Phylogenomic tree construction

ISMB Tutorial: Computational methods for comparative regulatory genomics

Lecture 3

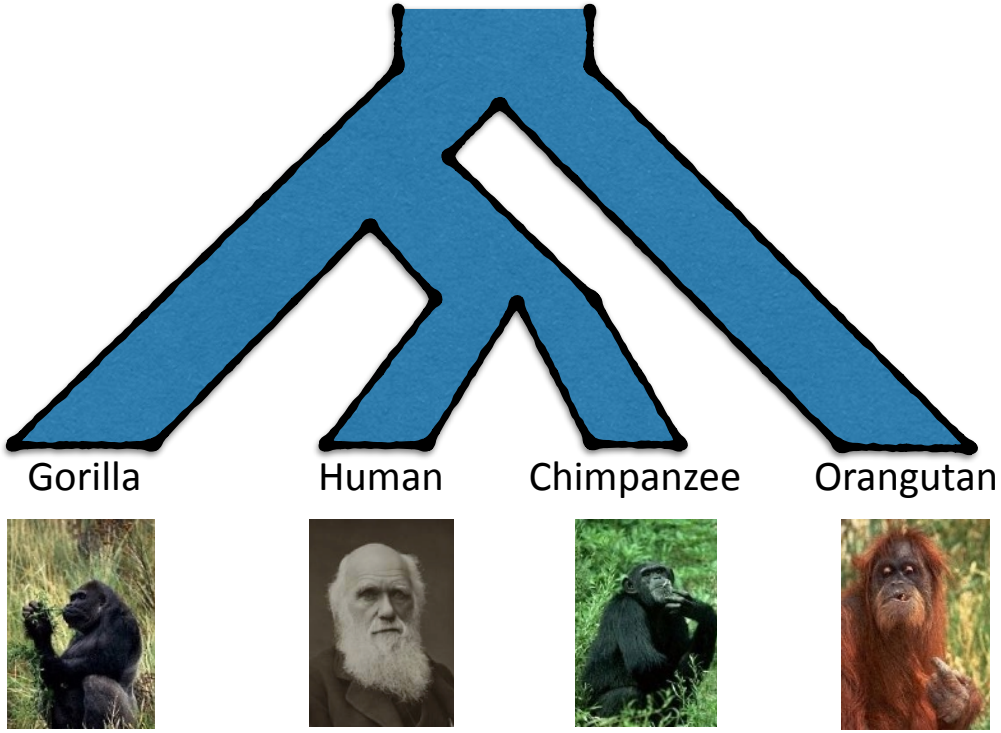
Siavash Mirarab

smirarab@ucsd.edu

Topics in this lecture

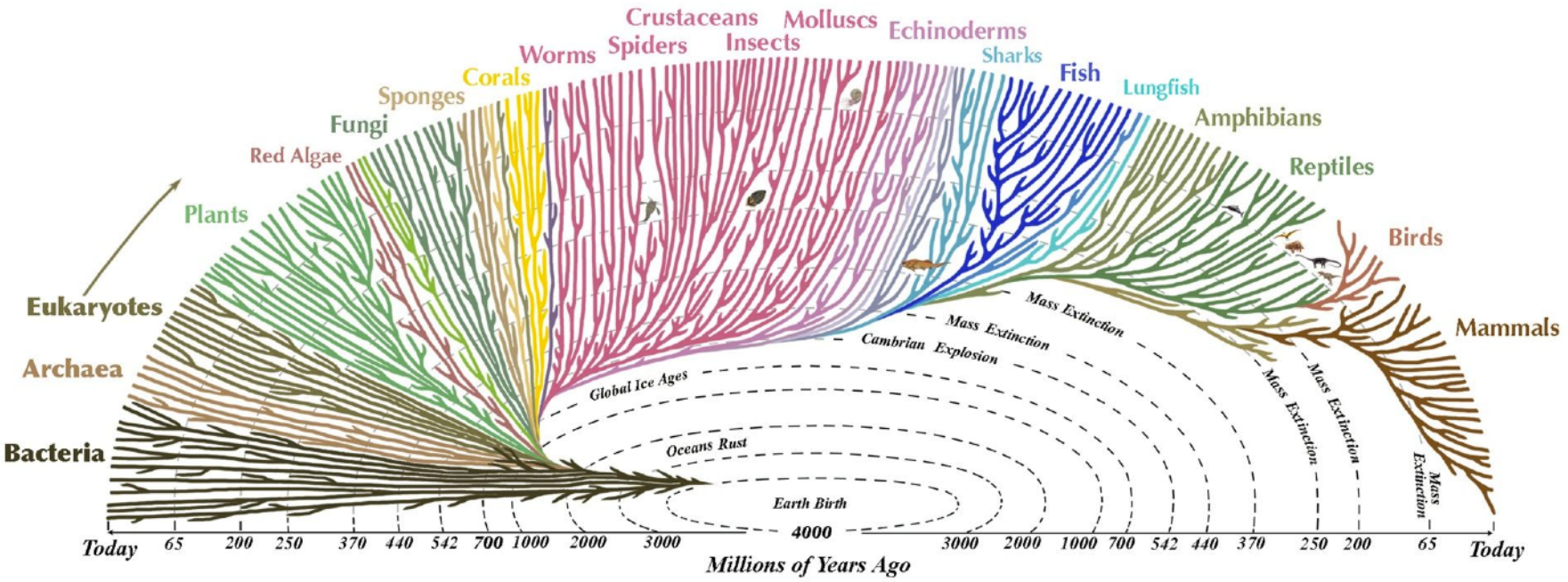
- Phylogenomics: premise and challenges
- Species trees versus gene trees
  - Causes for discordance
- Phylogenetic inference despite discordance
  - Questions
  - Models
  - Method choices

Phylogeny



Tree of life

"***Nothing in biology makes sense except in the light of evolution.***" Dobzhansky, 1973

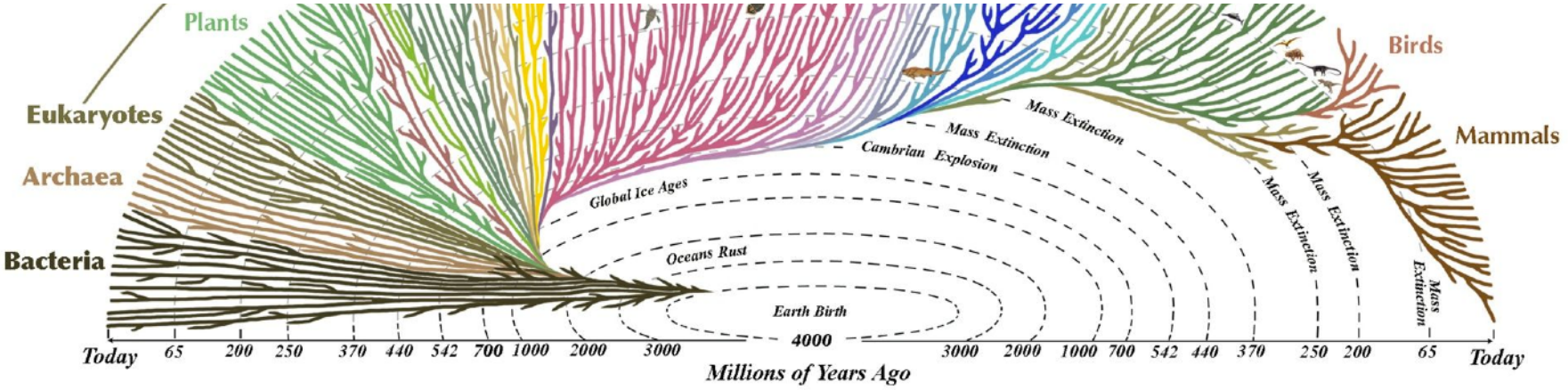


source: http://www.evogeneao.com/

Tree of life

"***Nothing in biology makes sense except in the light of evolution.***" Dobzhansky, 1973

"***Nothing in evolution makes sense except in the light of phylogeny.***" multiple coinage

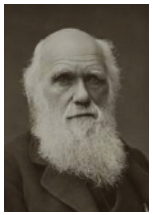


source: http://www.evogeneao.com/

Phylogenetic reconstruction from DNA

Gorilla

ACTGCACACCG



Human

ACTGCCCCCG

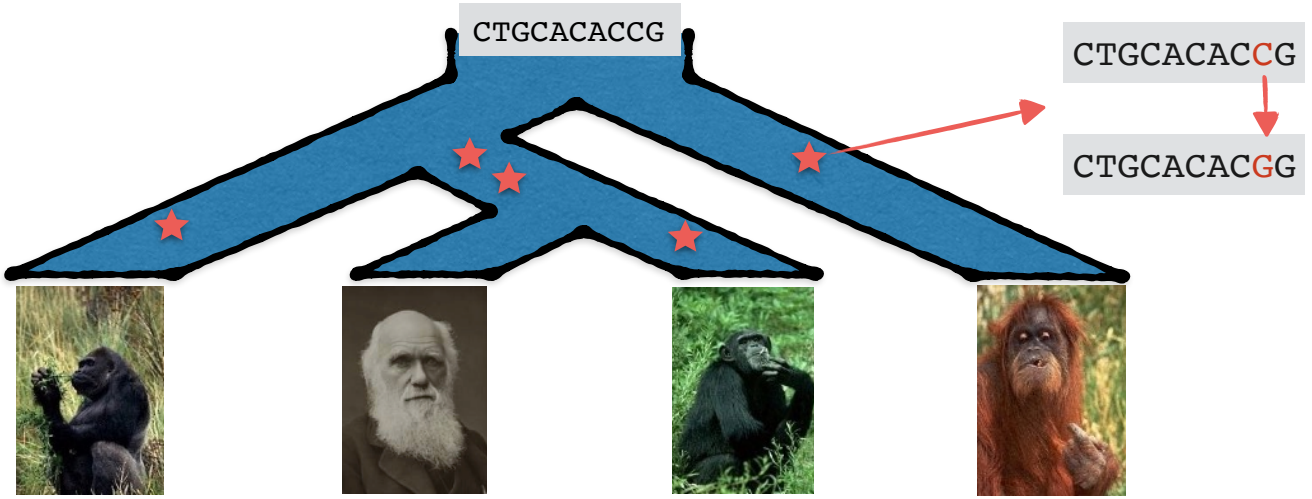
Chimpanzee

AATGCCCCCG

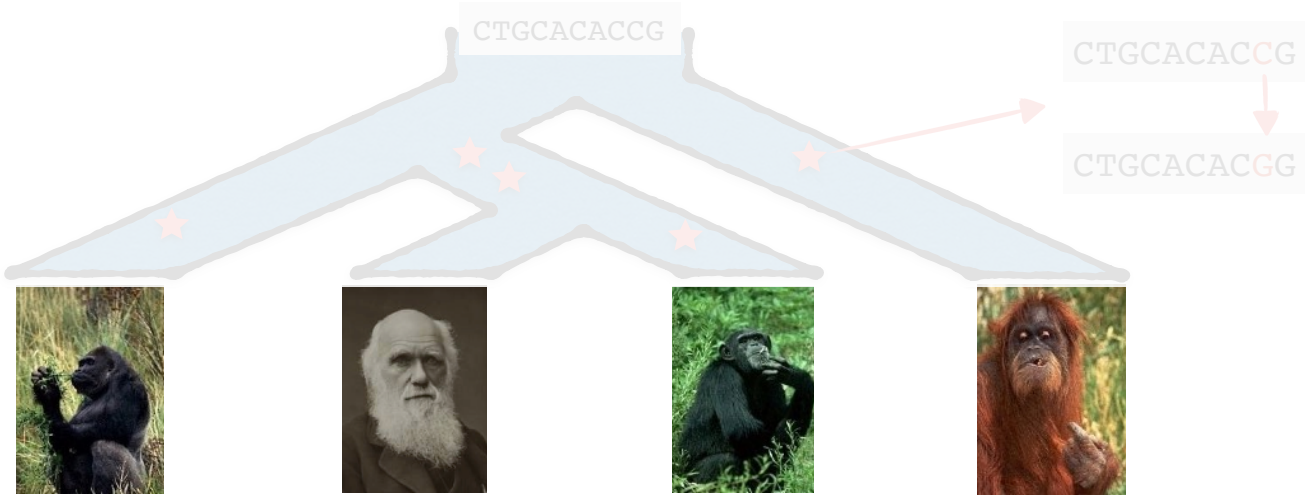
Orangutan

CTGCACACGG

Phylogenetic reconstruction from DNA

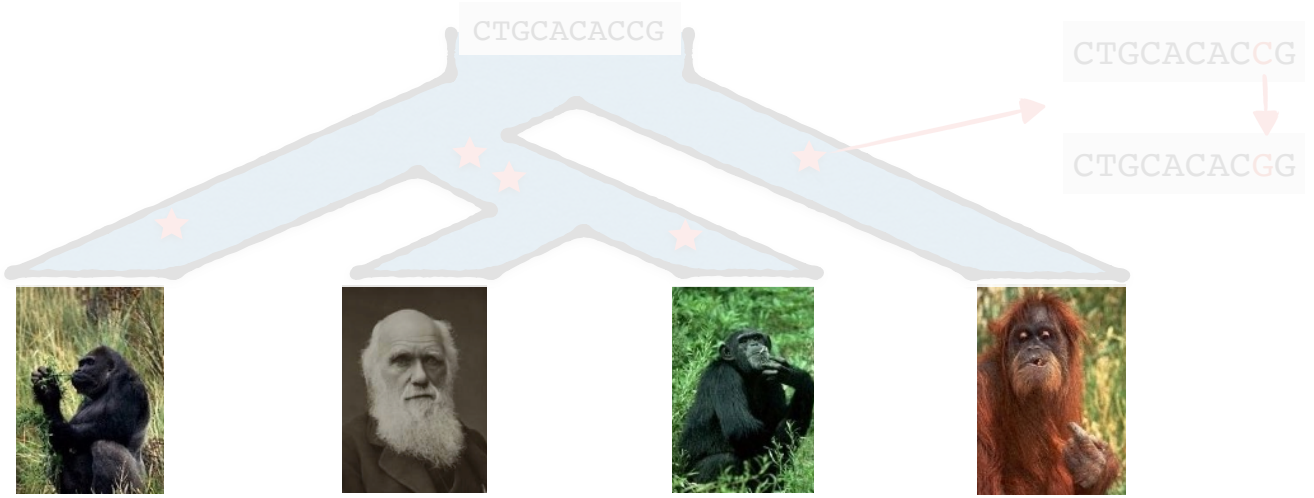


Phylogenetic reconstruction from DNA



Gorilla	Human	Chimpanzee	Orangutan
ACTGCACACCG	ACTGCCCCCG	AATGCCCCCG	CTGCACACGG

Phylogenetic reconstruction from DNA



Gorilla
ACTGCACACCG

Human
ACTGCCCCCG

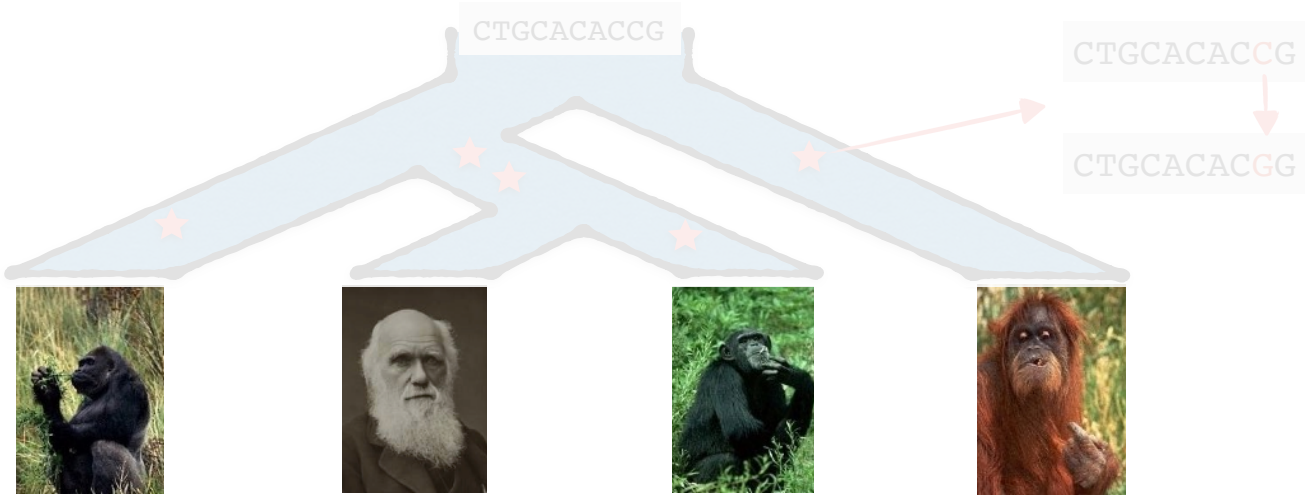
Chimpanzee
AATGCCCCCG

Orangutan
CTGCACACGG

Gorilla	ACTGCACACCG
Human	ACTGC-CCCCG
Chimpanzee	AATGC-CCCCG
Orangutan	-CTGCACACGG

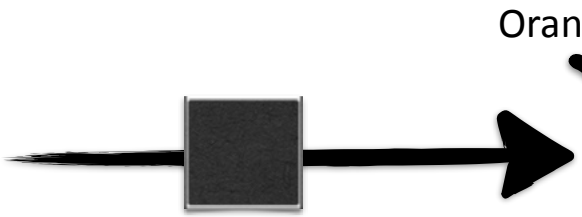
$$D$$

Phylogenetic reconstruction from DNA

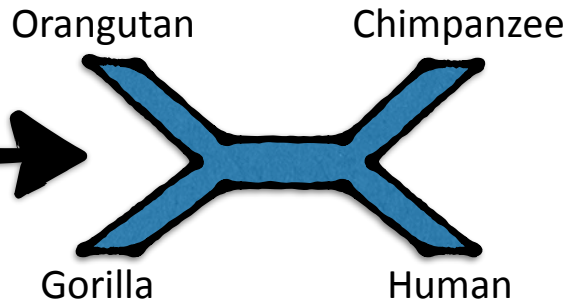


Gorilla	ACTGCACACCG
Human	ACTGC-CCCCG
Chimpanzee	AATGC-CCCCG
Orangutan	-CTGCACACGG

D



$P(D|T)$

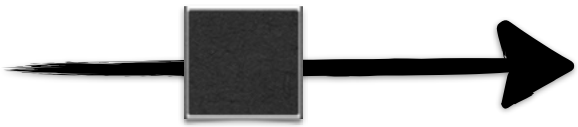


T

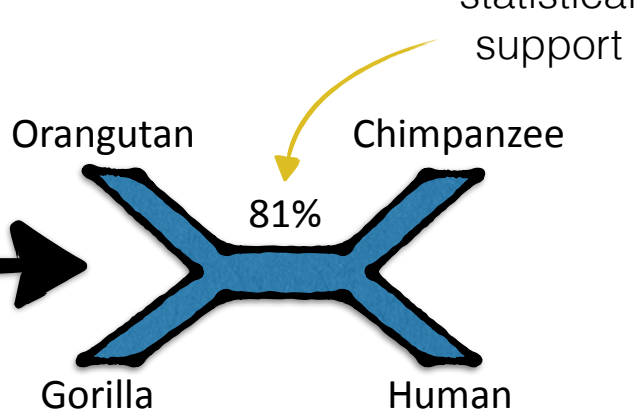
Phylogenetic reconstruction from DNA

Gorilla	ACTGCACACCG
Human	ACTGC-CCCCG
Chimpanzee	AATGC-CCCCG
Orangutan	-CTGCACACGG

D

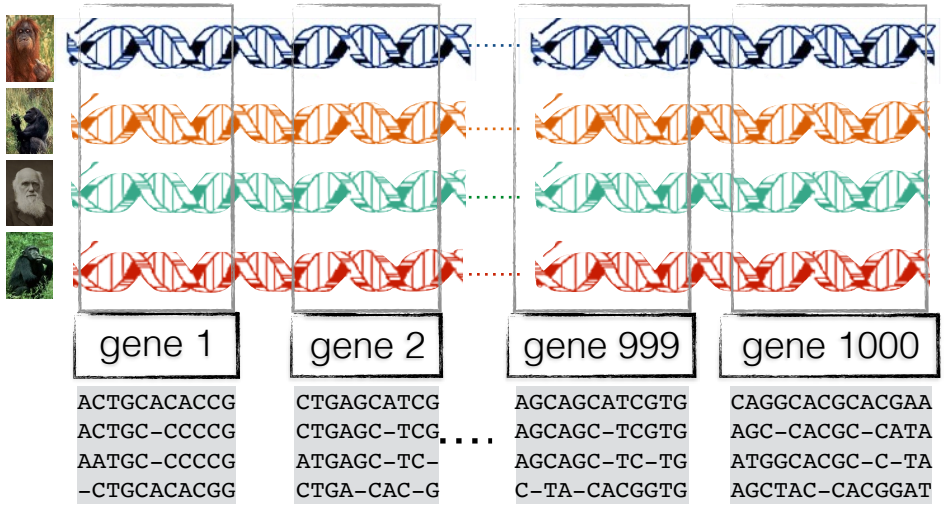


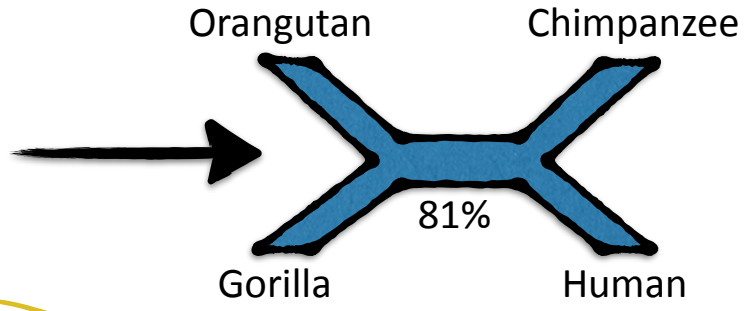
$P(D|T)$



T

Phylogenomics: promise

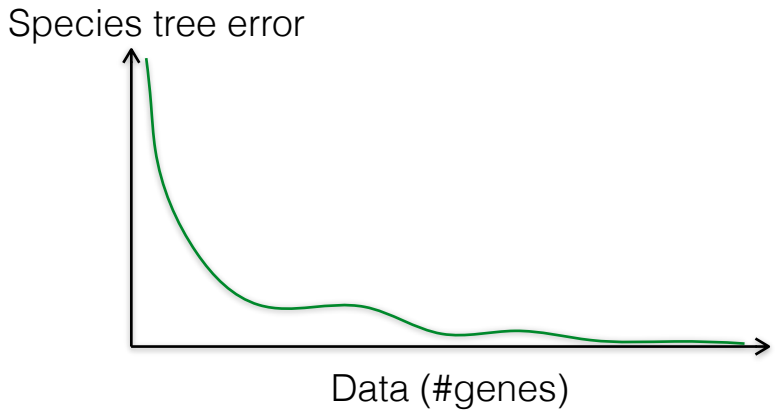




"gene" here simply means a (recombination-free) parts of the genome

more data ⟶

better inference



Phylogenomics: promise

nature
International journal of science

Evolution

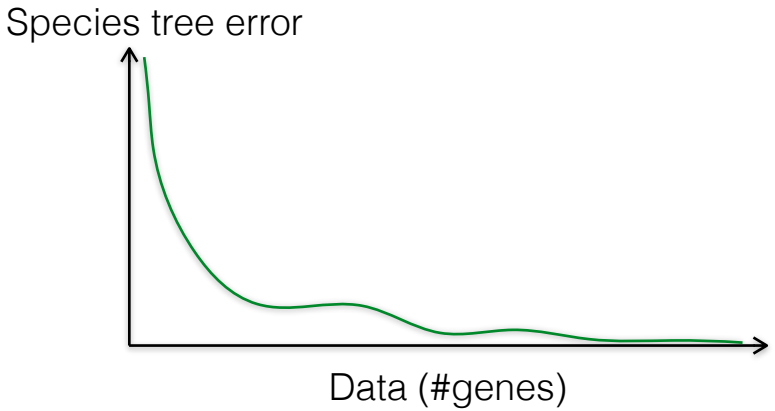
Ending incongruence

Henry Gee

Recovering the true evolutionary history of any group of organisms has seemed impossible. The availability of large amounts of genomic data promises an era in which the uncertainties are better constrained.

more data ⟶

better inference



Phylogenomics:
only a few years later

Trends in Genetics

Volume 22, Issue 4, April 2006, Pages 225–231

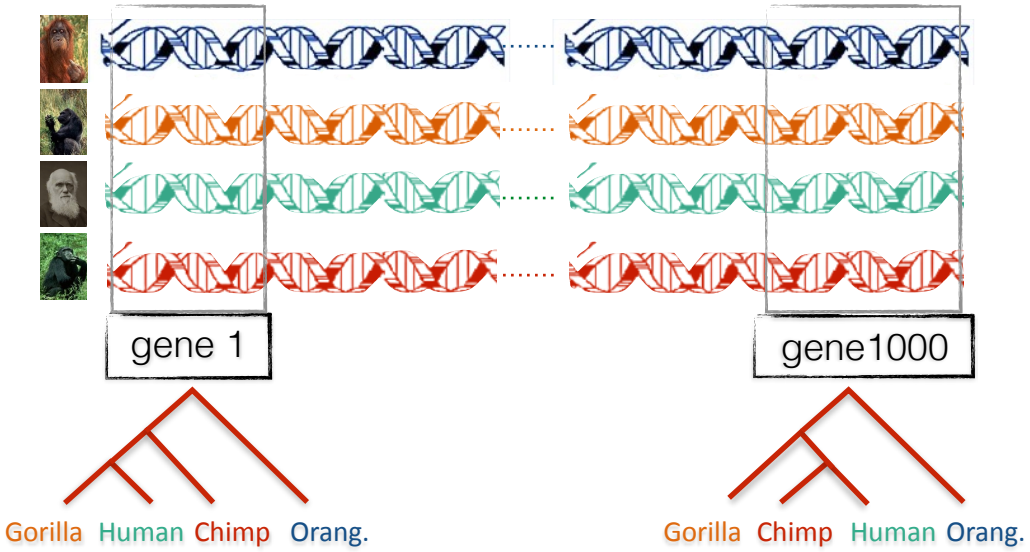
Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, Hervé Philippe

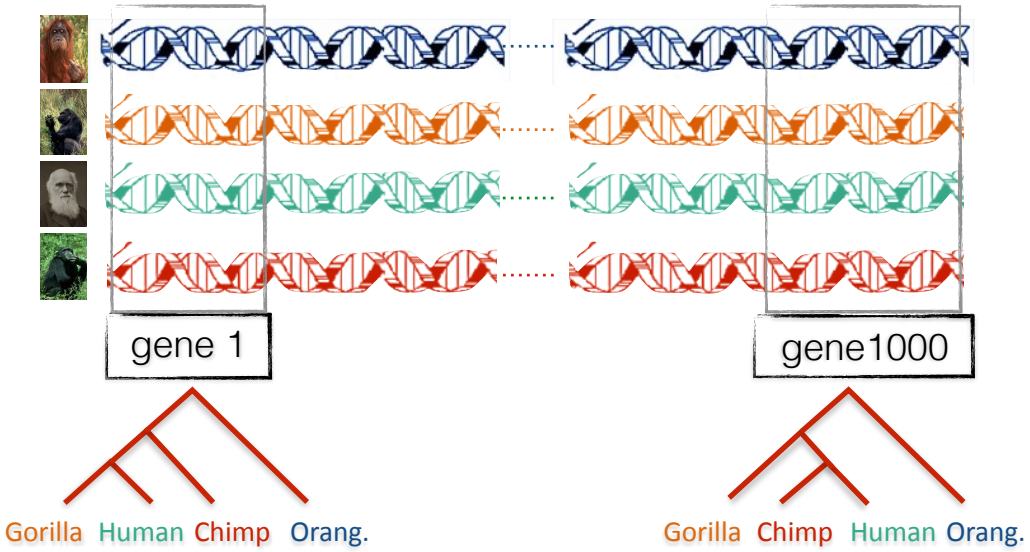
Show more

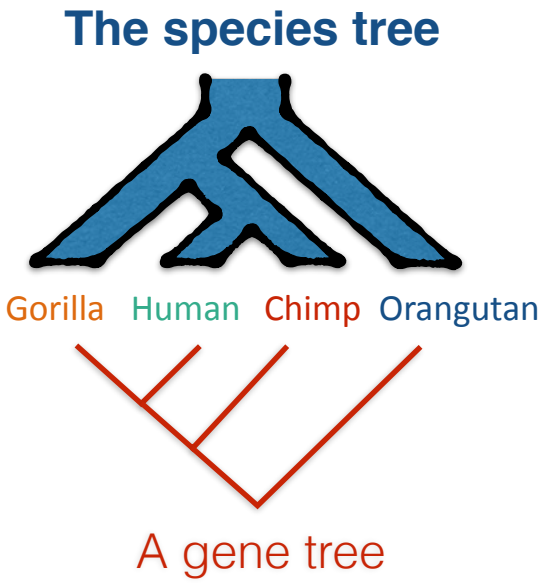
https://doi.org/10.1016/j.tig.2006.02.003

Gene tree discordance

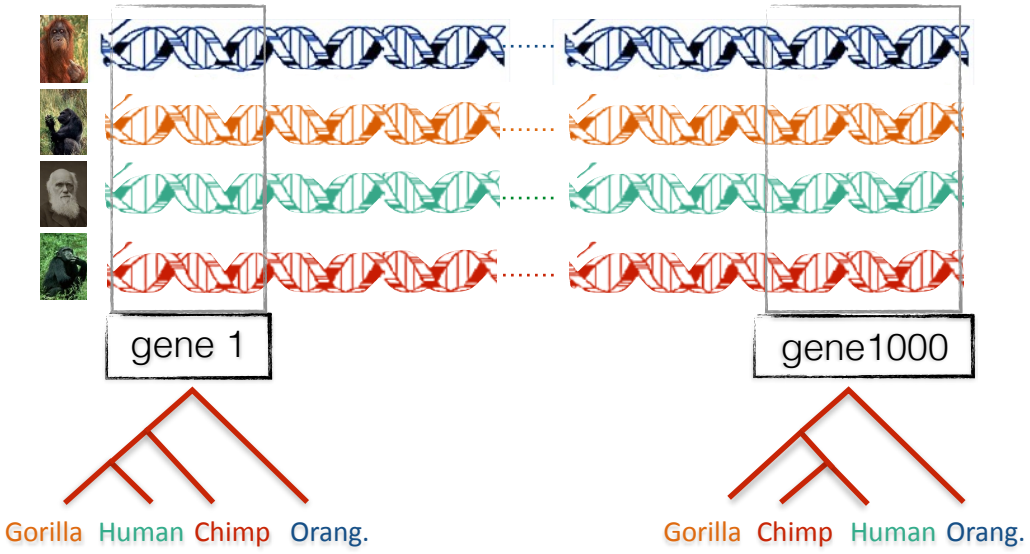


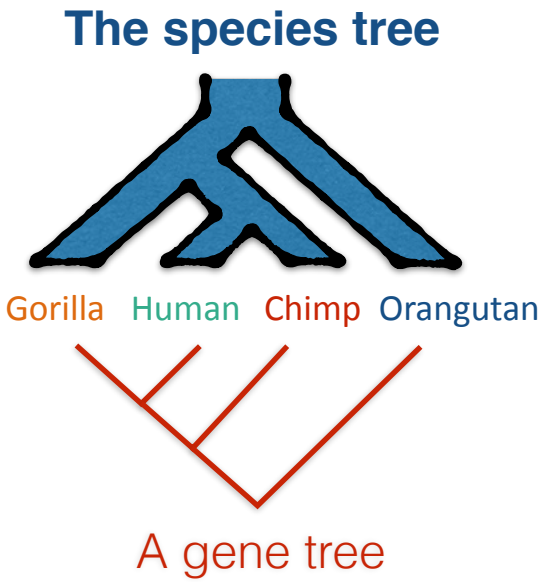
Gene tree discordance





Gene tree discordance



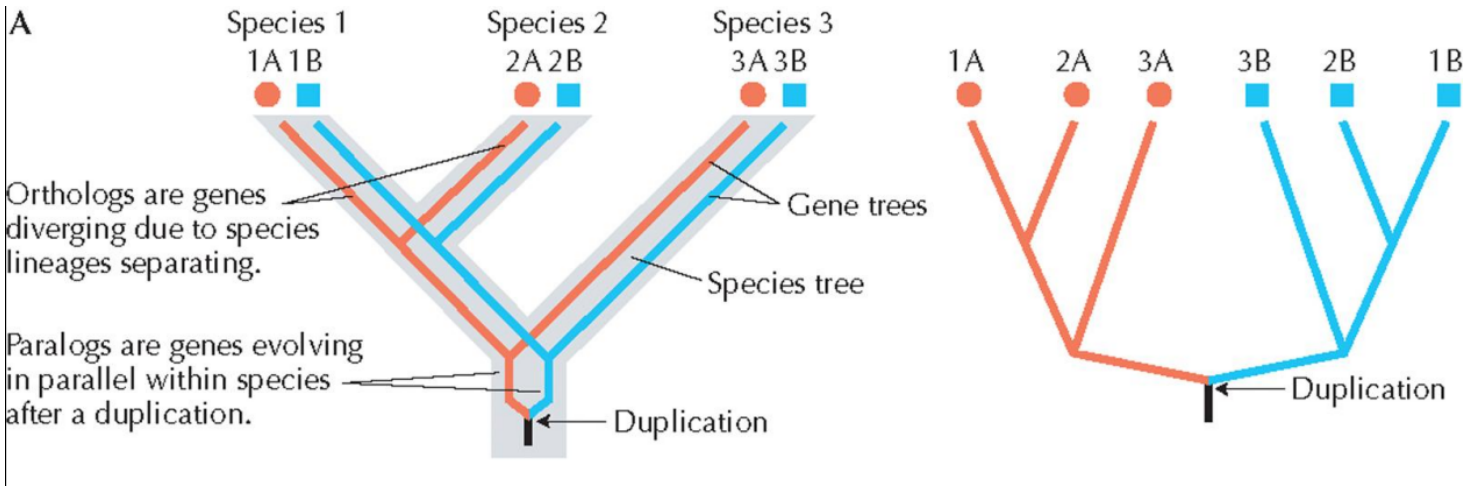


Causes of gene tree discordance include:

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

Gene duplication and loss

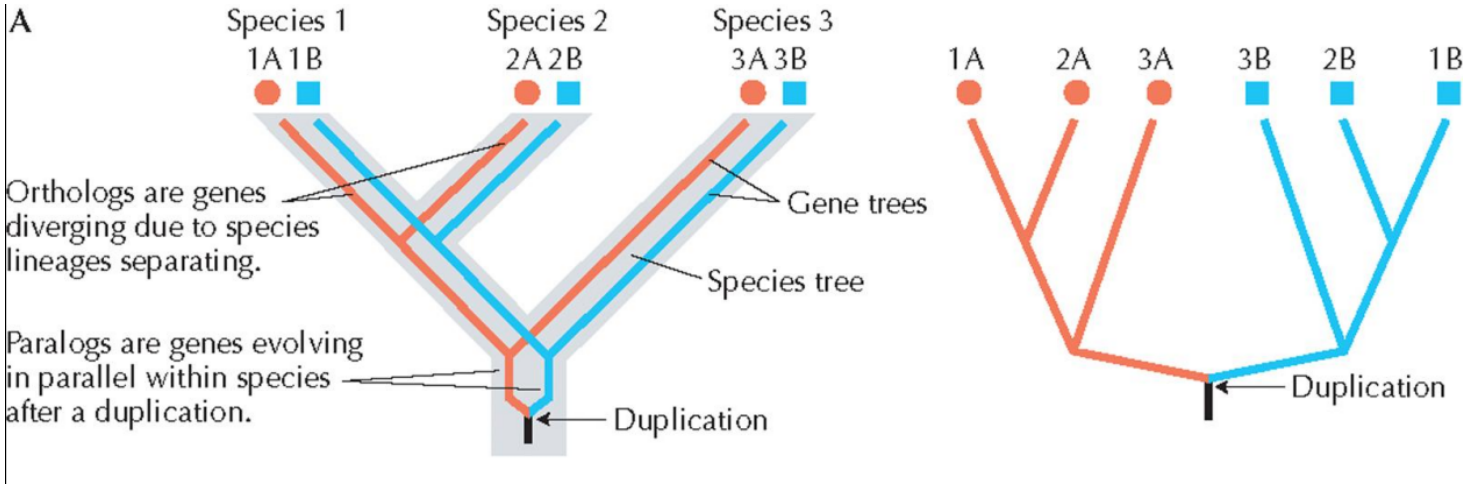
- In some species (e.g. plants) gene duplication and loss is rampant.



picture from evolution-textbook.org (Fig 5-20), redrawn from Eisen, Genome Research, 1998

Gene duplication and loss

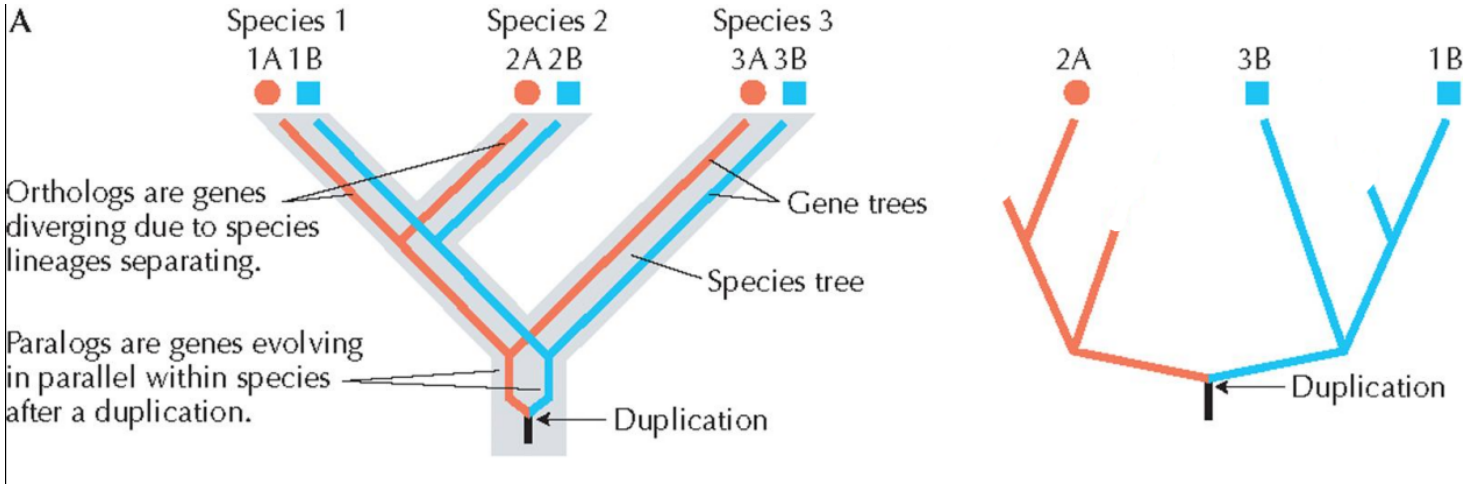
- In some species (e.g. plants) gene duplication and loss is rampant.
- Recall from Colin's talk:
  Paralogs: genes that diverged in a duplication event; e.g: 2A and 3B
  Orthologs: genes that diverged in a speciation event; e.g: 1B and 3B



picture from evolution-textbook.org (Fig 5-20), redrawn from Eisen, Genome Research, 1998

Gene duplication and loss

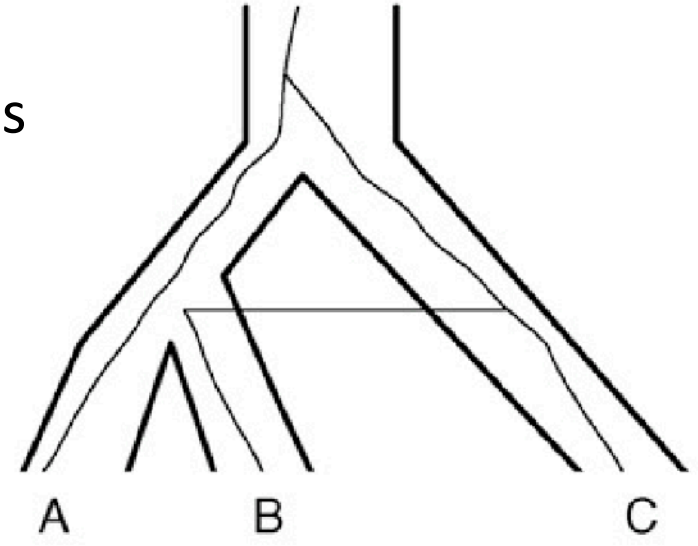
- In some species (e.g. plants) gene duplication and loss is rampant.
- Recall from Colin's talk:
  Paralogs: genes that diverged in a duplication event; e.g: 2A and 3B
  Orthologs: genes that diverged in a speciation event; e.g: 1B and 3B
- A gene tree that includes paralogous genes may differ from the species tree
  - We strive for finding orthologous genes, but we may fail



picture from evolution-textbook.org (Fig 5-20), redrawn from Eisen, Genome Research, 1998

Horizontal Gene Transfer (HGT)

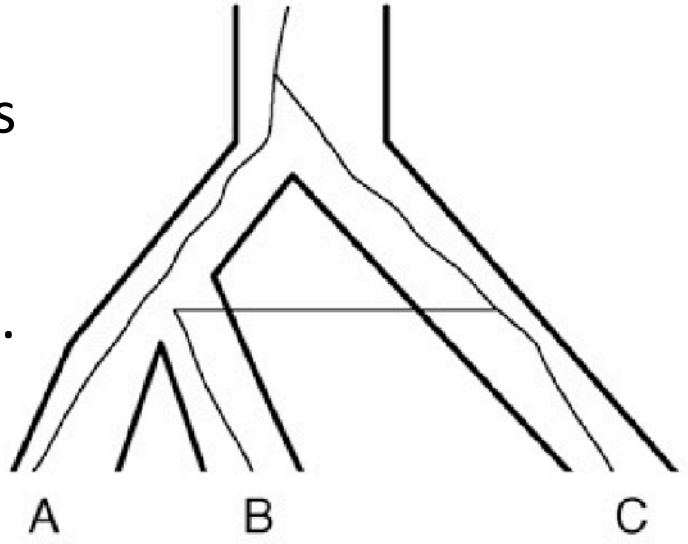
- Horizontal gene transfer: an organism picks up DNA from another organism or the environment rather from its ancestors
  - It may replace a similar gene present in the target or may create a new copy
- Rampant in prokaryotes;
  observed in plants and other eukaryotes



[Degnan & Rosenberg, 2009]

Horizontal Gene Transfer (HGT)

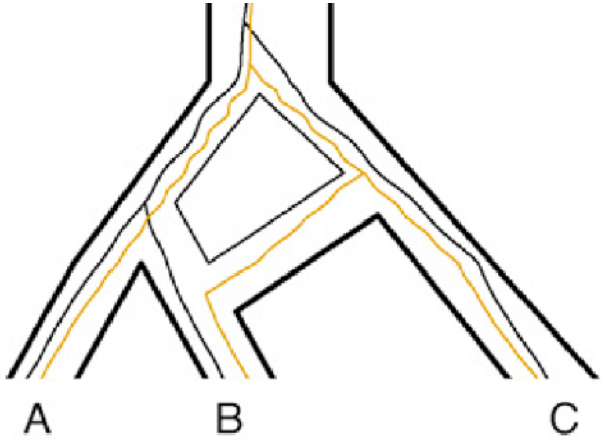
- Horizontal gene transfer: an organism picks up DNA from another organism or the environment rather from its ancestors
  - It may replace a similar gene present in the target or may create a new copy
- Rampant in prokaryotes;
  observed in plants and other eukaryotes
- A tree may be insufficient.
  A phylogenetic network may be needed.



[Degnan & Rosenberg, 2009]

Hybridization

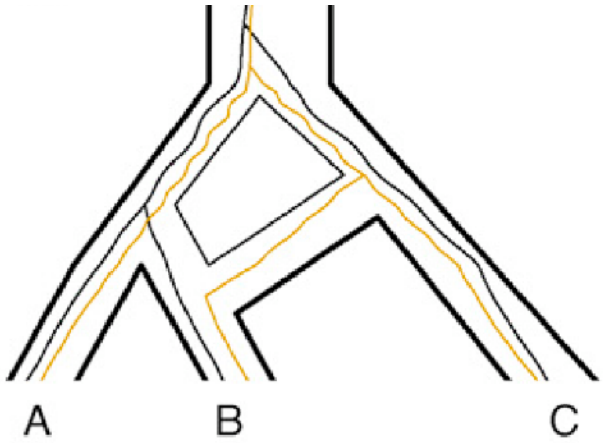
- A viable species is created as a results of hybridization between two different species
  - The contribution of the parent species to the new species need not be equal.
  - A tree is not sufficient; networks needed



[Degnan & Rosenberg, 2009]

Hybridization

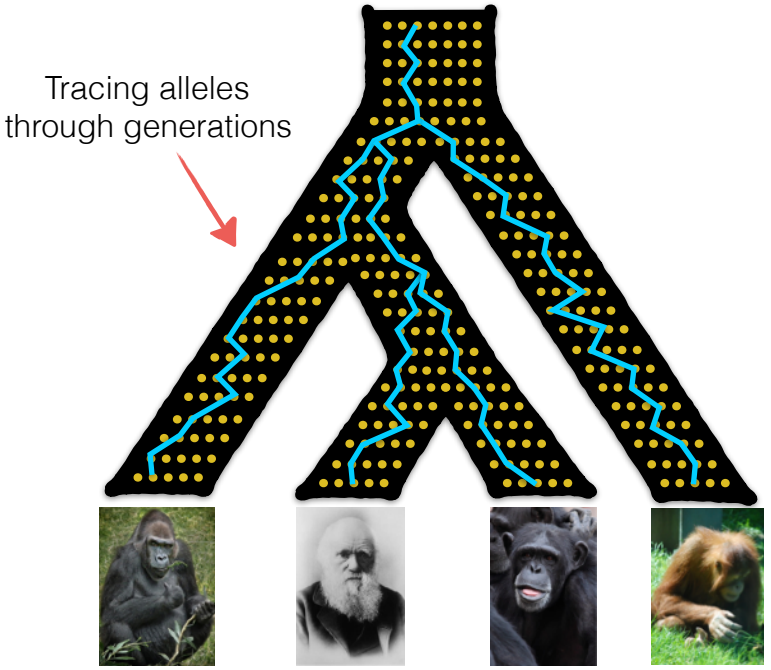
- A viable species is created as a results of hybridization between two different species
  - The contribution of the parent species to the new species need not be equal.
  - A tree is not sufficient; networks needed
- What does a species even mean?
  - 17 different definitions … a matter of great debate
  - Species delineation is an active area of research



[Degnan & Rosenberg, 2009]

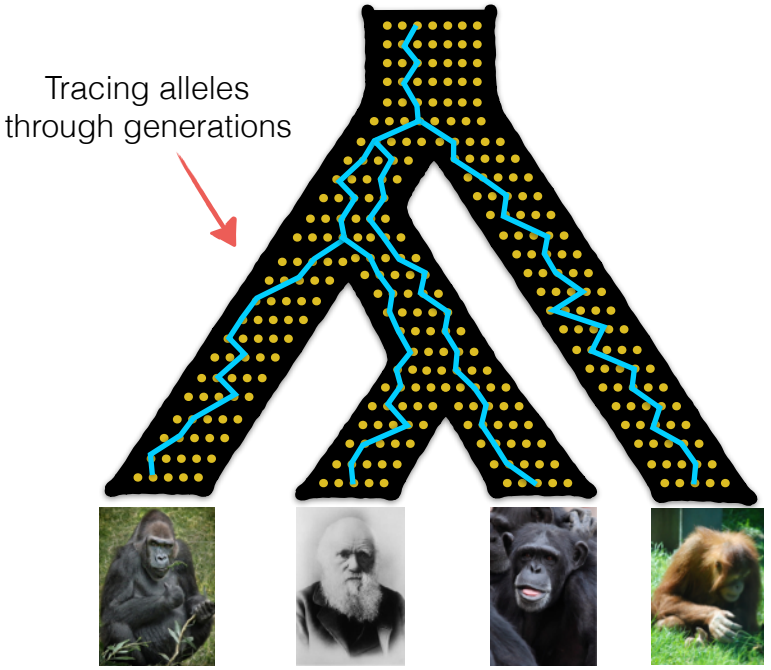
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations



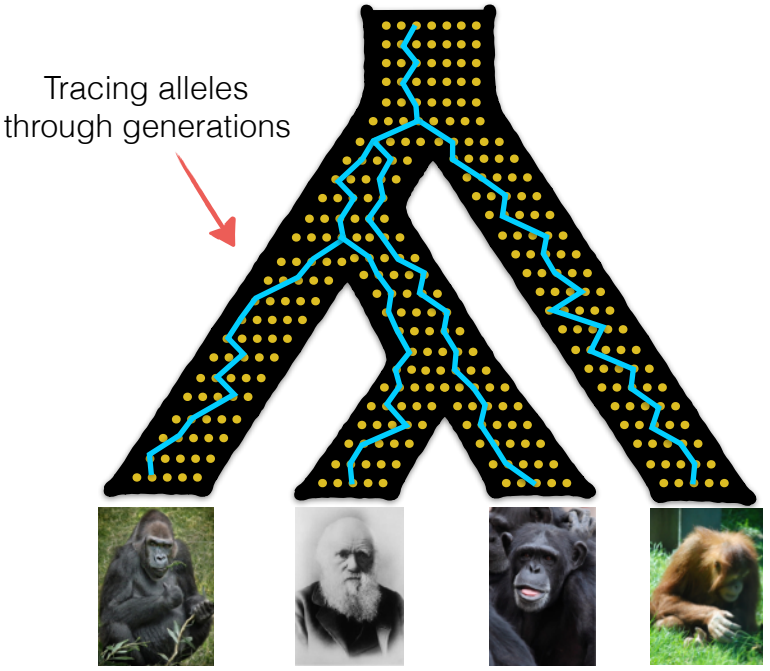
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations



Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations
- Omnipresent:
  - Possible for every gene tree
  - Likely for short branches or large population sizes



Phylogenetic inference despite discordance

So far …

we learned various reasons for discordance.

Gene tree discordance has to be accounted for.

How so?

Four main questions in phylogenomics

- Reconciliation:
  - Map a given (i.e., known) gene tree on to a species tree
  - Explains how a gene tree evolved inside the species tree

Four main questions in phylogenomics

- Reconciliation:
  - Map a given (i.e., known) gene tree on to a species tree
  - Explains how a gene tree evolved inside the species tree
- Infer the species tree given a collection of known (or inferred) gene trees

Four main questions in phylogenomics

- Reconciliation:
  - Map a given (i.e., known) gene tree on to a species tree
  - Explains how a gene tree evolved inside the species tree
- Infer the species tree given a collection of known (or inferred) gene trees
- Infer a gene tree given a known species tree and sequence data for that gene (a.k.a tree fixing)

Four main questions in phylogenomics

- Reconciliation:
  - Map a given (i.e., known) gene tree on to a species tree
  - Explains how a gene tree evolved inside the species tree
- Infer the species tree given a collection of known (or inferred) gene trees
- Infer a gene tree given a known species tree and sequence data for that gene (a.k.a tree fixing)
- Co-estimate gene trees and the species tree given the sequence data for a collection of genes

Four main questions in phylogenomics

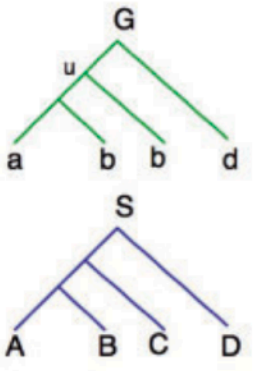
- Reconciliation:
  - Map a given (i.e., known) gene tree on to a species tree
  - Explains how a gene tree evolved inside the species tree
- Infer the species tree given a collection of known (or inferred) gene trees
- Infer a gene tree given a known species tree and sequence data for that gene (a.k.a tree fixing)
- Co-estimate gene trees and the species tree given the sequence data for a collection of genes
- … and others …

Approaches to phylogenomics inference

A. Parsimony-based
B. Model-based
C. Summary-based

Parsimony-based

- Describe discordances between gene trees and the species tree using the minimum number of "events"
  - Events: Duplications, losses, transfers, deep coalescences
  - Relies heavily on fast (linear time) parsimonious reconciliation between gene trees and the species tree



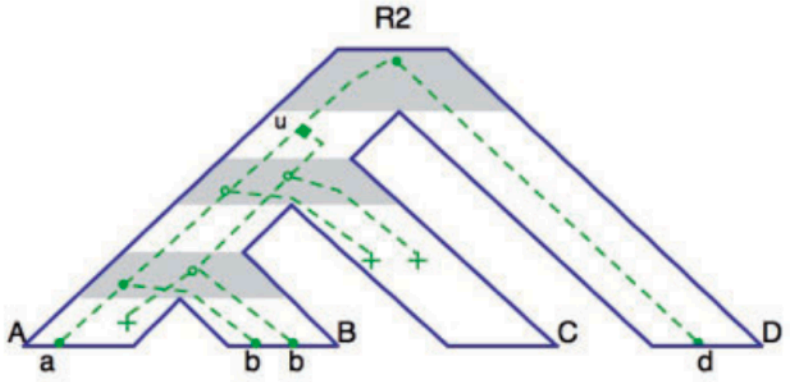


Figure from Doyon et al., Briefings in Bioinformatics, 2011 doi: 10.1093/bib/bbr045

Parsimony-based

- Describe discordances between gene trees and the species tree using the minimum number of "events"
  - Events: Duplications, losses, transfers, deep coalescences
  - Relies heavily on fast (linear time) parsimonious reconciliation between gene trees and the species tree

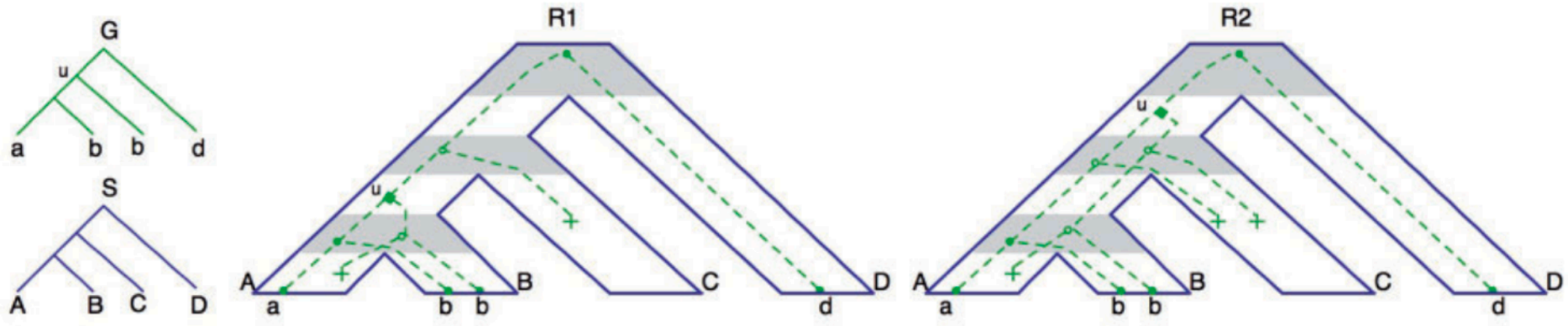
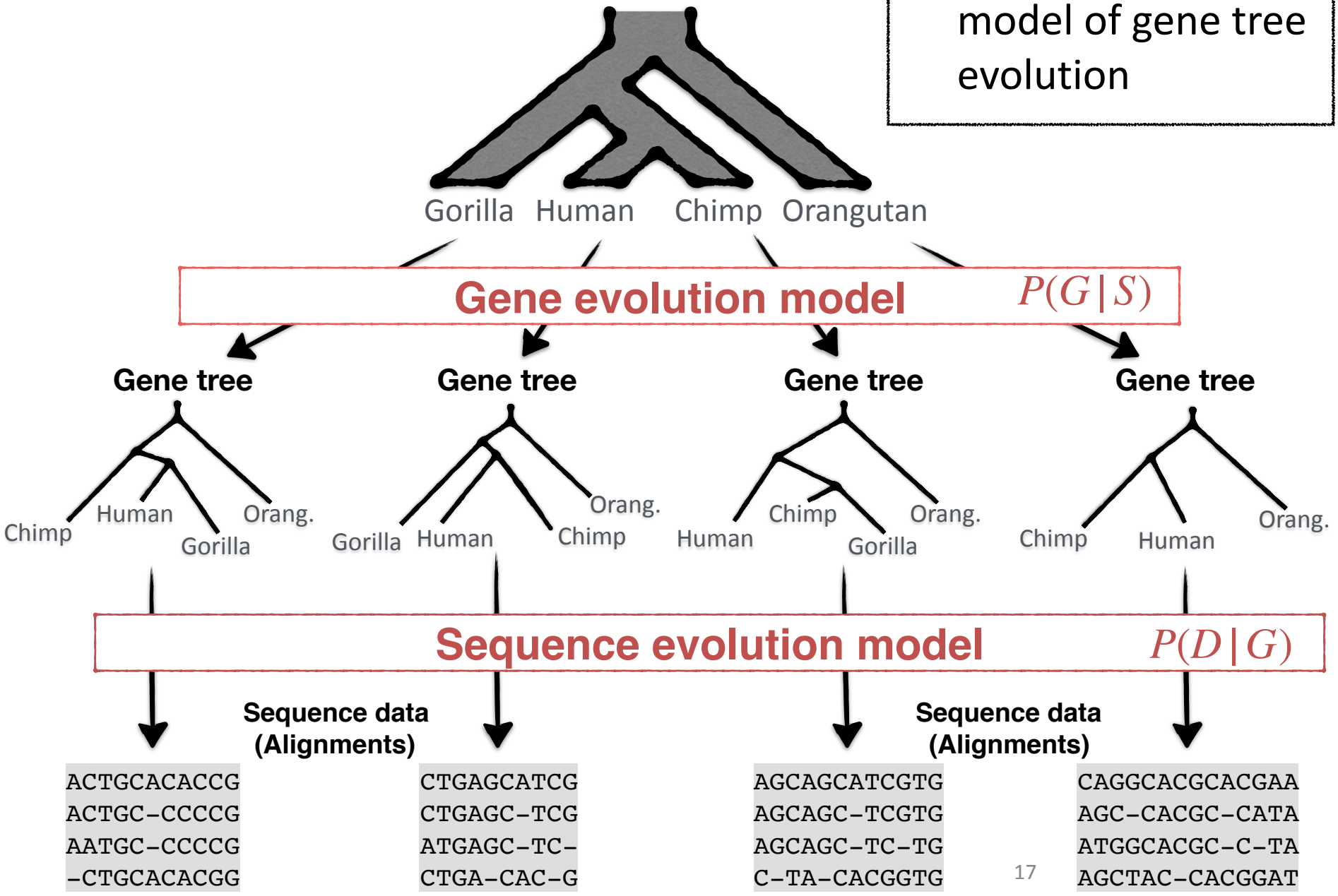


Figure from Doyon et al., Briefings in Bioinformatics, 2011 doi: 10.1093/bib/bbr045

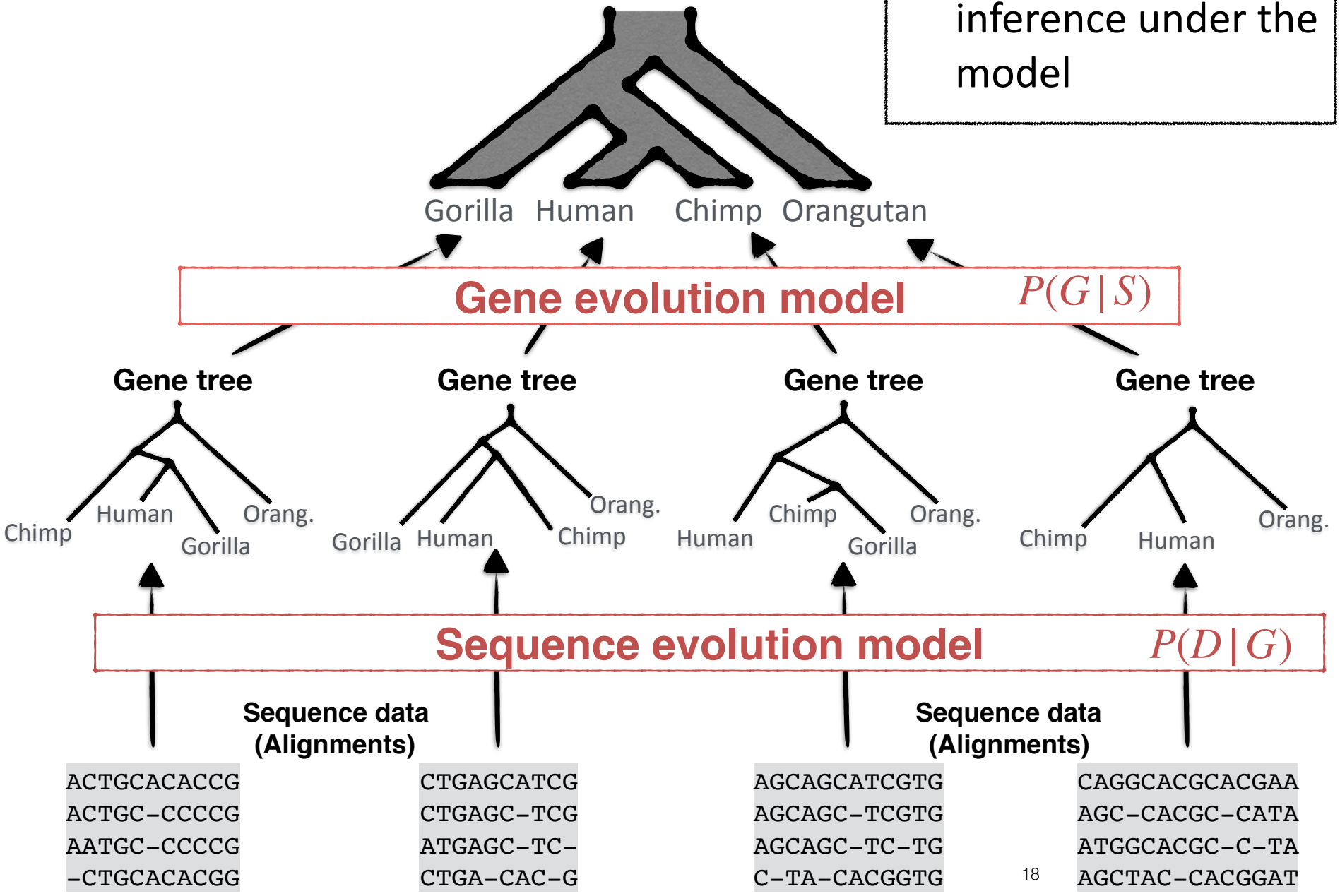
Model-based

A. Design a generative model of gene tree evolution



Model-based

B. ML or Bayesian inference under the model



Models of gene tree evolution

- ILS: modeled by the Multi-Species Coalescent model (MSC): an extension of the Kingman's coalescent to multiple species

Models of gene tree evolution

- ILS: modeled by the Multi-Species Coalescent model (MSC): an extension of the Kingman's coalescent to multiple species
- Duploss: typically modeled using birth death processes, requiring a rate of birth and death (often fixed)

Models of gene tree evolution

- ILS: modeled by the Multi-Species Coalescent model (MSC): an extension of the Kingman's coalescent to multiple species
- Duploss: typically modeled using birth death processes, requiring a rate of birth and death (often fixed)
- HGT, gene flow/hybridization: the species phylogeny is modeled as a network (DAG) and gene trees are stochastically embedded in the network.

Models of gene tree evolution

- ILS: modeled by the Multi-Species Coalescent model (MSC): an extension of the Kingman's coalescent to multiple species
- Duploss: typically modeled using birth death processes, requiring a rate of birth and death (often fixed)
- HGT, gene flow/hybridization: the species phylogeny is modeled as a network (DAG) and gene trees are stochastically embedded in the network.
- Models of combined effects also exist
  - DTLSR, DLCoal, ODT, Hybridization+ILS

Models of gene tree evolution

- ILS: modeled by the Multi-Species Coalescent model (MSC): an extension of the Kingman's coalescent to multiple species
- Duploss: typically modeled using birth death processes, requiring a rate of birth and death (often fixed)
- HGT, gene flow/hybridization: the species phylogeny is modeled as a network (DAG) and gene trees are stochastically embedded in the network.
- Models of combined effects also exist
  - DTLSR, DLCoal, ODT, Hybridization+ILS

- Caution: inference under many of these models is difficult

Summary-based methods

- Use expectations under a statistical model but avoid computing the likelihood
  - Often are statistically consistent under some model of gene tree evolution
  - Often based on summary statistics or distance measures

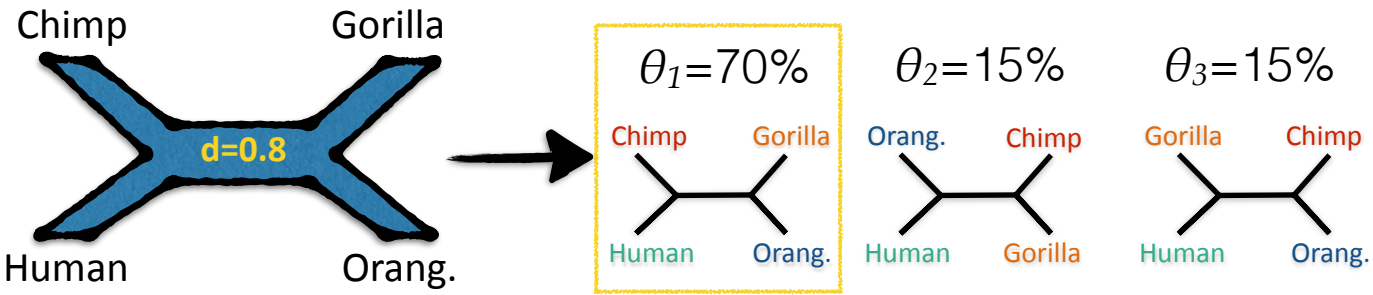
Summary-based methods

- Use expectations under a statistical model but avoid computing the likelihood
  - Often are statistically consistent under some model of gene tree evolution
  - Often based on summary statistics or distance measures
- Usually work in two steps:
  - Gene trees are independently inferred from sequence data
  - Gene trees are combined to build the species tree

- Let’s see an example …

Under MSC model of ILS

For a quartet (4 species), the *unrooted* species tree topology has at least 1/3 probability of appearing in gene trees (Allman, et al. 2010)



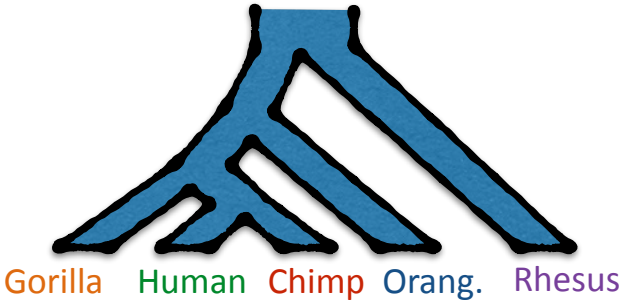
The most frequent gene tree

=

The most likely species tree

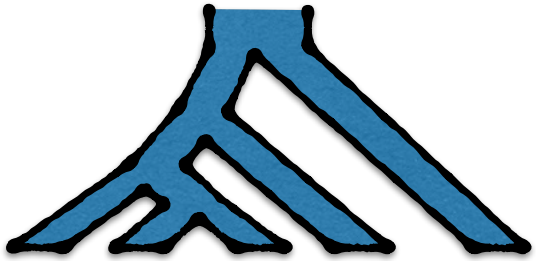
More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)



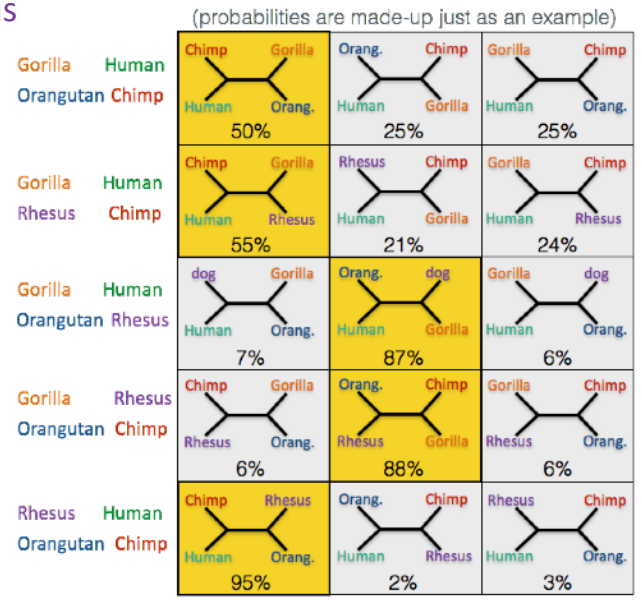
More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)



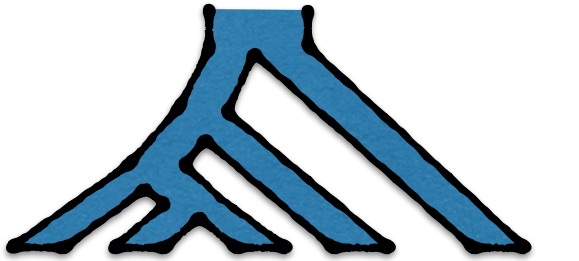
1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the dominant tree for all quartets of taxa
3. Combine quartet trees

   Some tools (e.g.. BUCKy-p [Larget, et al., 2010])



More than 4 species

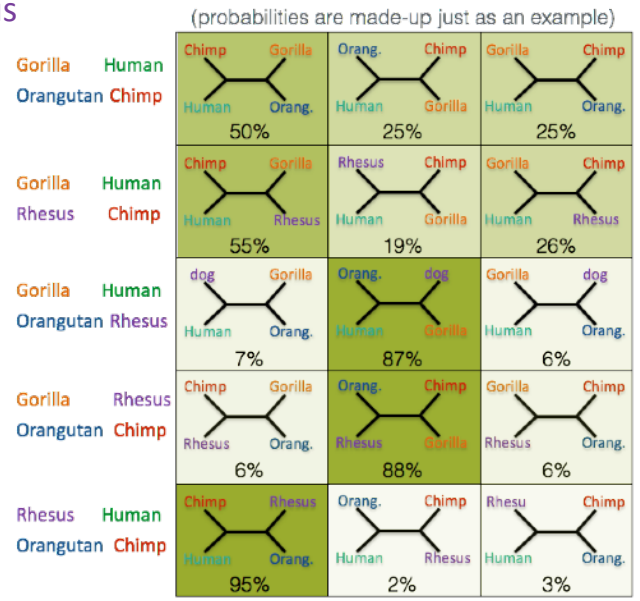
For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)



1. Break gene trees into $\binom{n}{4}$ quartets of species

Alternative:

Weight all $3\binom{n}{4}$ quartet topologies by their frequency and find the optimal tree



Maximum Quartet Support Species Tree

- Optimization problem:

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)|$$

the set of quartet trees induced by T

a gene tree

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

Maximum Quartet Support Species Tree

- Optimization problem:

NP-Hard [Lafond & Scornavaccaori, 2016]

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)|$$

the set of quartet trees induced by T

a gene tree

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

ASTRAL

[Mirarab, et al., Bioinformatics, 2014] [Mirarab and Warnow, Bioinformatics, 2015]

- Solve the Maximum Quartet Support problem exactly using dynamic programming

ASTRAL

[Mirarab, et al., Bioinformatics, 2014] [Mirarab and Warnow, Bioinformatics, 2015]

- Solve the Maximum Quartet Support problem exactly using dynamic programming
- Constrains the search space to make large datasets feasible
  - The constrained version remains statistically consistent
  - Running time of constrained version increases polynomially with the input size

ASTRAL

[Mirarab, et al., Bioinformatics, 2014] [Mirarab and Warnow, Bioinformatics, 2015]

- Solve the Maximum Quartet Support problem exactly using dynamic programming
- Constrains the search space to make large datasets feasible
  - The constrained version remains statistically consistent
  - Running time of constrained version increases polynomially with the input size
- Is adopted widely for incomplete lineage sorting

So far …

Three types of discordance:

A. Duplication and losses
B. HGT & Hybridization
C. ILS

* Each captured by many statistical models

Four Questions:

A. Reconciliation
B. Species tree inference
C. Gene tree inference
D. Co-estimation

Three approaches:

A. Parsimony-based
B. Model-based
C. Summary-based

Many many tools

- Species tree inference
  - ILS: ASTRAL, STAR, MP-EST, GLASS, NJst/ASTRID, DISTIQUE, STELLS, …
  - Duploss: DupTree, iGTP, DynaDup, MulRF, …
  - Hybridization: Phylonet, PhyloNetworks, SNaQ, …
  - Agnostic: Bucky, MRP, MRL, guenomu, …
- Gene tree correction
  - TreeFix, Giga, RefineTree, SPIDIR, SPIMAP, PrIME-GSR, SYNERGY, ALE, ODT (see also EnsemblCompara)
- Reconciliation
  - Notung, DLCoalRecon, PrIME, Phylonet, TreeMap, Korak, CoRe-Pa, Jane, Mowgli, AnGST, …
- Co-estimation
  - PHYLDOG, *BEAST, BEST, BBCA, PhyloNet, …

Many many tools

- Species tree inference
  - ILS: ASTRAL, STAR, MP-EST, GLASS, NJst/ASTRID, DISTIQUE, STELLS, …
  - Duploss: DupTree, iGTP, DynaDup, MulRF, …
  - H
  - A
- Gene
  - T E, ODT
- Reconciliation
  - Notung, DLCoalRecon, PrIME, Phylonet, TreeMap, Korak, CoRe-Pa, Jane, Mowgli, AnGST, …
- Co-estimation
  - PHYLDOG, *BEAST, BEST, BBCA, PhyloNet, …

No scalable species/gene tree inference method can currently address all causes of discordance

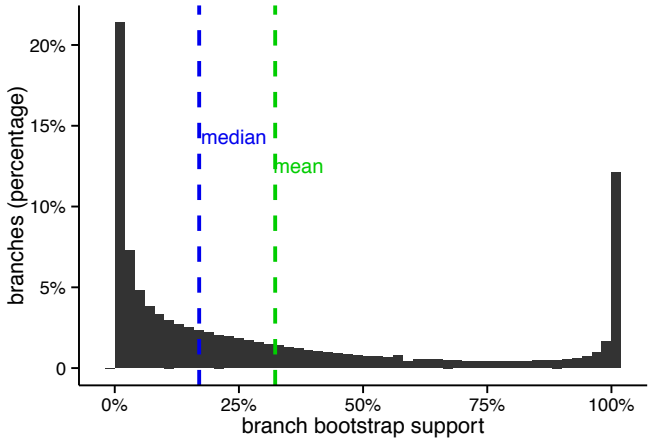
What do you choose?

- What cause(s) of discordance do you believe to be prevalent?

- Do you need to worry about duplication and loss?
  - Or perhaps you have a relatively reliable way of finding orthology? Maybe using whole genome alignments.
- Do you expect hybridization, gene flow, or HGT for your taxa?
- Is ILS likely to be present?
  - short internal branches
  - very high population sizes

What do you choose?

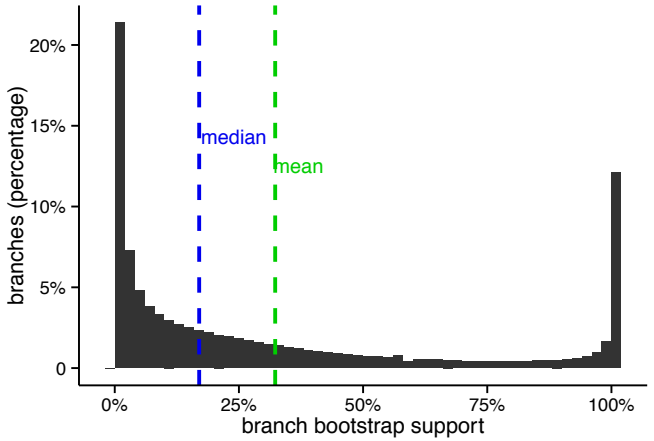
- Statistical noise cannot be ignored



[Jarvis, et al., Science, 2014]

What do you choose?

- Statistical noise cannot be ignored
- Inferring gene trees from short sequences is often error-prone
  - Tree-fixing methods may help
  - Co-estimation is less prone to gene tree estimation error



[Jarvis, et al., Science, 2014]

What do you choose?

- What is the dataset size?

- Methods differ vastly in terms of scalability
  - Fast/scalable: parsimony-based and summary methods (+ ML gene trees)
  - Slower: statistical models, especially when considering multiple causes of discordance
  - Slowest: co-estimation
- Not all approaches are implemented in an optimized manner
  - Some methods are easier to parallelize than others

Take away messages

- Causes of gene tree discordance are varied and can co-exist
- Inference under discordance is an ongoing area of research
  - Dataset size matters!
- Thinking about uncertainty and error is important

Reference

- Maddison, Wayne P. “Gene Trees in Species Trees.” Systematic Biology 46, no. 3 (September 1, 1997): 523–36. https://doi.org/10.2307/2413694.

- Degnan, James H., and Noah A. Rosenberg. “Gene Tree Discordance, Phylogenetic Inference and the Multispecies Coalescent.” Trends in Ecology and Evolution 24, no. 6 (June 1, 2009): 332–40. https://doi.org/10.1016/j.tree.2009.01.009.

- Doyon, J.-P. JP, Vincent Ranwez, Vincent Daubin, and V. Berry. “Models, Algorithms and Programs for Phylogeny Reconciliation.” Briefings in Bioinformatics 12, no. 5 (September 22, 2011): 392–400. https://doi.org/10.1093/bib/bbr045.

- Szöllõsi, G J, E Tannier, Vincent Daubin, and Bastien Boussau. “The Inference of Gene Trees with Species Trees.” Systematic Biology 64, no. 1 (July 28, 2014): e42–62. https://doi.org/10.1093/sysbio/syu048.

- Warnow, Tandy. Computational phylogenetics: An introduction to designing methods for phylogeny estimation. Cambridge University Press, 2017.