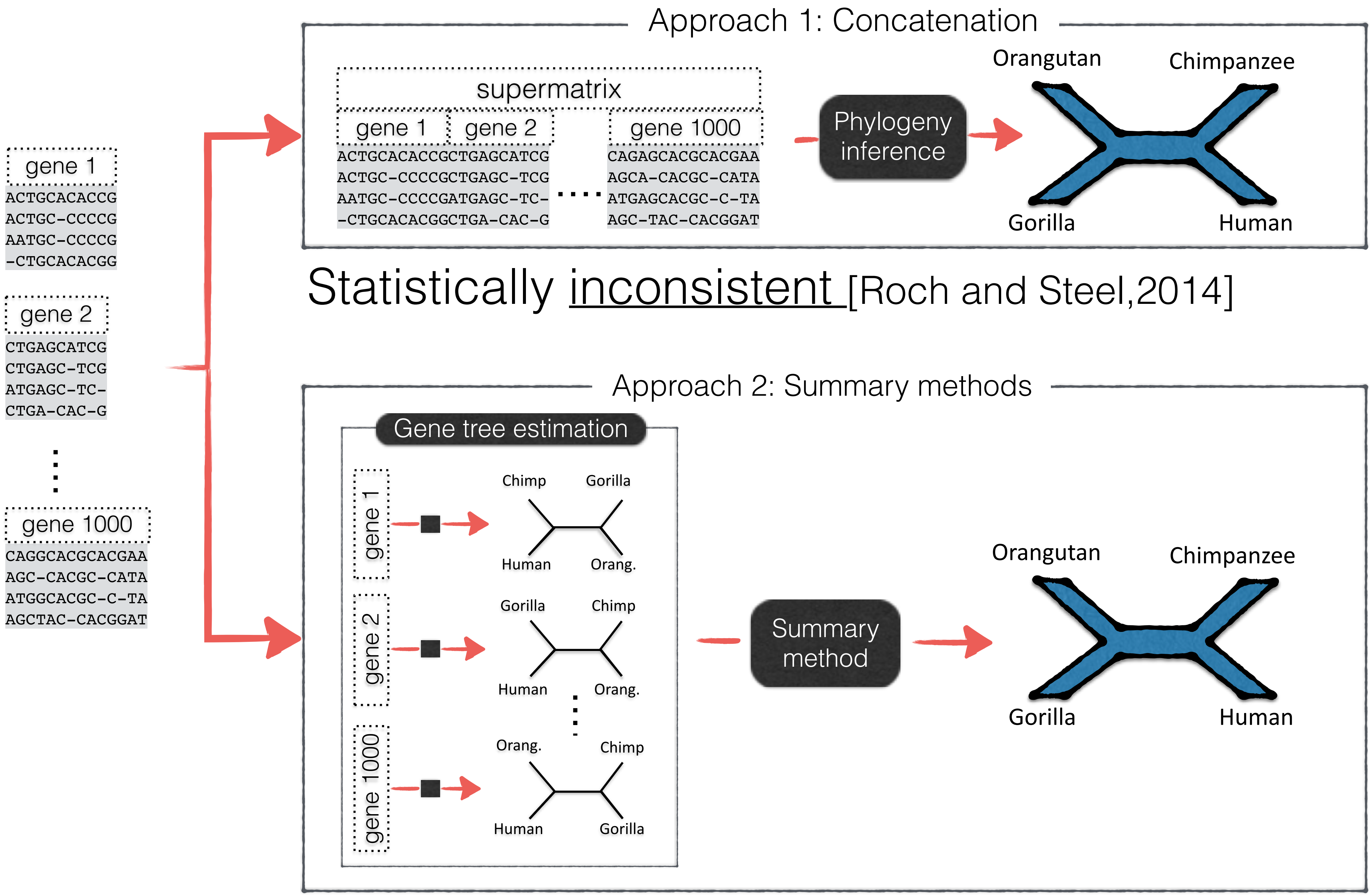
Phylogeny inference

Approach 2: Summary methods

Gene tree estimation

Summary method

Multi-gene species tree estimation



gene 1

ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG

gene 2

CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G

⋮

gene 1000

CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

Approach 1: Concatenation

supermatrix

gene 1 | gene 2 | gene 1000

ACTGCACACCGCTGAGCATCG .... CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG .... AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC- .... ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G .... AGC-TAC-CACGGAT

Phylogeny inference

Statistically inconsistent [Roch and Steel,2014]

Approach 2: Summary methods

Gene tree estimation

gene 1

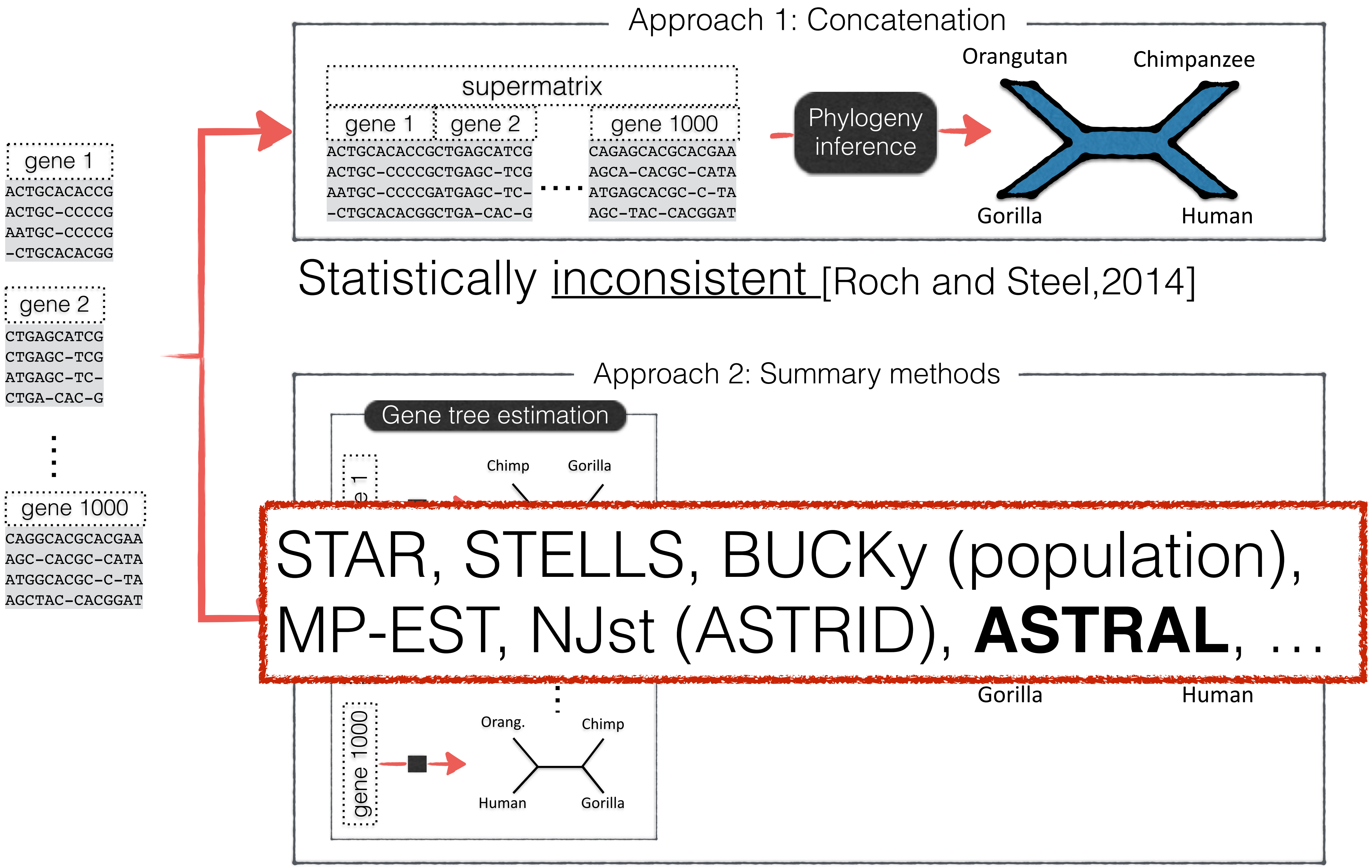
gene 2

gene 1000

Summary method

Can be statistically consistent given true gene trees

Multi-gene species tree estimation

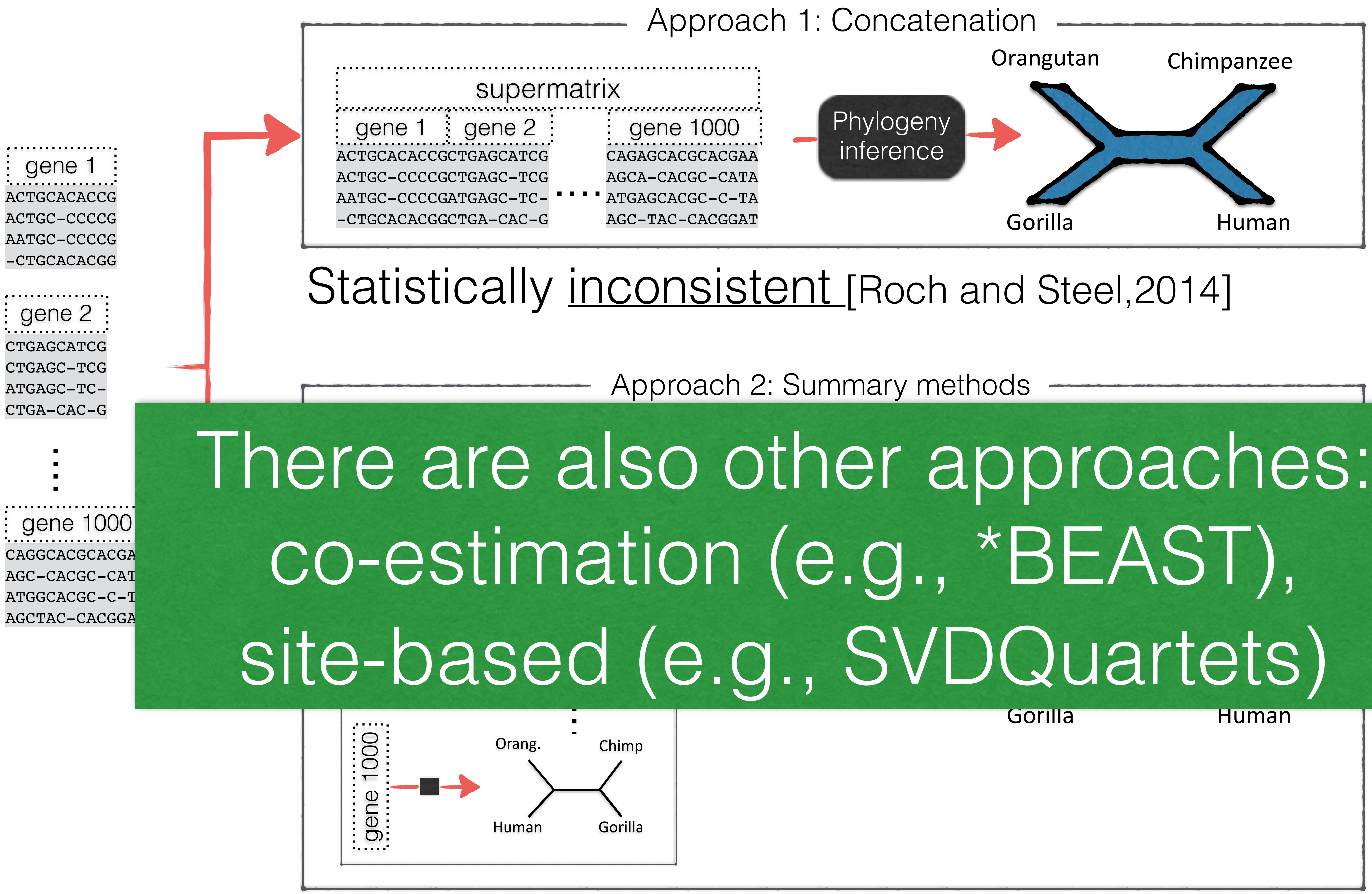


Statistically inconsistent [Roch and Steel,2014]

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

Can be statistically consistent given true gene trees

Multi-gene species tree estimation



Statistically inconsistent [Roch and Steel,2014]

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

Can be statistically consistent given true gene trees

ASTRAL

- **Input:** A set of inferred unrooted gene trees
- **Output:** A species tree with branch lengths in coalescent units and branch support values

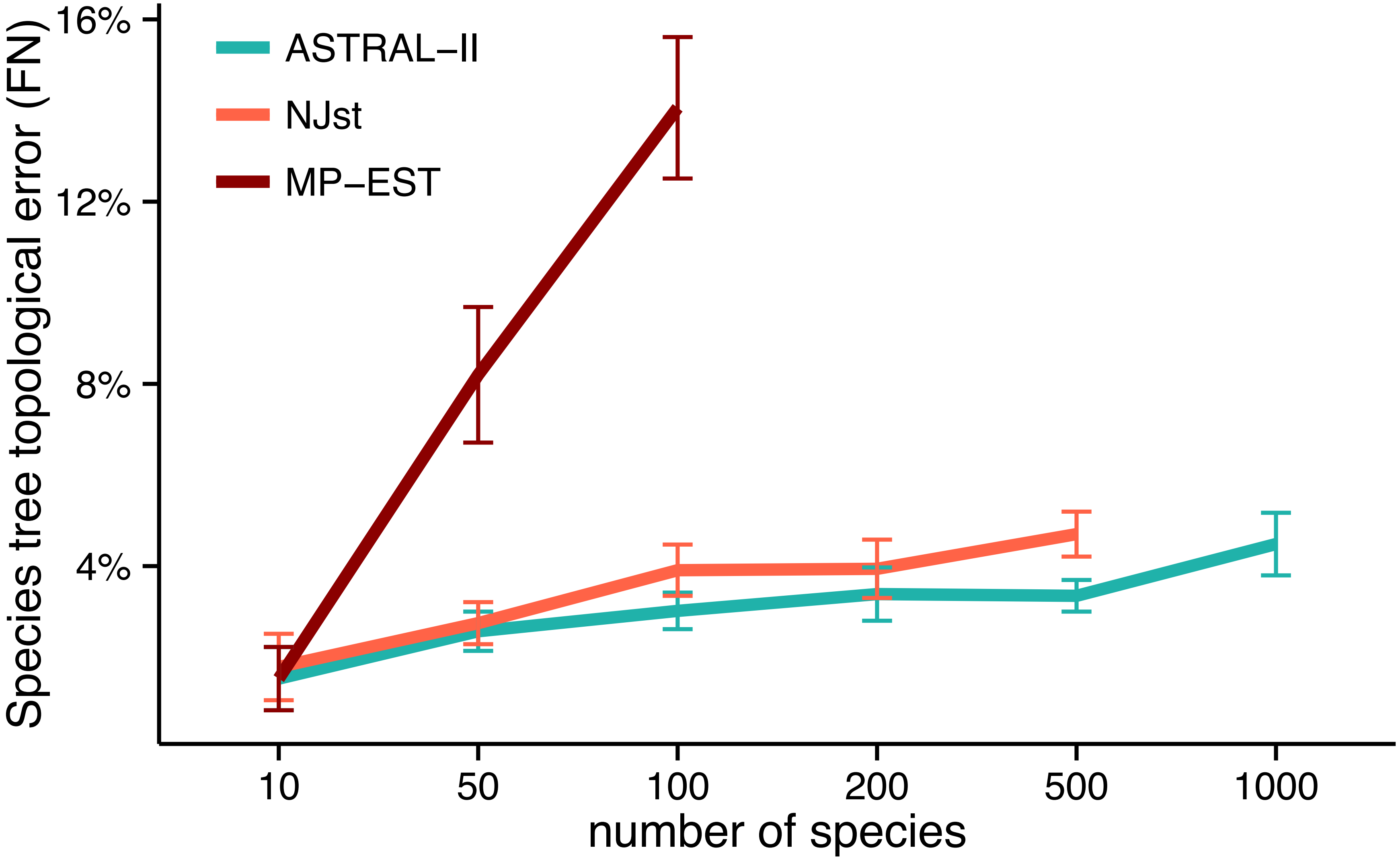
ASTRAL

- **Input:** A set of inferred unrooted gene trees
- **Output:** A species tree with branch lengths in coalescent units and branch support values
- **Approach**: try to find the species tree that shares the maximum number of quartet trees with input gene trees

ASTRAL

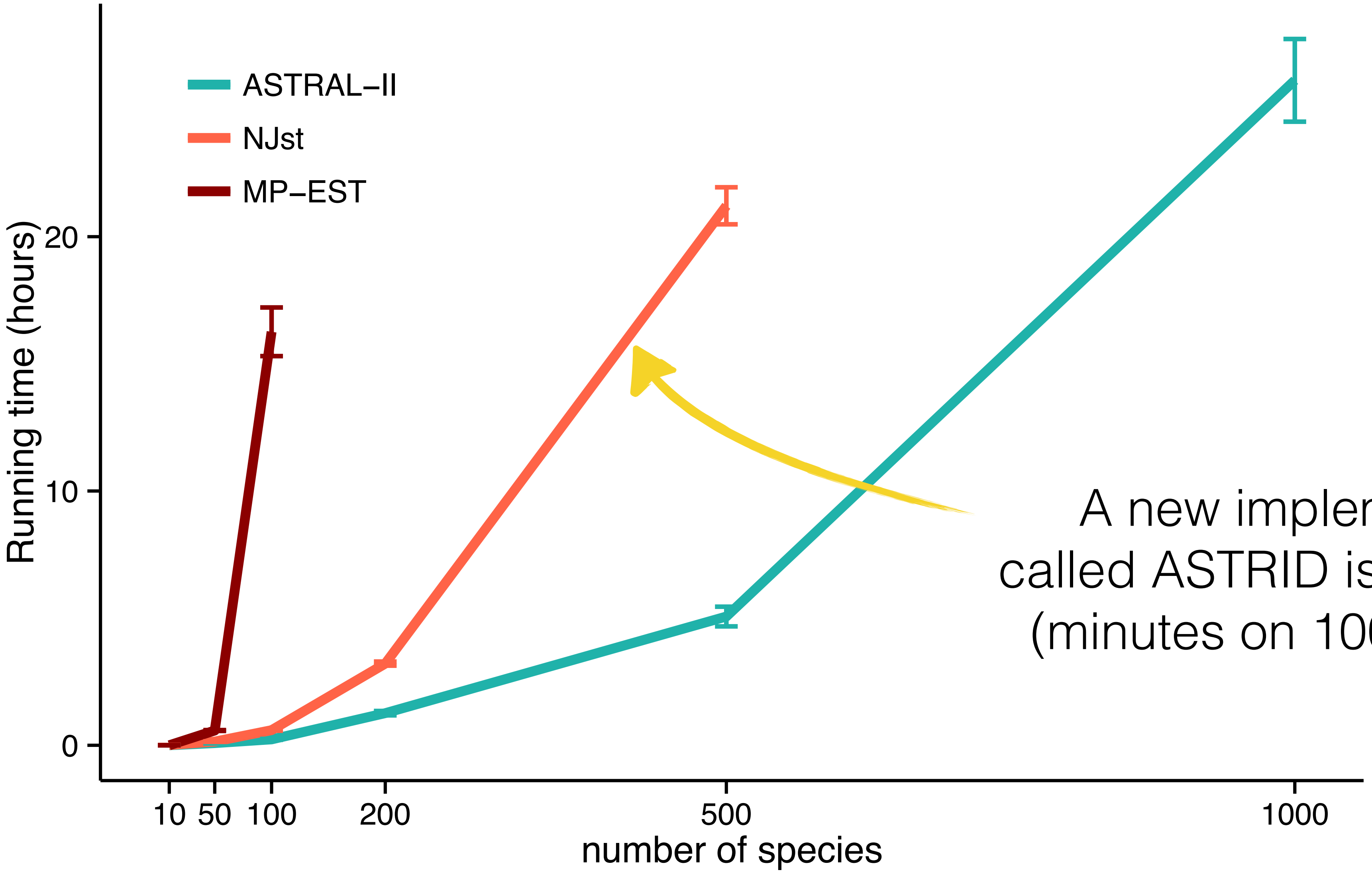
- **Input:** A set of inferred unrooted gene trees
- **Output:** A species tree with branch lengths in coalescent units and branch support values
- **Approach**: try to find the species tree that shares the maximum number of quartet trees with input gene trees
- **Designed for:**
  - Accuracy (established in simulation studies)
  - Scalability: the default version runs on a thousand genes from a thousand species in a day
    —> Important for next phases of B10K

ASTRAL: accurate and scalable



1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

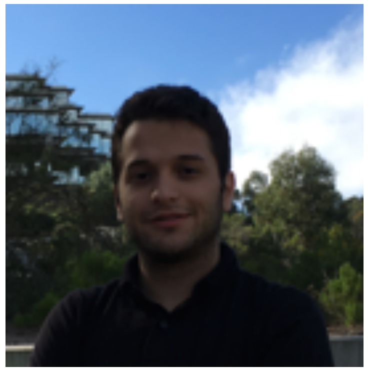
Running time as function of # species



A new implementation called ASTRID is much faster (minutes on 1000 species)

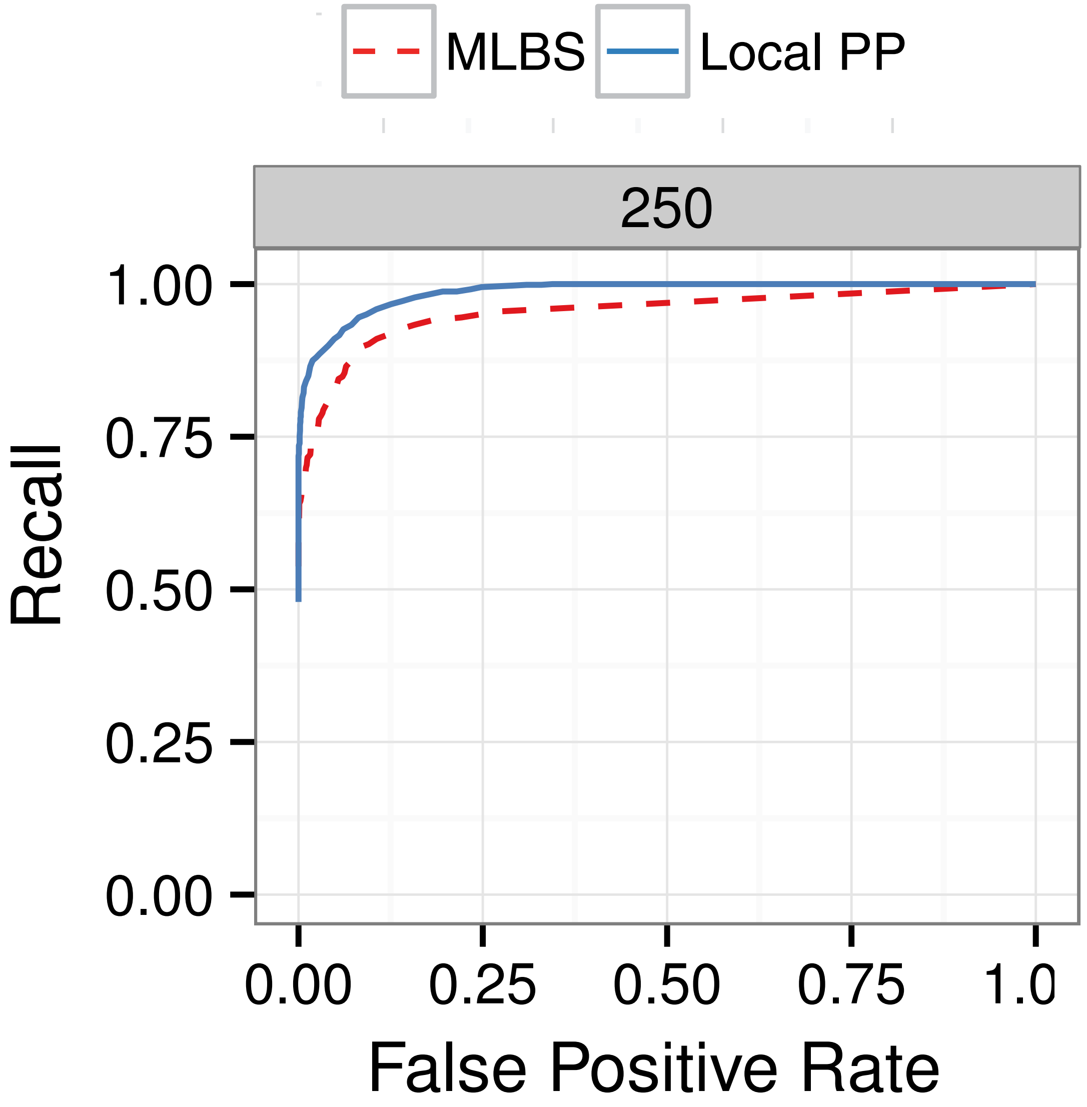
1000 genes, "medium" levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

Local branch support



Erfan Sayyari

- Use frequency of quartets defined *around* each species tree branch and some strong assumptions to compute statistical support for each species tree branch
- Extremely scalable.
  - Doesn't need bootstrapped gene trees
  - Minutes on 1K species
- More accurat than traditional bootstrapping



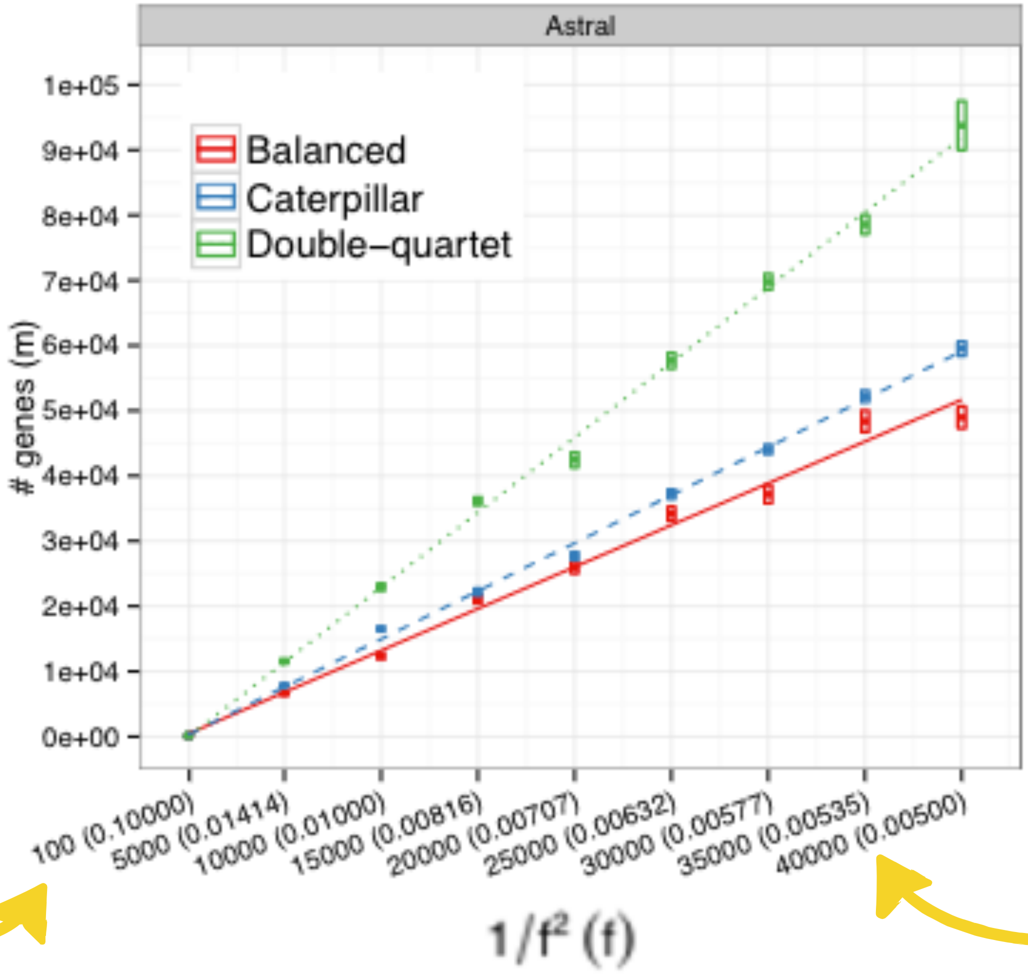
Avian simulated dataset
(48 taxa, 1000 genes)

[Sayyari, Mirarab, MBE 2016]

How many genes does ASTRAL need?

Depends on the branch length (f) and the number of species (n):

$$log(n)\, f^{-2}$$



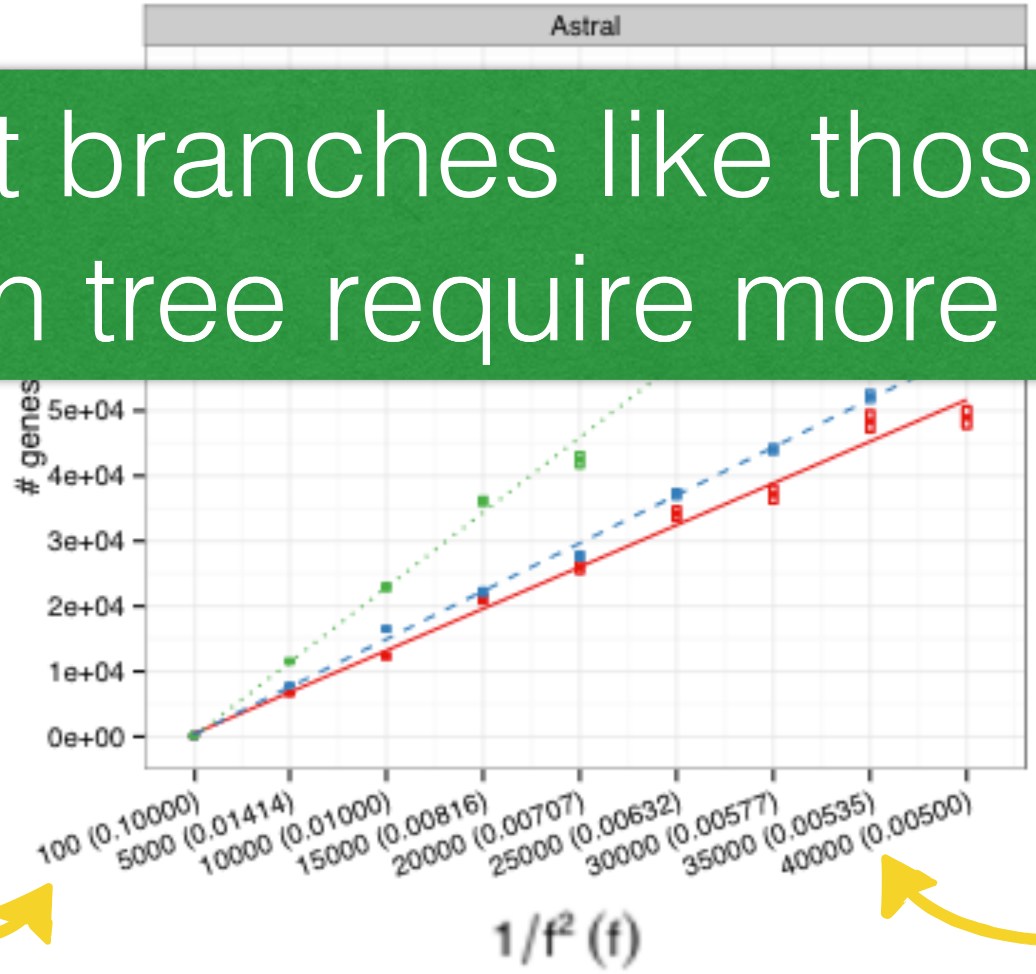
100,000 years

5000 years

How many genes does ASTRAL need?

Depends on the branch length (f) and the number of species (n):

$$log(n)\, f^{-2}$$



Short branches like those in the avian tree require more genes!

100,000 years

5000 years

Ongoing improvements to ASTRAL

- A GPU implementation is 10-20X faster (less than an hour on 1000 species).
- Improved measures of support (fewer and weaker assumptions)
- Better ways of dealing with multiple individuals from the same species
- Divide-and-conquer to enable analyzing many tens of thousands of species

Unsolved challenges for the next phase of B10K

- **Hybridization:** Inferring species networks is doable now, but not on the scales targeted by B10K

- **Gene tree error/uncertainty :** gene tree uncertainty is notoriously high for avian genomes.

    - We previously proposed statistical binning for this problem. It is not clear that binning will scale to hundreds of species

- **Standard pipelines for quality control:** for example, for dealing with fragmentary data, codon bias, etc.

- And many more …