Bioinformatics: Network Analysis

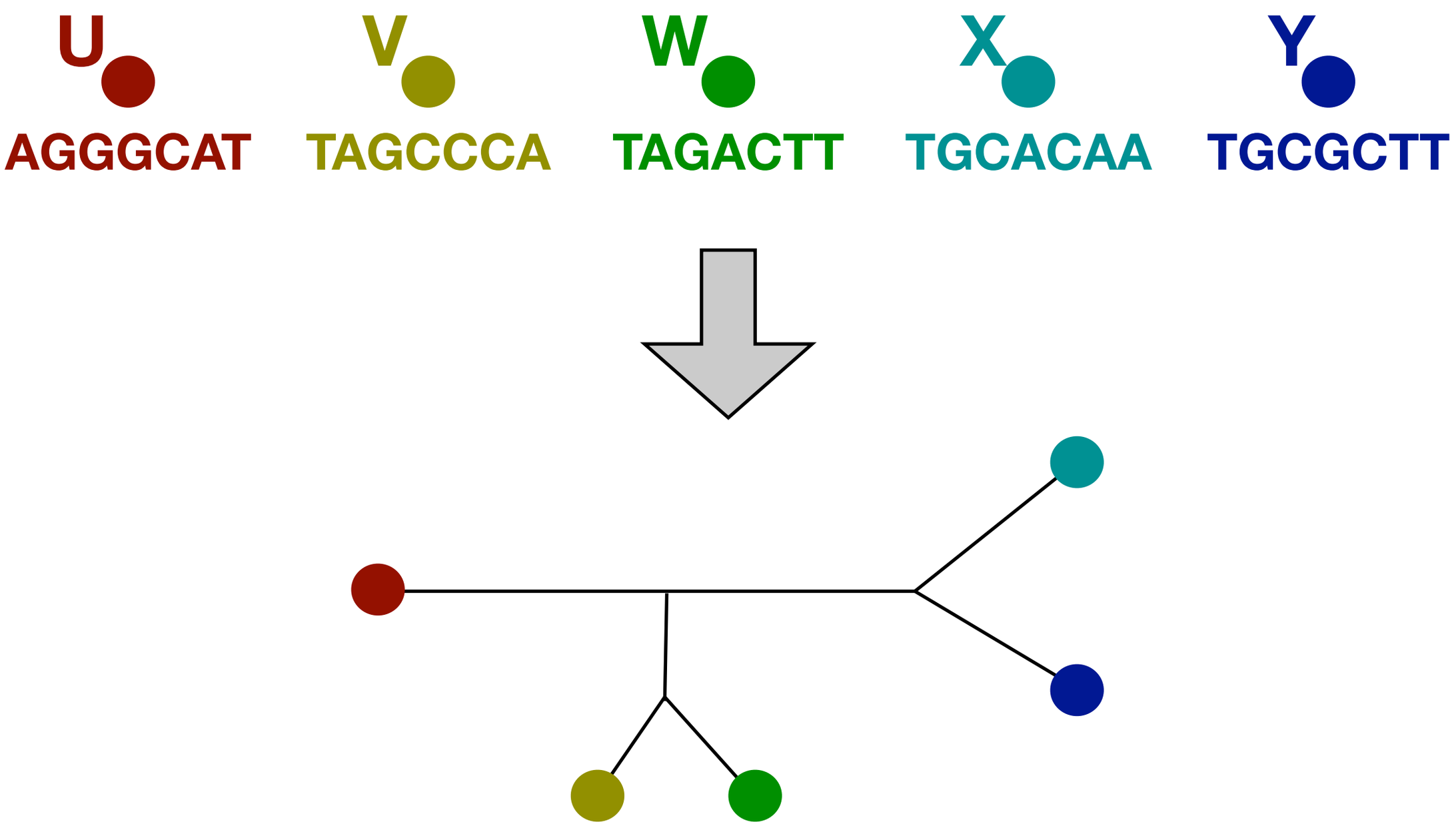
Evolution of Genes and Genomes

COMP 572 (BIOS 572 / BIOE 564) - Fall 2013

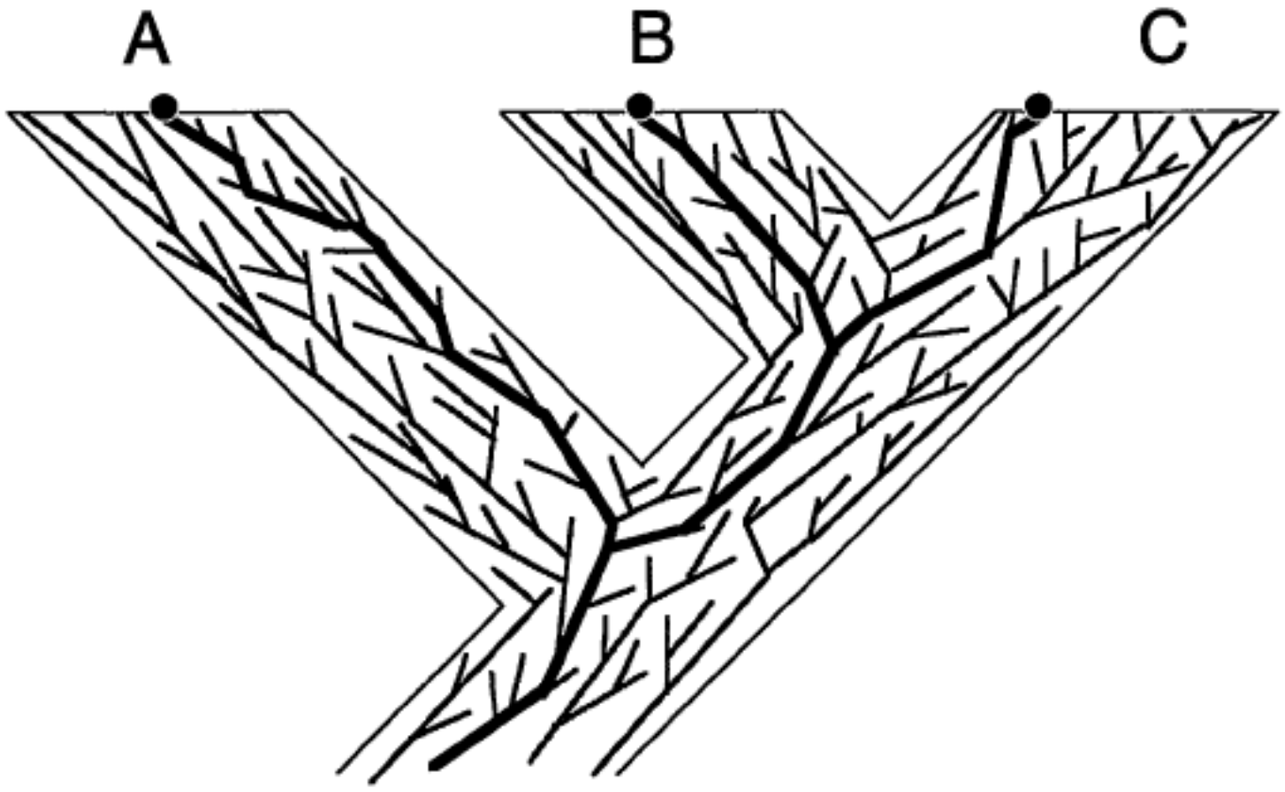
Luay Nakhleh, Rice University

The “Traditional” Phylogeny Reconstruction Problem

The “Traditional” Phylogeny Reconstruction Problem

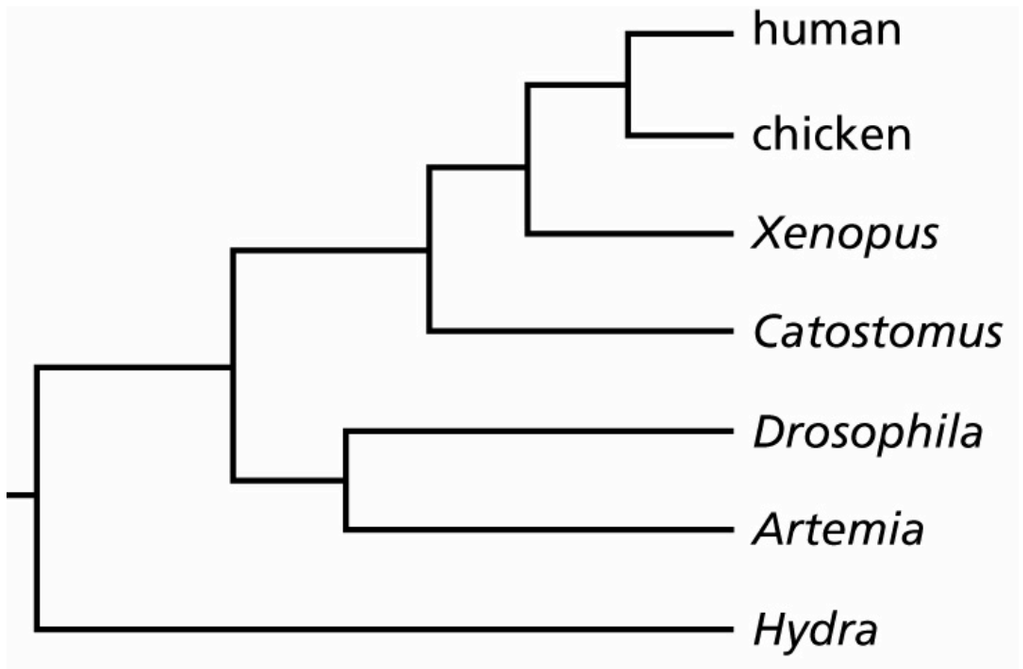


The Evolution of Genes Within the Branches of a Species Tree

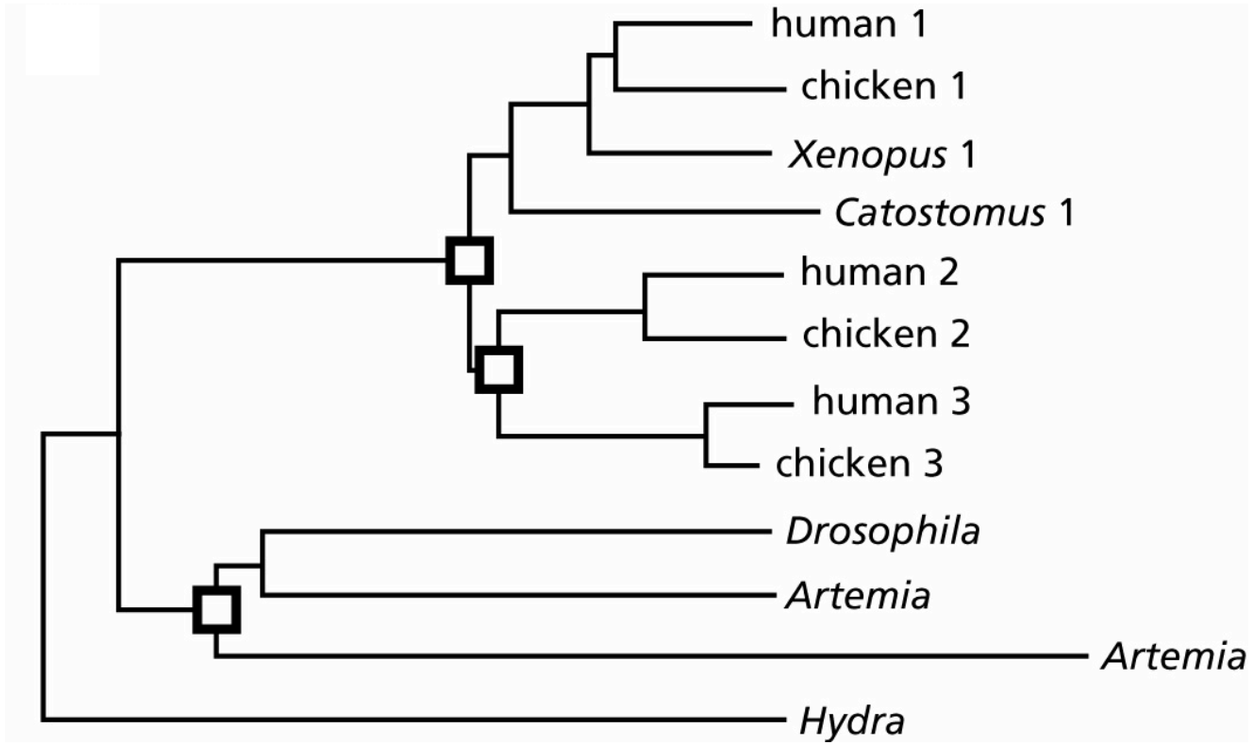


[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

So, What Tree is Being Reconstructed?

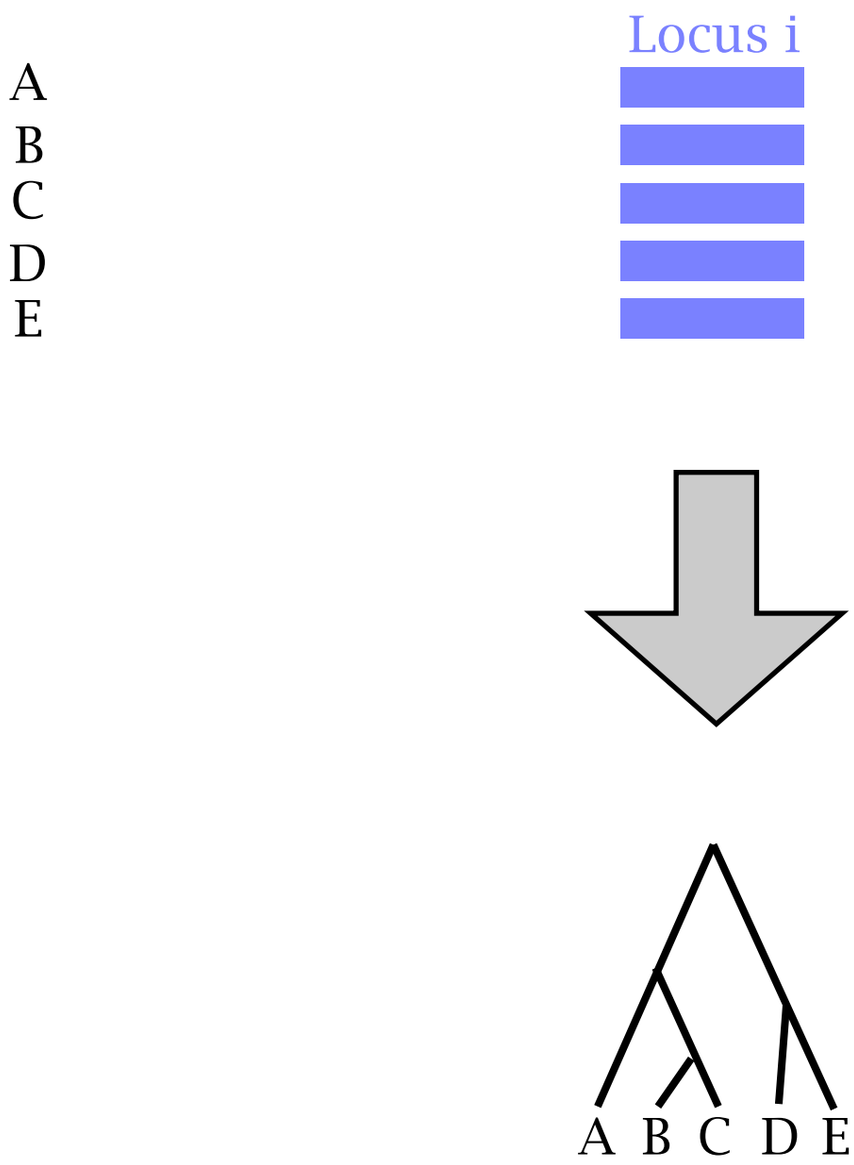


Species tree



Gene tree

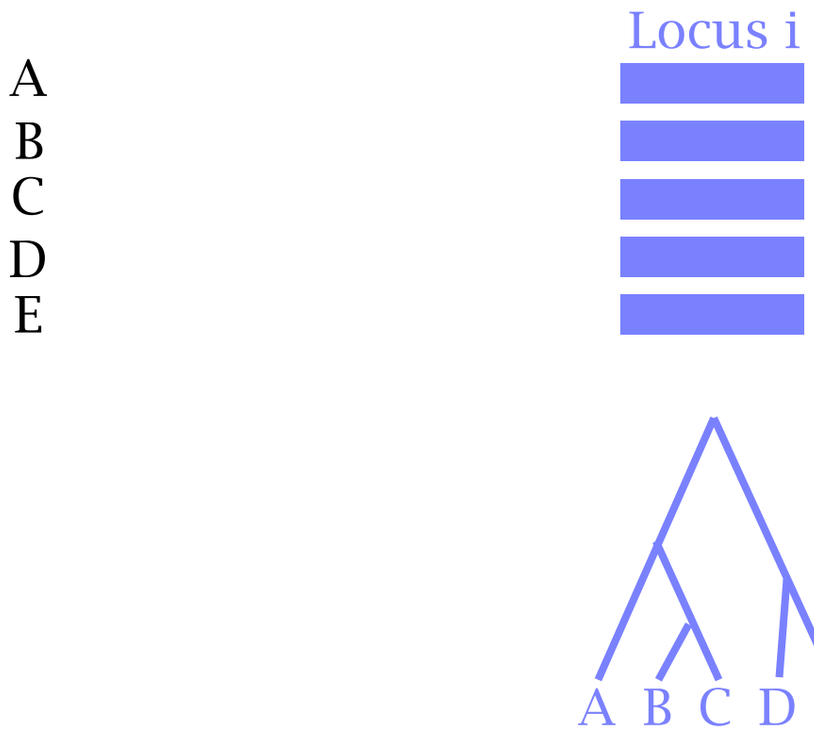
The Pre-Genomic Era



Species
Phylogeny

The Pre-Genomic Era

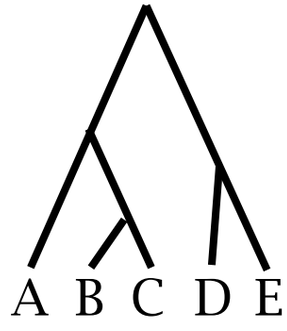
The Pre-Genomic Era



The Pre-Genomic Era

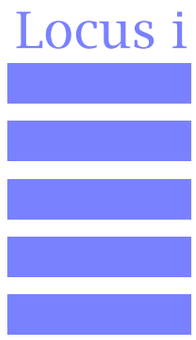
Gene Tree

Species Phylogeny



The Pre-Genomic Era

A
B
C
D
E



The "traditional" phylogeny reconstruction problem

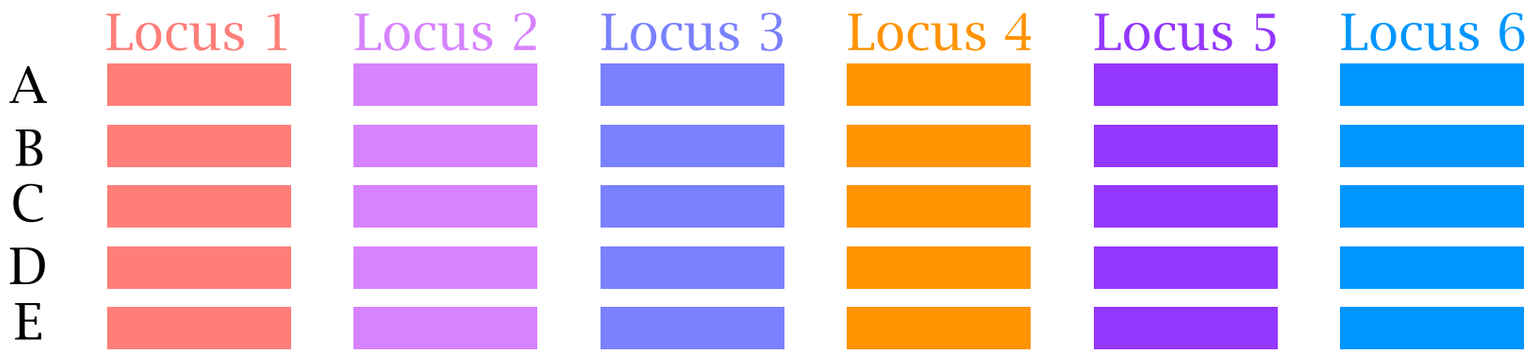
Gene Tree

Species Phylogeny

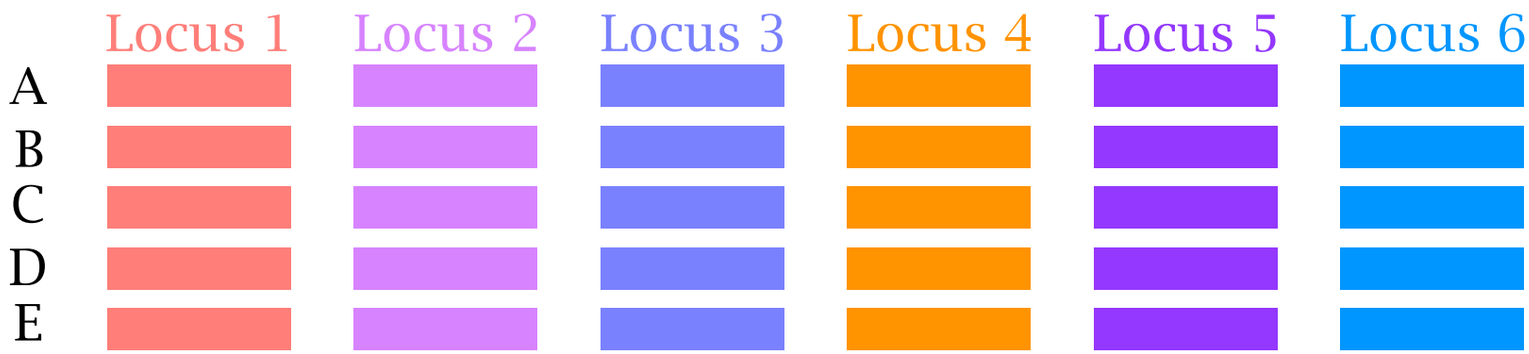
The Post-genomic Era

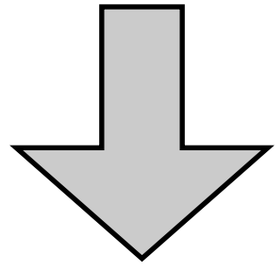
A
B
C
D
E

The Post-genomic Era

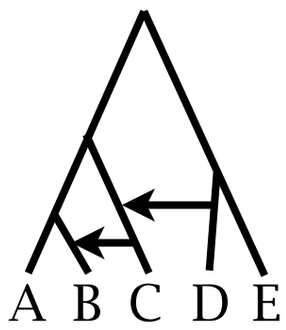


The Post-genomic Era

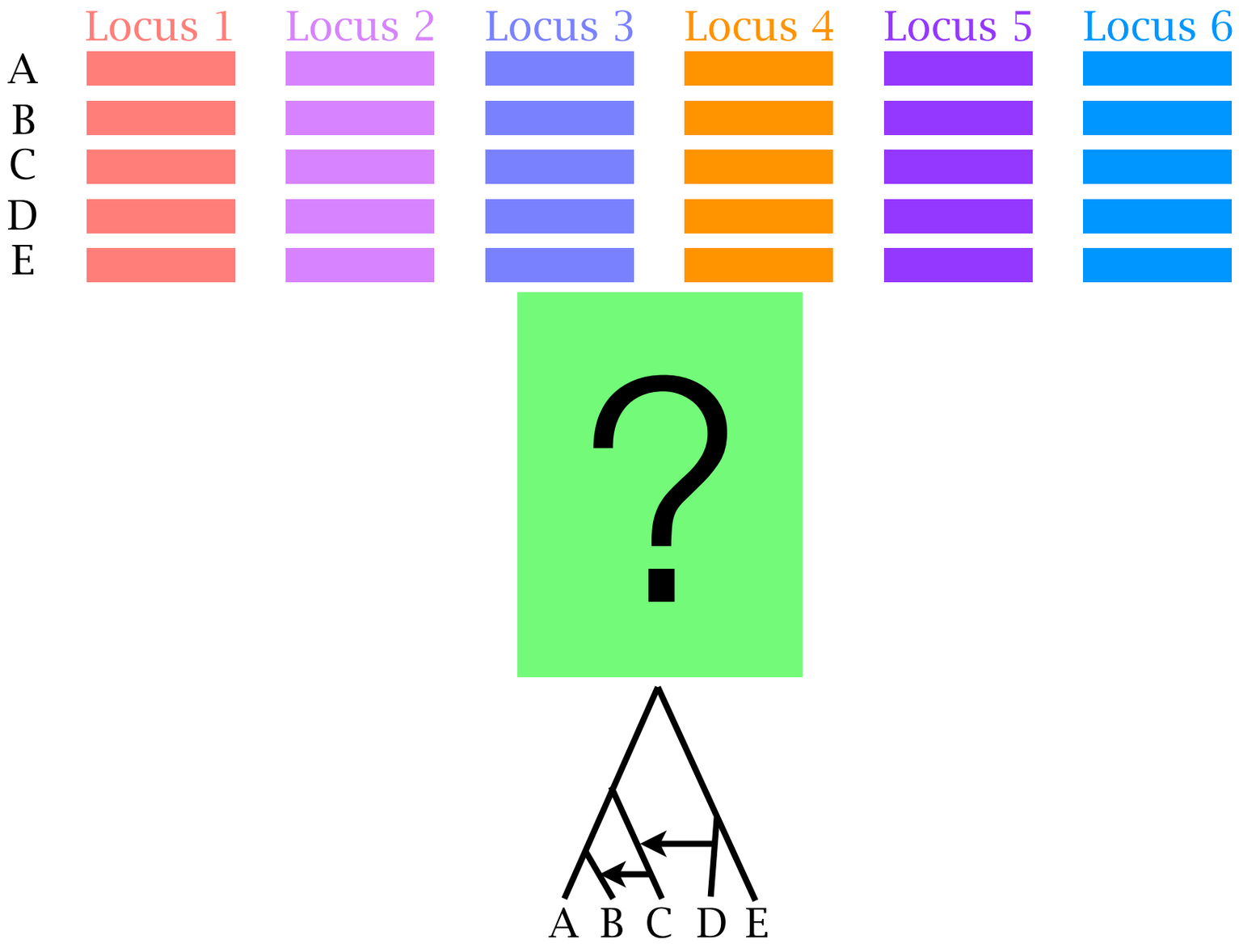




Species
Phylogeny

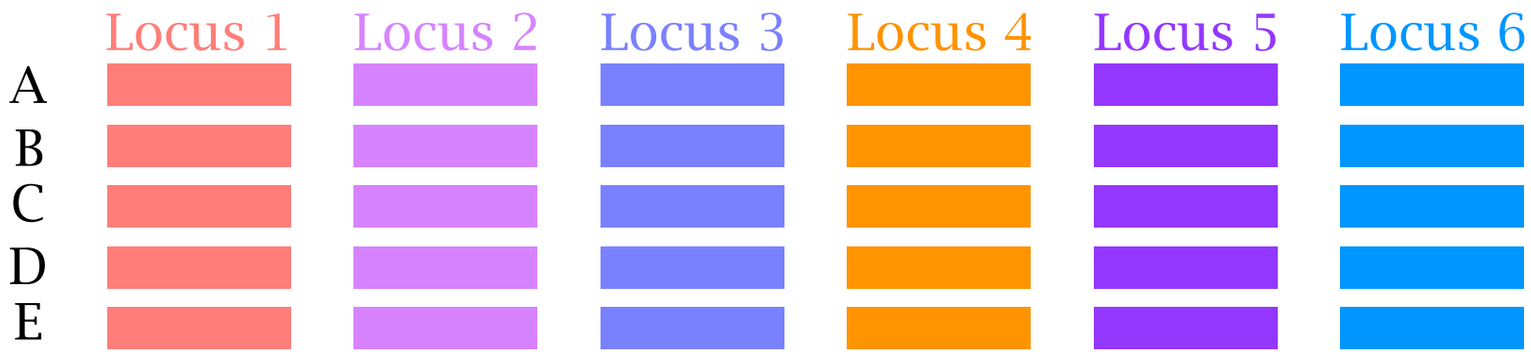


The Post-genomic Era

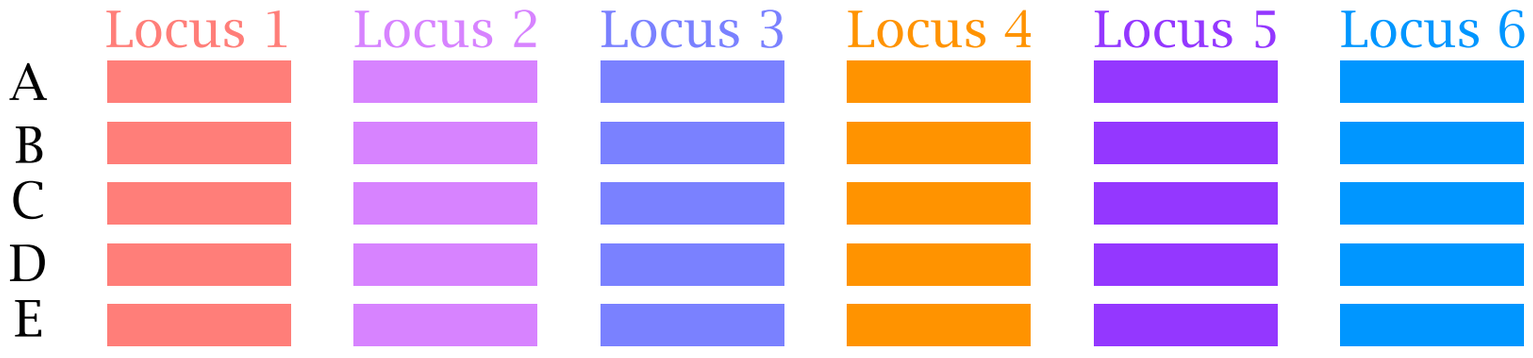


Species
Phylogeny

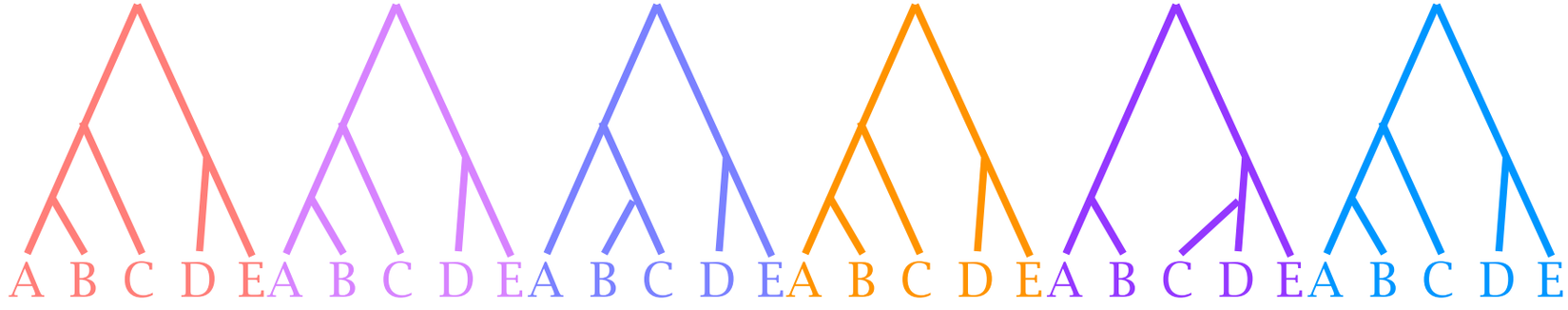
The Post-genomic Era:
I. Gene Tree Incongruence



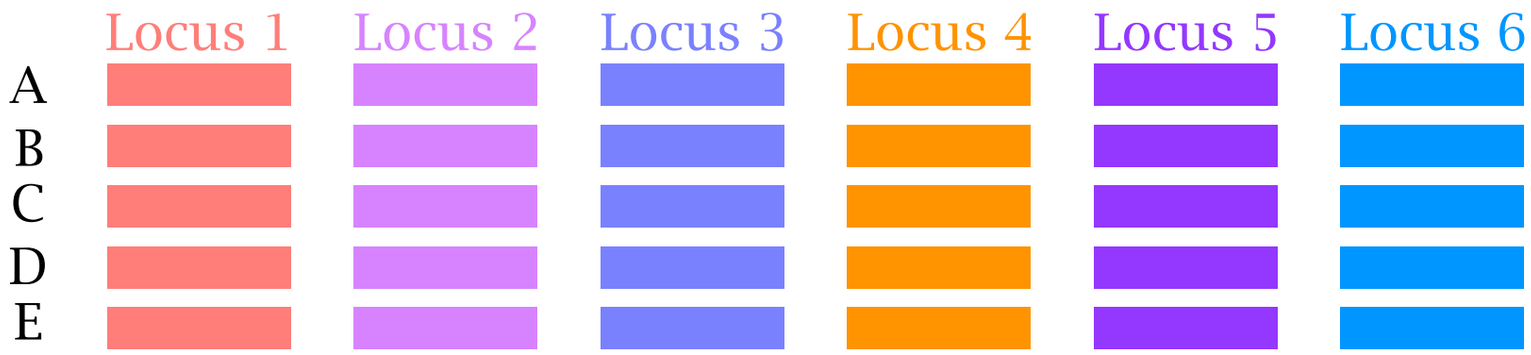
The Post-genomic Era:
I. Gene Tree Incongruence



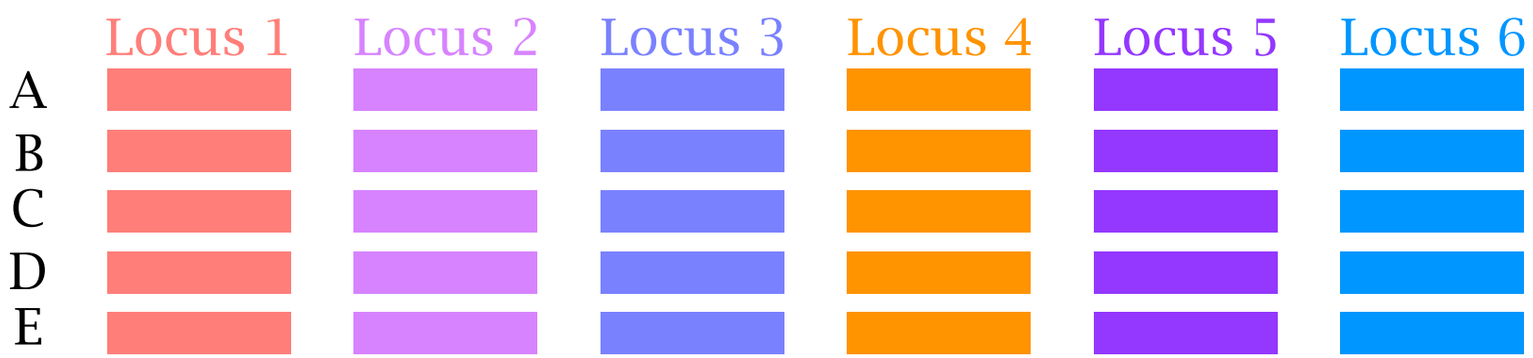
Gene Trees



The Post-genomic Era:
II. Genome Rearrangements



The Post-genomic Era:
II. Genome Rearrangements



The Genomic Context

The Post-genomic Era: Incongruence and Rearrangements

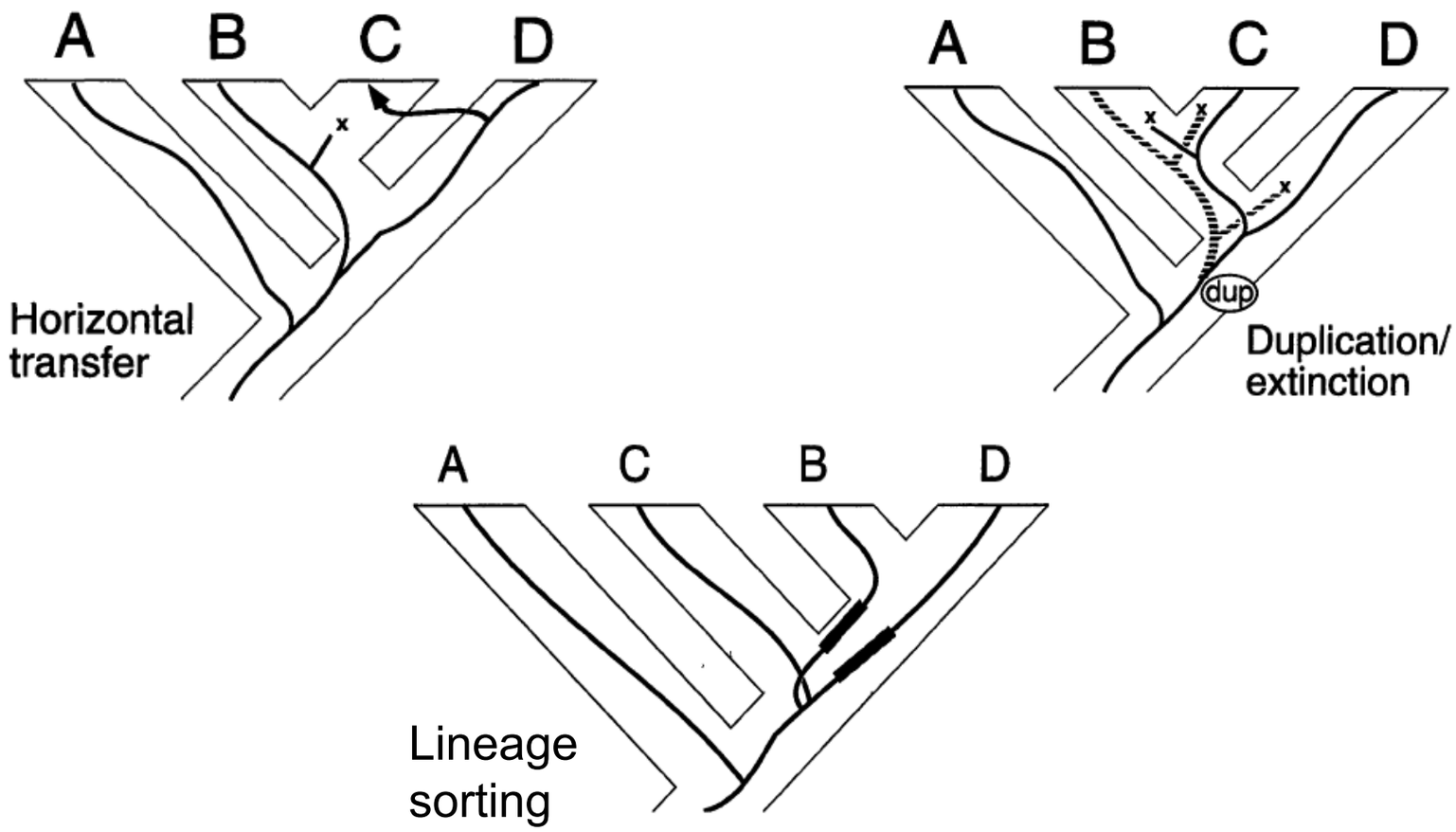
- Gene tree incongruence and genome rearrangements pose challenges and opportunities:
  - **Challenges:** how to model the events, how to infer the events, how to infer species phylogeny while accounting for these events, ...
  - **Opportunities:** resolve very shallow and very deep evolutionary relationships, inform about gene function, understand genomic structural variations and their role in disease (e.g., cancer), ...

Outline of the Rest of this Tutorial

- Gene tree incongruence
    - Biological causes
    - General mathematical frameworks
- Genome rearrangement
    - Rearrangement events
    - General mathematical frameworks

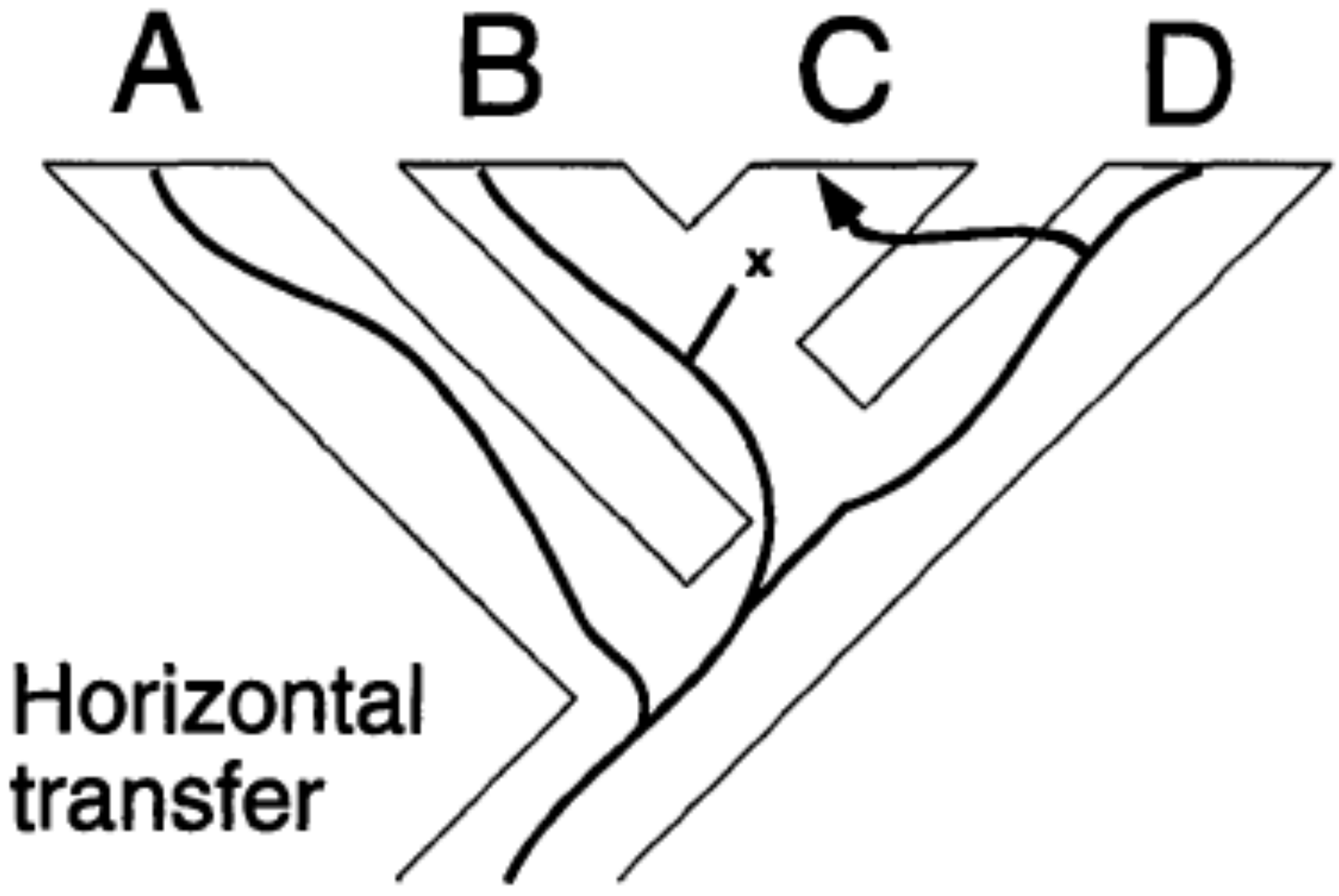
Gene Tree Incongruence

Three Main Biological Events



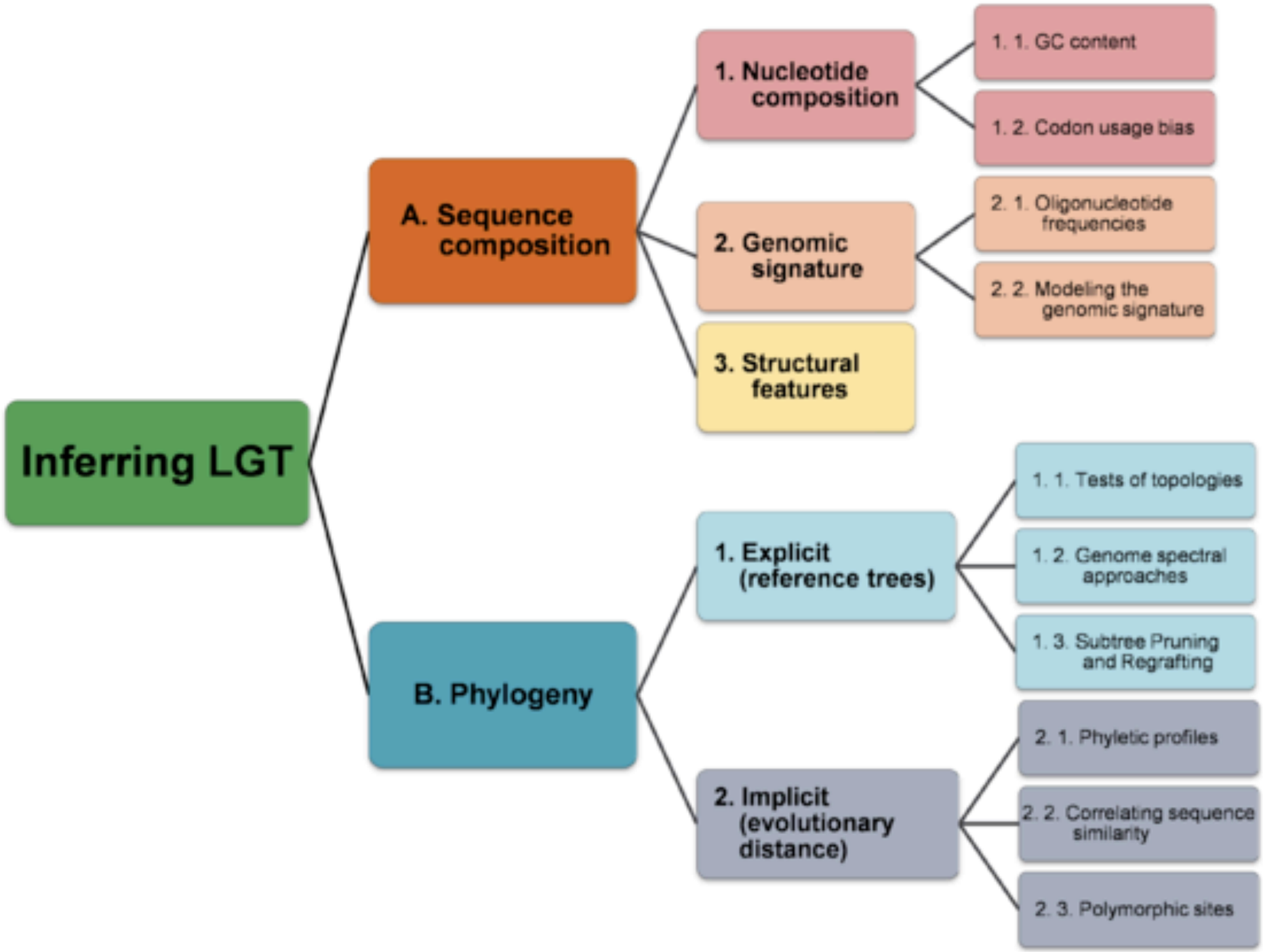
[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

Horizontal (or, Lateral) Gene Transfer (HGT/LGT)



[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

Detecting HGT



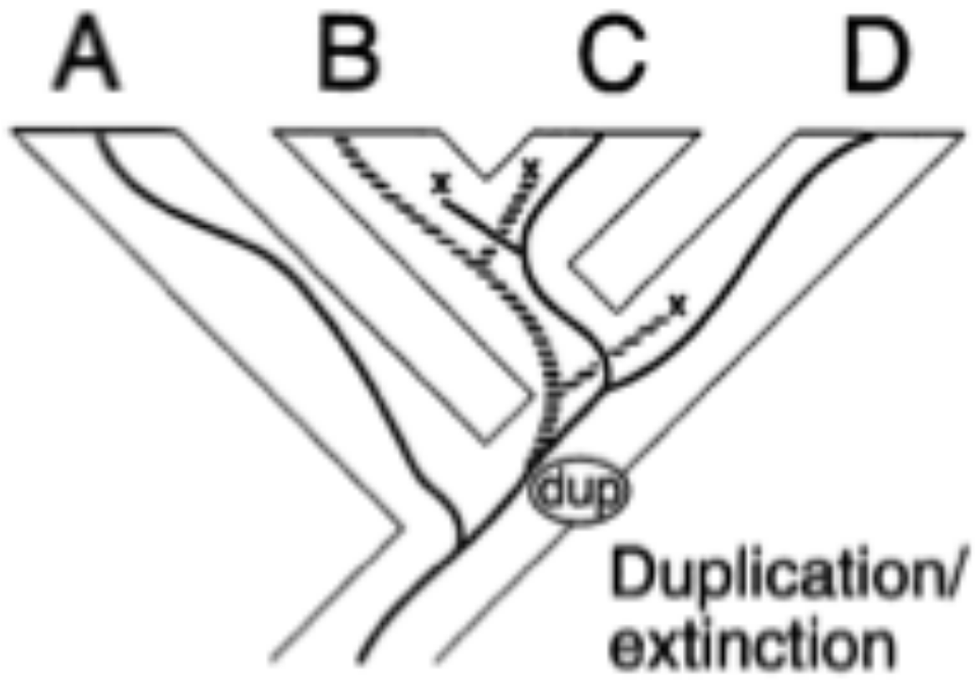
[Source: http://topicpages.ploscompbiol.org/wiki/Detection_of_horizontal_gene_transfer]

Detecting HGT

- The explicit phylogeny-based approach for detecting HGT mostly seeks the minimum number of tree transformation operations (often, the “subtree prune and regraft” operation) that reconciles a gene tree with a species tree.

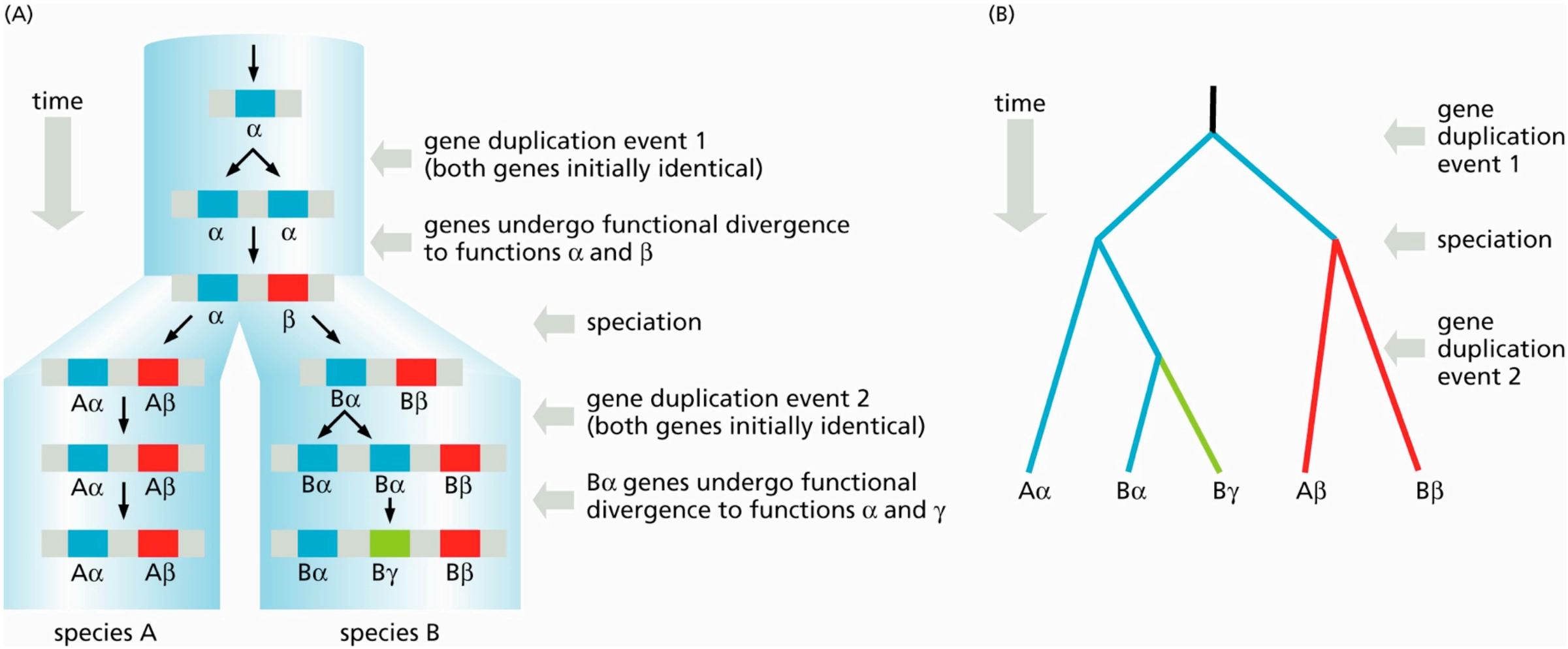
- This number is taken as a lower bound on the number of HGT events required to explain the evolutionary history of the gene under study.

Gene Duplication and Loss



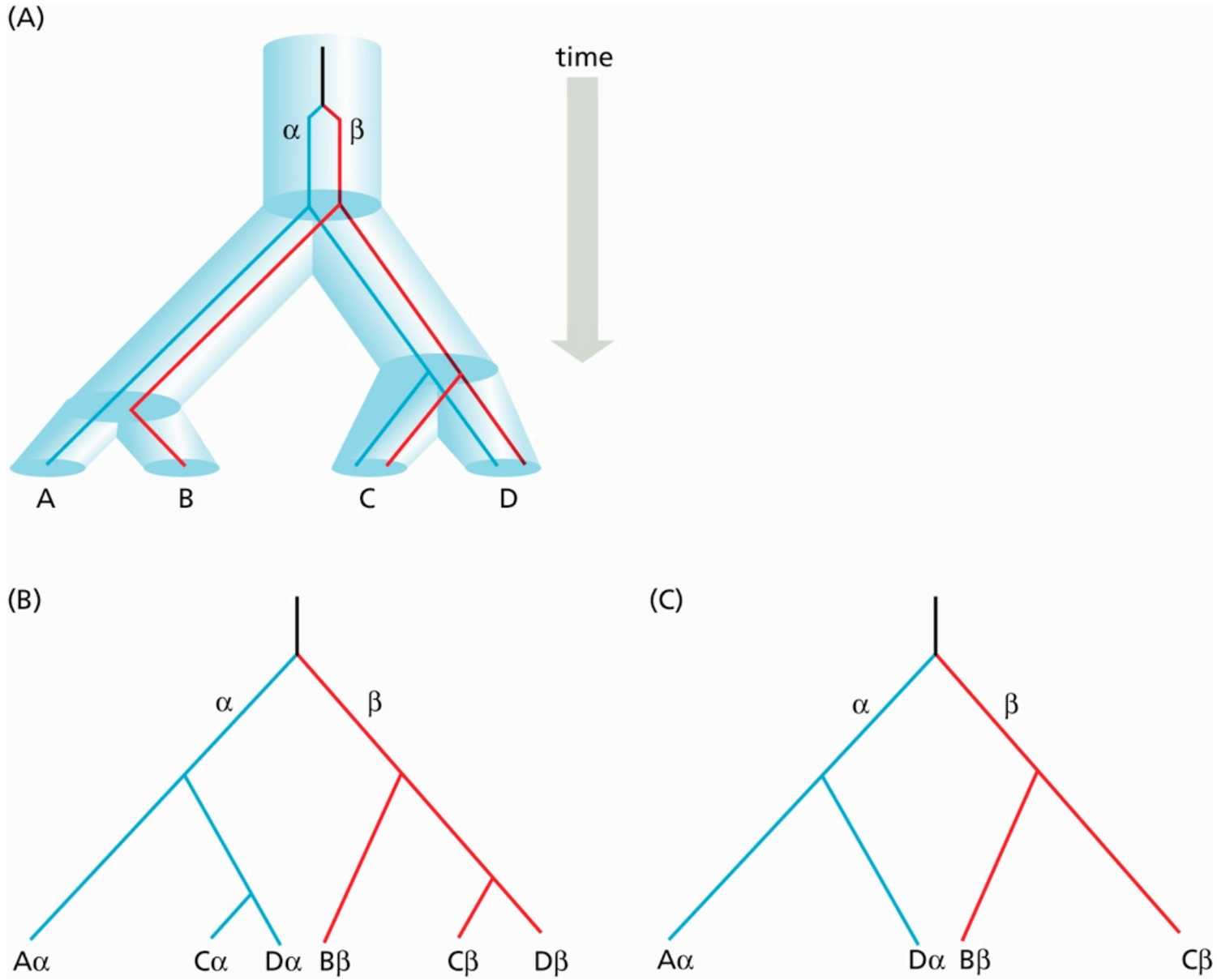
[Source: W.P. Maddison, Syst. Biol. 46(3):523-536,1997.]

Gene Duplication and Loss



[Source: Understanding Bioinformatics]

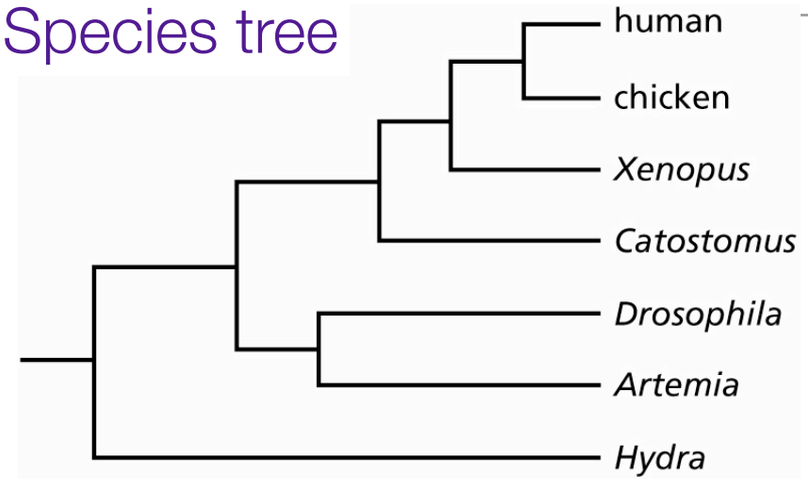
Gene Duplication and Loss



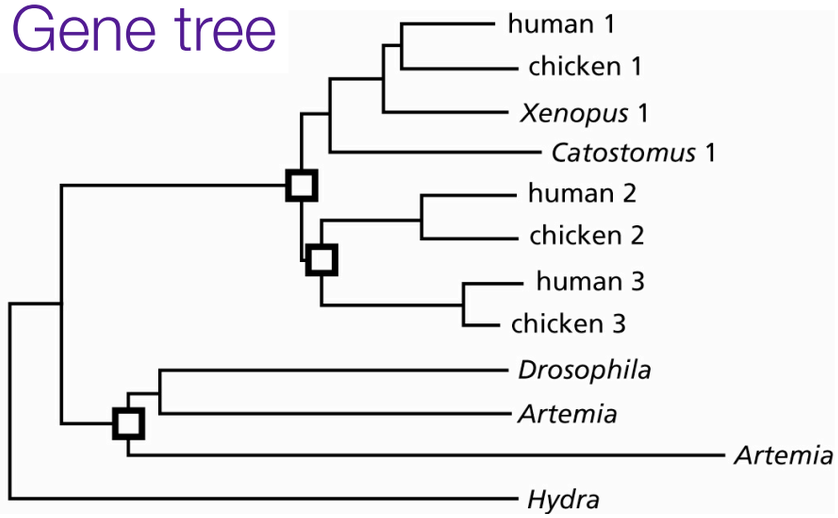
[Source: Understanding Bioinformatics]

Gene Duplication and Loss

Species tree

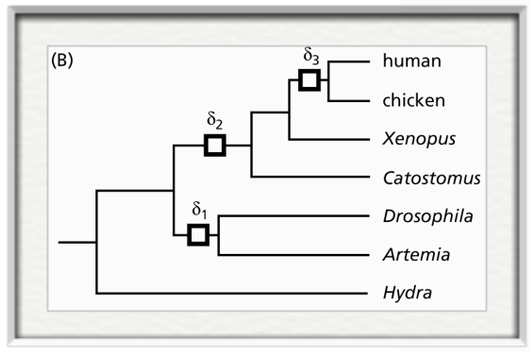


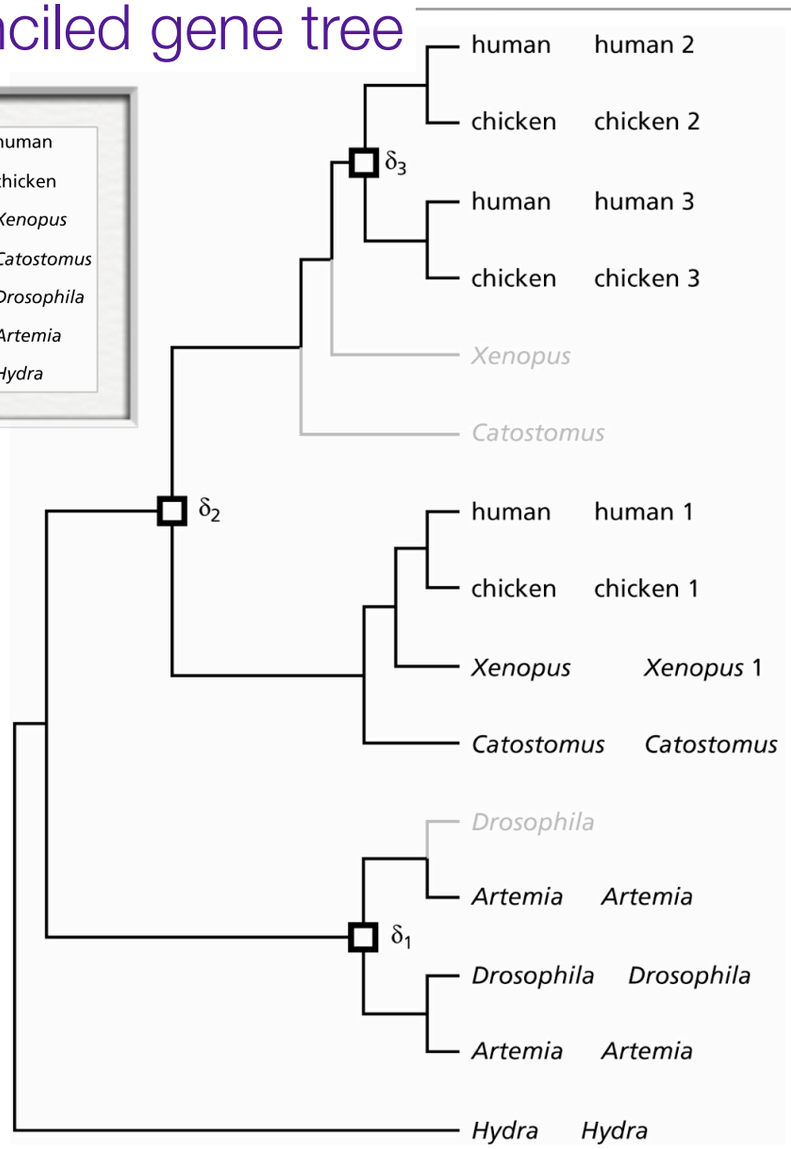
Gene tree



Reconcile

Reconciled gene tree





[Source: Understanding Bioinformatics]

Gene Duplication and Loss

- The parsimony approach to the reconciliation problem seeks the minimum number of duplications and losses (or a weighted sum thereof) to explain the incongruence between the gene tree and species tree.

  - Beginning with Goodman et al., 1979

- Probabilistic models of gene duplication/loss are now emerging, allowing for probabilistic reconciliations.

The Gene Evolution Model and Computing Its Associated Probabilities

LARS ARVESTAD AND JENS LAGERGREN

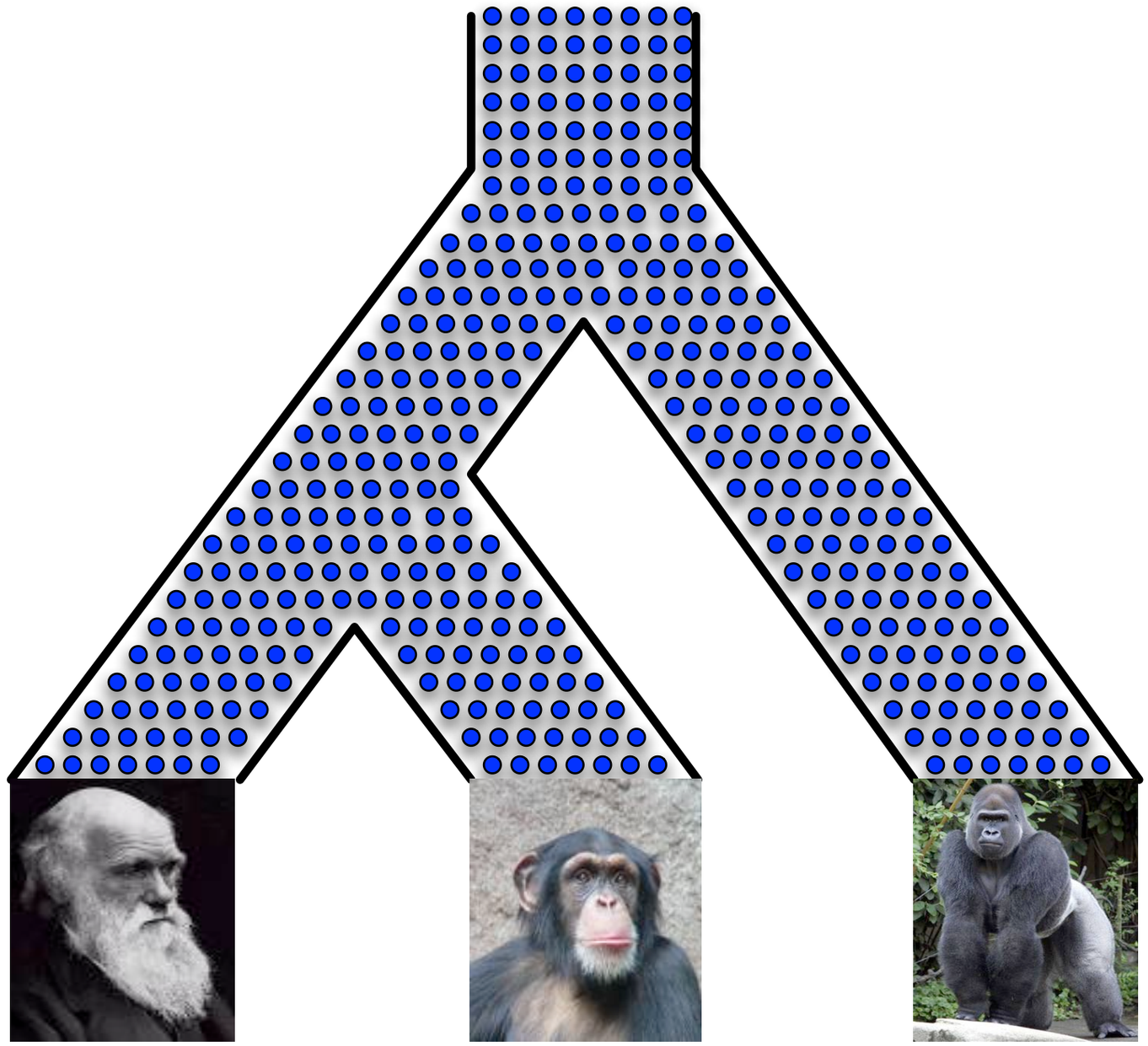
Royal Institute of Technology and Stockholm Bioinformatics Center, Stockholm, Sweden

AND

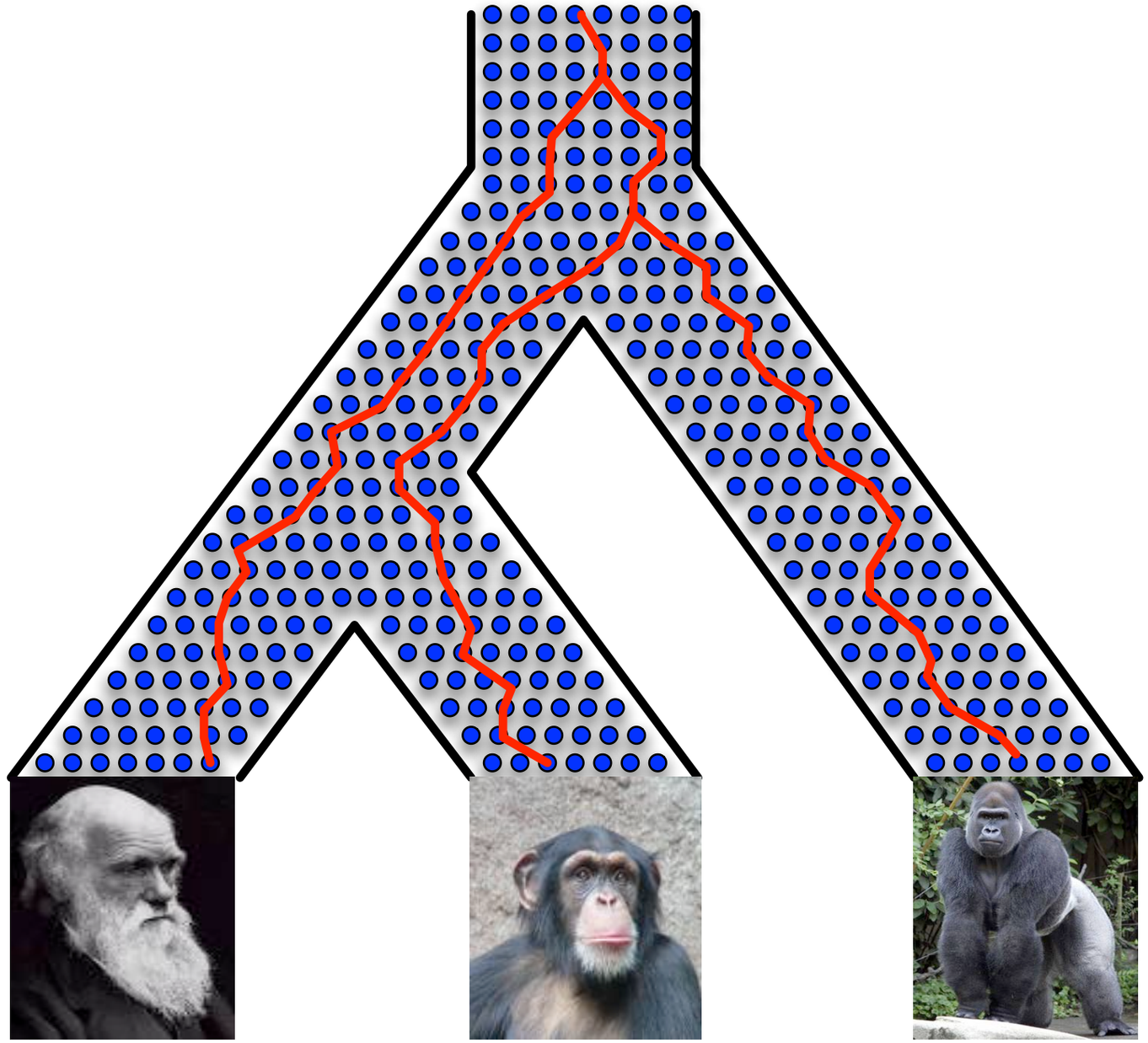
BENGT SENNBLAD

Stockholm University and Stockholm Bioinformatics Center, Stockholm, Sweden

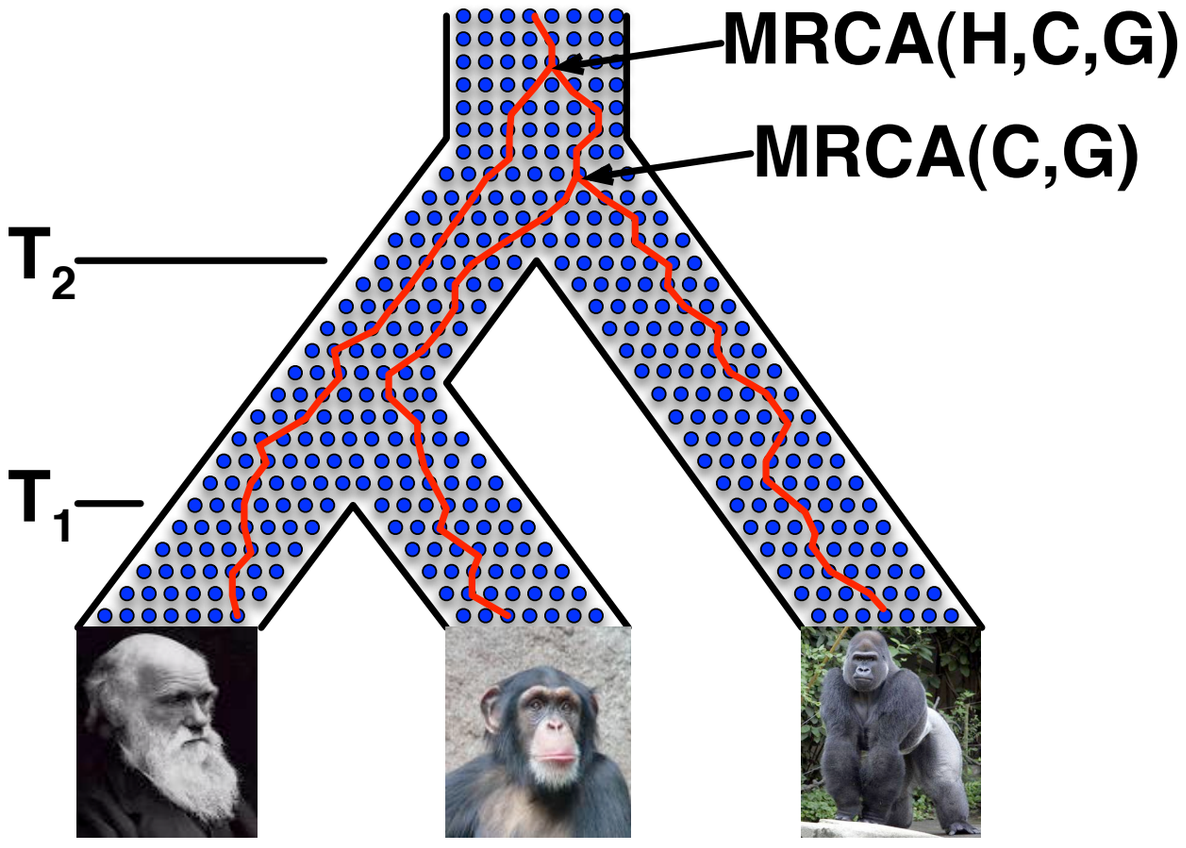
Incomplete Lineage Sorting (ILS)



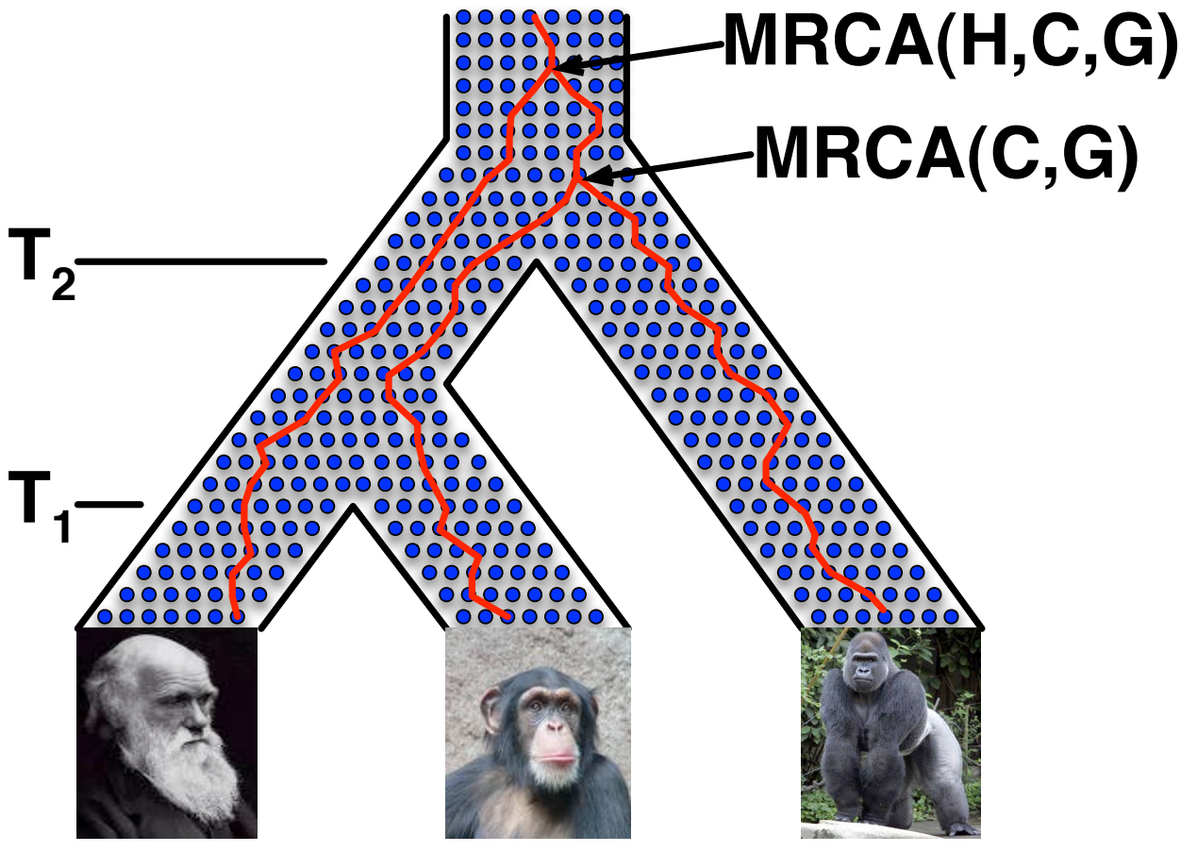
Incomplete Lineage Sorting (ILS)



Incomplete Lineage Sorting (ILS)



Incomplete Lineage Sorting (ILS)

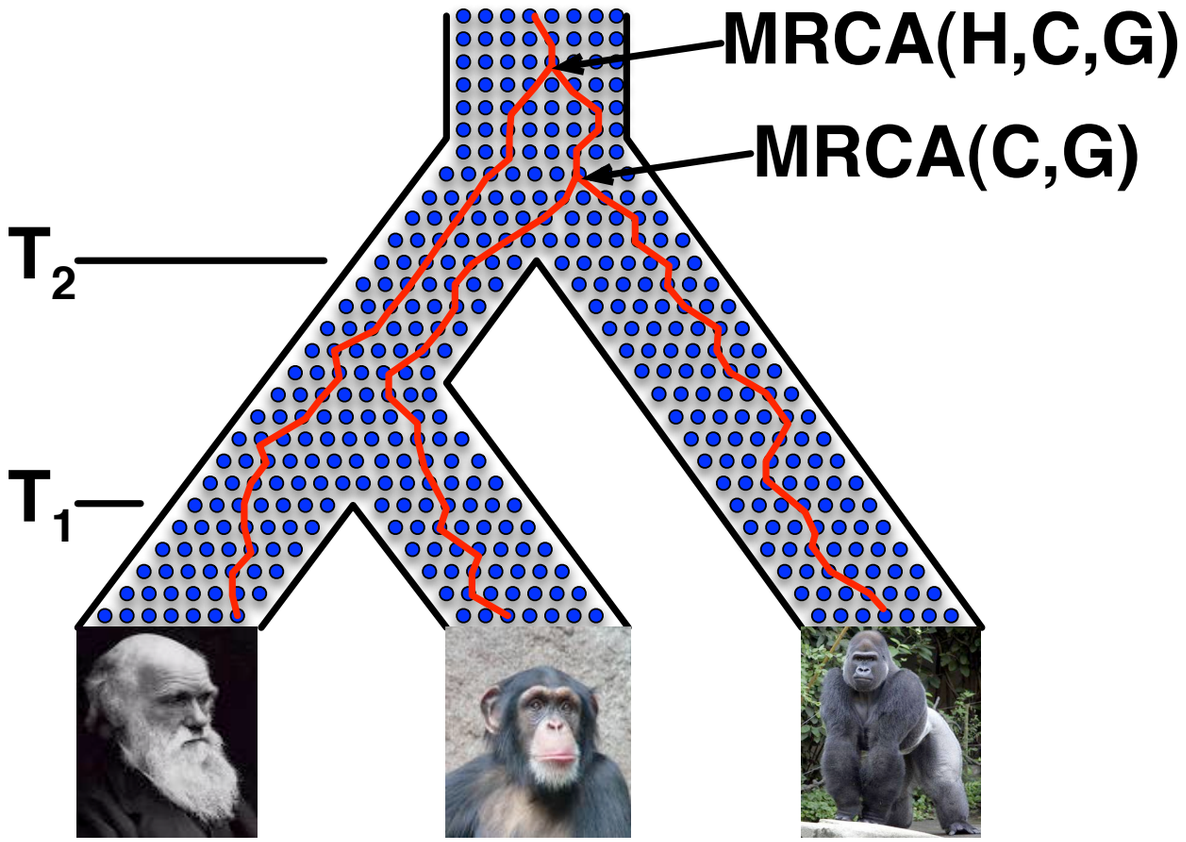


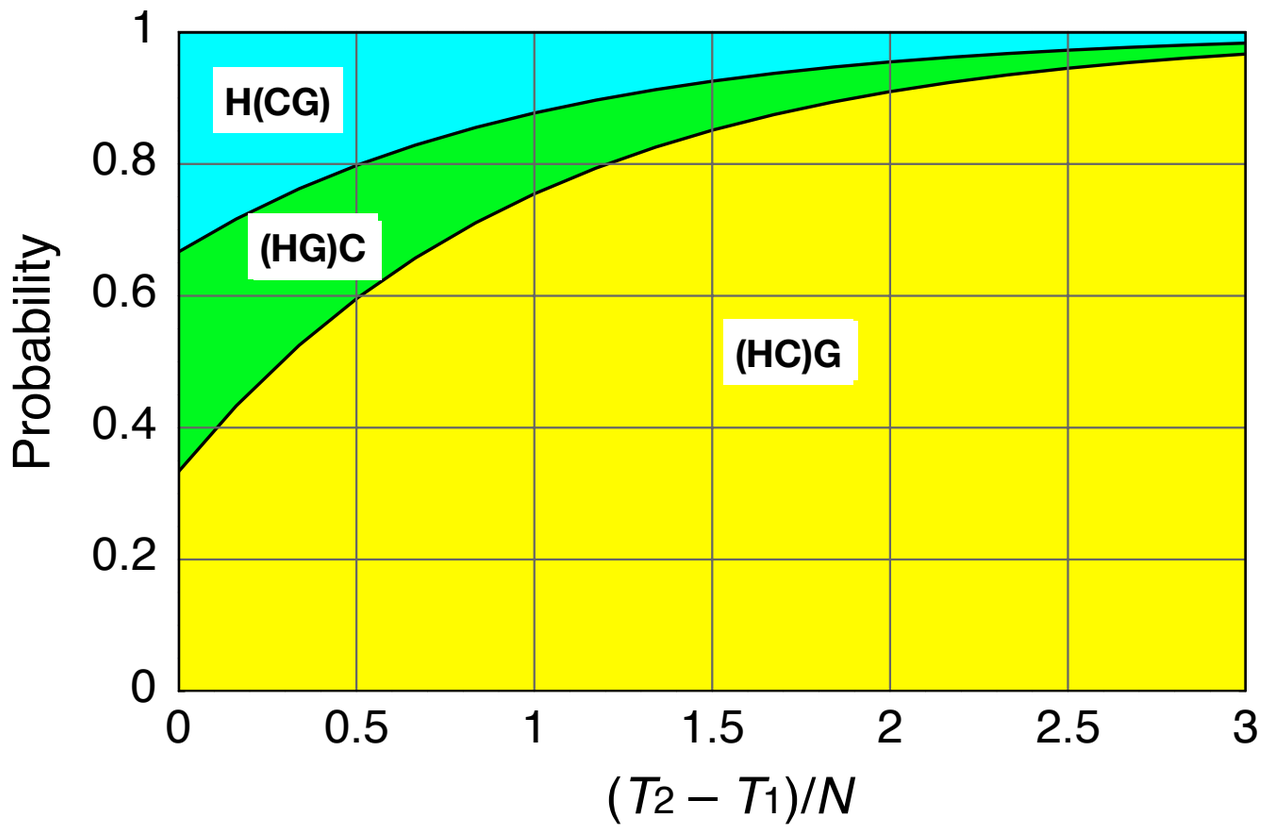
$$\mathbf{P}[((H,C),G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[((H,G),C)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[(H,(C,G))] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

Incomplete Lineage Sorting (ILS)





$$\mathbf{P}[((H,C),G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[((H,G),C)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[(H,(C,G))] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

Incomplete Lineage Sorting (ILS)

- A gene tree can be reconciled with a species tree under ILS using
    - a parsimony approach, which seeks to minimize the amount of “deep coalescence” of the gene tree within the branches of the species tree, and
    - a probabilistic approach, which seeks to maximize the probability of observing the gene tree given the species tree, using the coalescent framework.

Incomplete Lineage Sorting (ILS)

- The inference problem seeks a species tree from a collection of gene trees (or sequence alignments).

- Many approaches have been proposed: parsimony, likelihood, Bayesian, distance-based, and summary statistics.

Inferring Phylogenetic Relationships in the Post-Genomic Era: A New Paradigm

- The increasing availability of multi-locus data is highlighting the extent of incongruence between a species tree and its "contained" gene trees, as well as among the gene trees themselves, and the need for new methods to establish phylogenetic relationships in light of this incongruence

- The result is the emergence of a new paradigm that simultaneously accounts for

  - mutations within a locus (base pair mutations and indels), and

  - incongruence among loci (HGT, dup/loss, and ILS).

Method

Unified modeling of gene duplication, loss, and coalescence using a locus tree

Matthew D. Rasmussen[1] and Manolis Kellis[1]

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; Broad Institute, Cambridge, Massachusetts 02139, USA

Dup/Loss + HGT

BIOINFORMATICS

Vol. 28 ISMB 2012, pages i283–i291
doi:10.1093/bioinformatics/bts225

Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss

Mukul S. Bansal [1,*], Eric J. Alm [2] and Manolis Kellis [1,3,*]

[1]Computer Science and Artificial Intelligence Laboratory, [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and [3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

ILS + Hybridization

OPEN ACCESS Freely available online

PLoS GENETICS

The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection

Yun Yu[1], James H. Degnan[2,3], Luay Nakhleh[1]*

1 Department of Computer Science, Rice University, Houston, Texas, United States of America, **2** Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, **3** National Institute of Mathematical and Biological Synthesis, Knoxville, Tennessee, United States of America

Dup/Loss + HGT + ILS

BIOINFORMATICS

Vol. 28 ECCB 2012, pages i409–i415
doi:10.1093/bioinformatics/bts386

Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees

Maureen Stolzer[1,*], Han Lai[1], Minli Xu[2], Deepa Sathaye[3], Benjamin Vernot[4] and Dannie Durand[1,3]

[1]Department of Biological Sciences, [2]Lane Center for Computational Biology, [3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA and [4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

Keep In Mind...

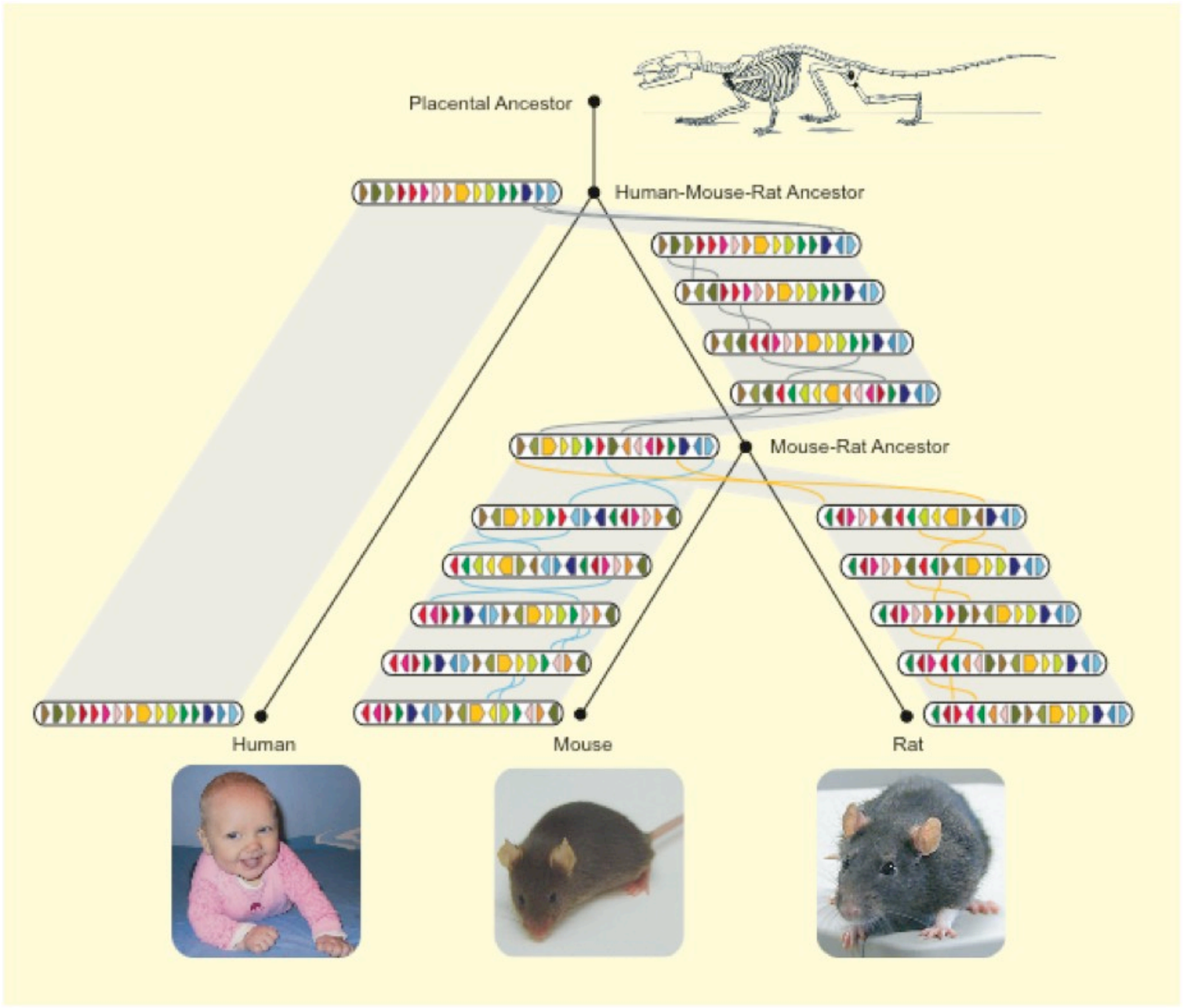
- In practice, gene trees are estimated from sequence data.
- Gene tree estimates may be inaccurate.
- These inaccuracies in the gene tree estimates give rise to incongruence similar to that caused by true evolutionary events.
- It is important to recognize this and account for errors in the gene tree estimates before or during the species phylogeny inference process.

Genome Rearrangements

Genome Rearrangements

- In addition to HGT and dup/loss, other “large” mutational events act on the genome:
    - transpositions
    - translocations
    - inversions
    - fusions
    - ...

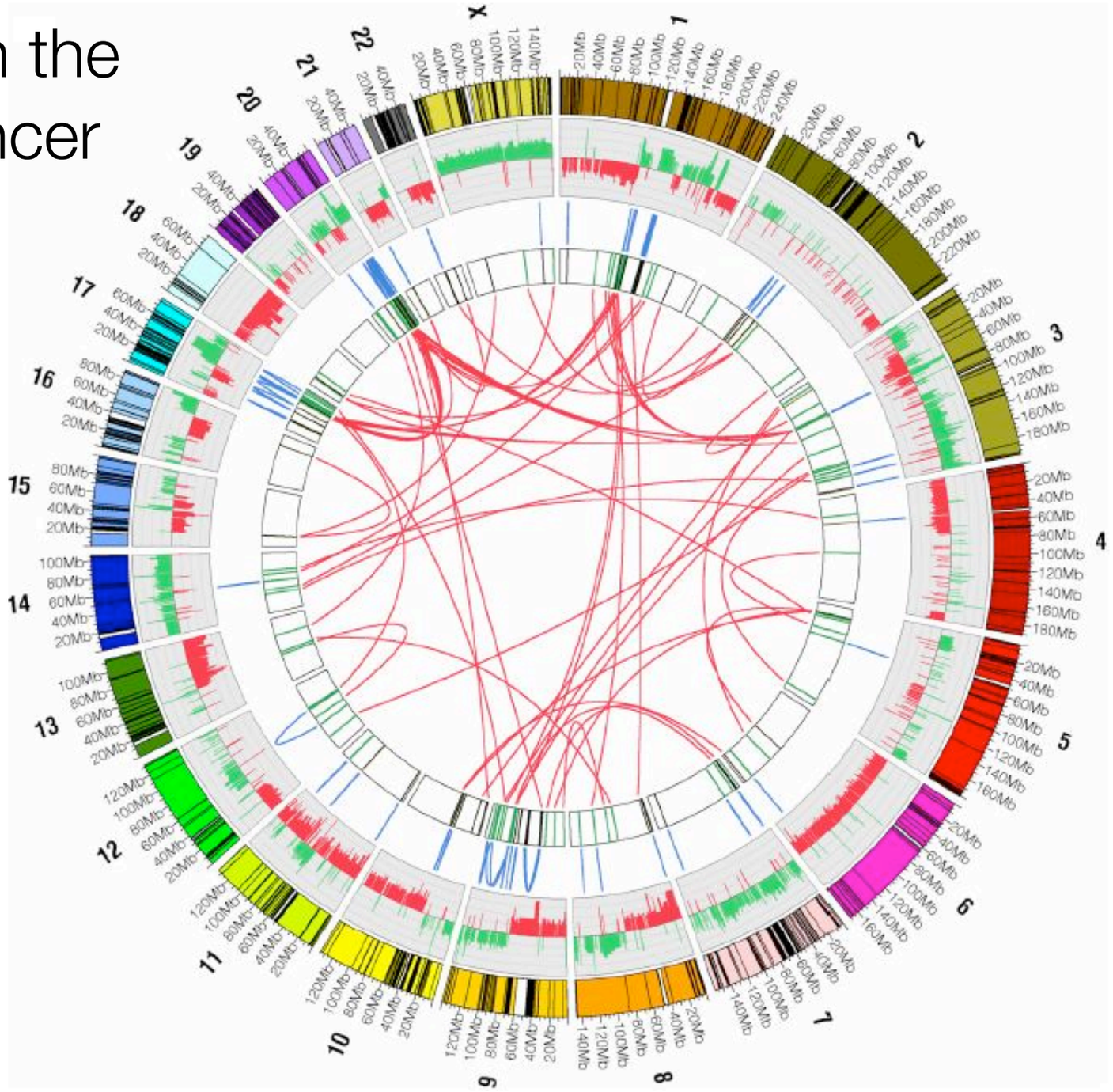
Genome Rearrangements



[Source: Bourque et al., Genome Research, 12(1):26-36,2002.]

Genome Rearrangements

Rearrangements in the MCF-7 breast cancer cell line



[Source: Hampton et al., Genome Research, 19(2):167-177,2009.]

Genome Rearrangements

- Genome representation: The molecular sequence information of genes is abstracted out, and the genome is turned into a list of signed numbers, where each element in the list corresponds to a gene, and the sign corresponds to the direction (strand) on the genome.

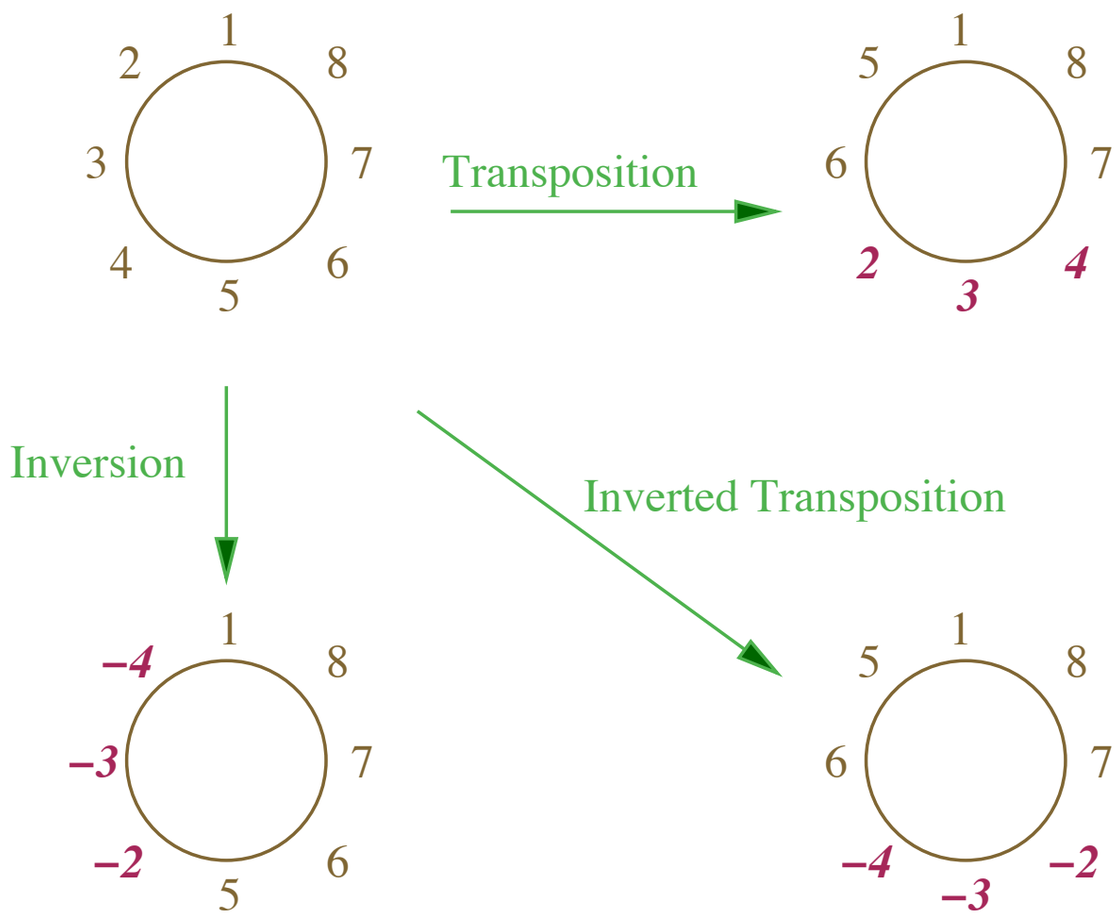
- Under this representation, a genome is viewed as a single character with a very large state space (all possible permutations of the list), and the evolution of this character is sought for a set of species.

Genome Rearrangements

```
G1=(1   2   3   4   5   6   7   8)
```

```
G2=(1   2  -5  -4  -3   6   7   8)
```

breakpoints (arrows) are missing adjacencies



[Source: Slides on Comparative Genomics, by B.M.E. Moret, PSB 2010]

Genome Rearrangements

- Evolutionary models for genome rearrangements include
  - parsimony: minimize the number of “allowed” rearrangement events that transform one genome into another
  - “weighted”:
    - The Nadeau-Taylor and Generalized Nadeau-Taylor models
    - The double-cut-and-join (DCJ) model

Genome Rearrangements

- Computational problems include:
    - Computing the distance between two genomes
    - Ancestral reconstruction of genomes
    - The “median problem” (given three genomes, find a genome that minimizes the sum of distances to all three genomes)
    - Multiple alignment of genomes
    - Phylogeny reconstruction from genomes

Challenges

- Currently, our best understanding, from a modeling perspective, is in the area of incomplete lineage sorting (thanks to the coalescent model). There is need for similar models for the other events.

- The holy grail: a model that captures all events simultaneously!

- Ultimately, we are interested in inference. Most methods currently employ the parsimony principle. There is need for probabilistic inference.

- Scalability: Current methods, even if accurate, are too slow to analyze large data sets. There is need for efficient algorithms and high-performance computing techniques.

Summary

- Phylogenomic analyses often involve dealing with incongruent gene trees and/or genome rearrangements.

- It is important to infer the evolutionary mechanism, or mechanisms, that gave rise to the incongruence.

- Inferring species phylogenies (trees or networks) in the post-genomic era requires accounting for “evolution within a gene” (e.g., nucleotide evolution) and “evolution within and across the branches of the species phylogeny” (i.e., gene tree incongruence).

- The evolution of networks should be tied to the evolution of genes/genomes!