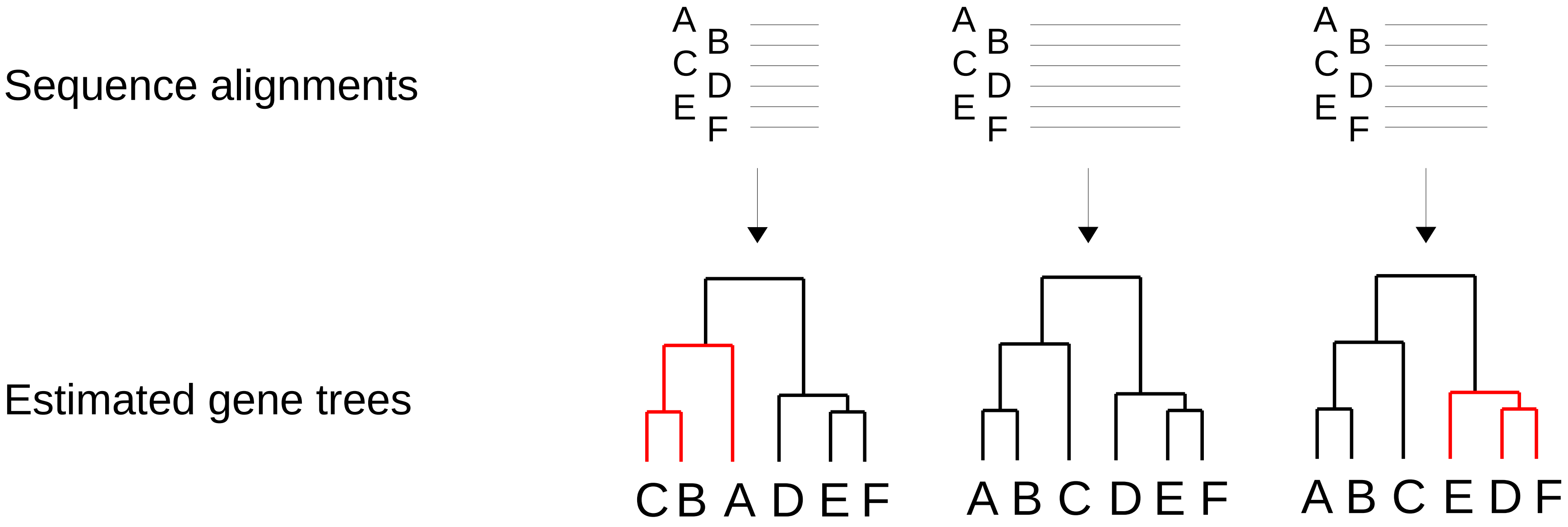
MCEB 2016

Incomplete lineage sorting in mammalian phylogenomics

Céline Scornavacca, Nicolas Galtier

Institute of Evolutionary Sciences
CNRS – University Montpellier

Phylogenetic conflict



Alignment errors

Tree-building errors

GENE TREES TRULY IN AGREEMENT

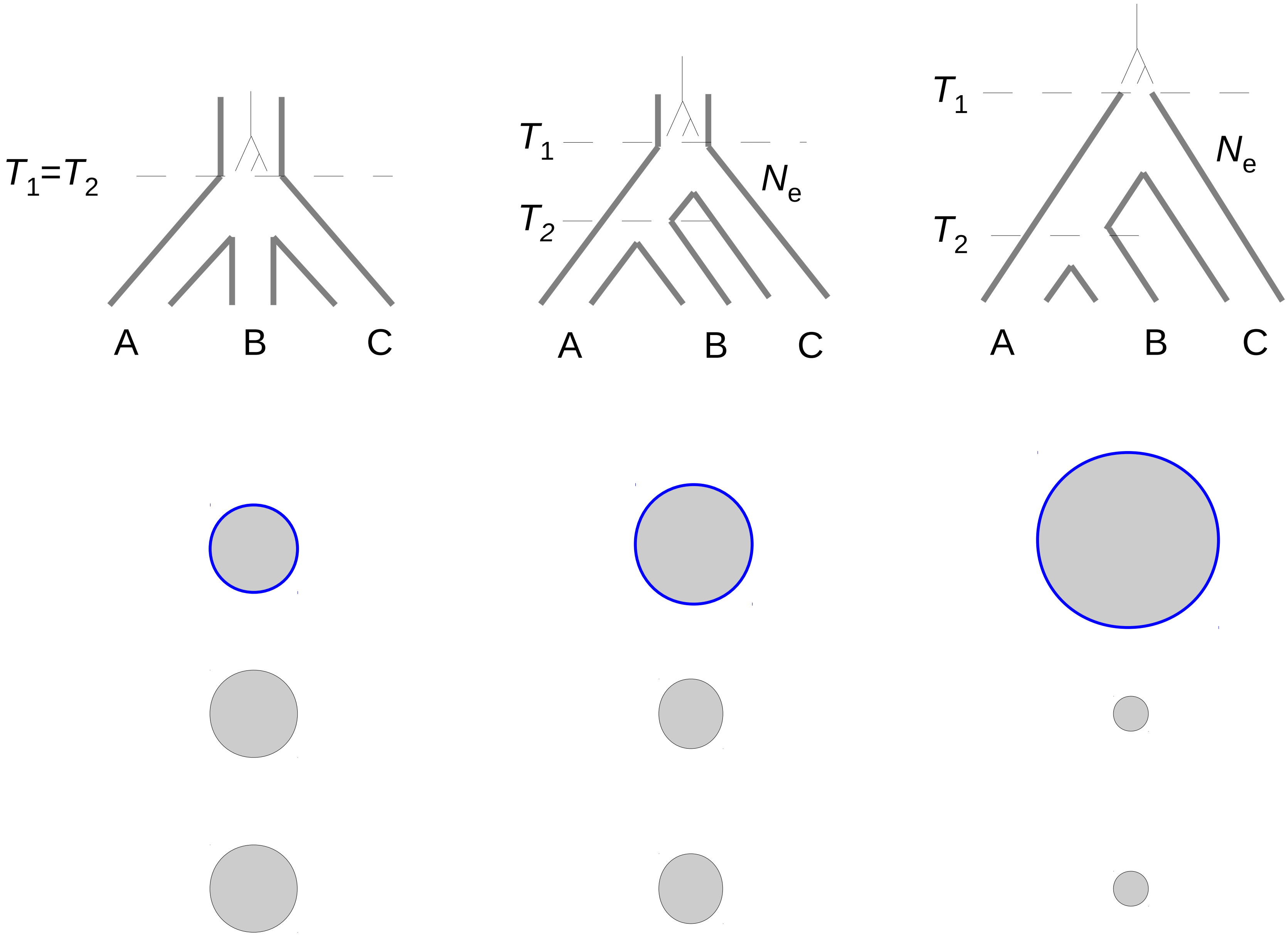
Horizontal gene transfer

Hidden paralogy

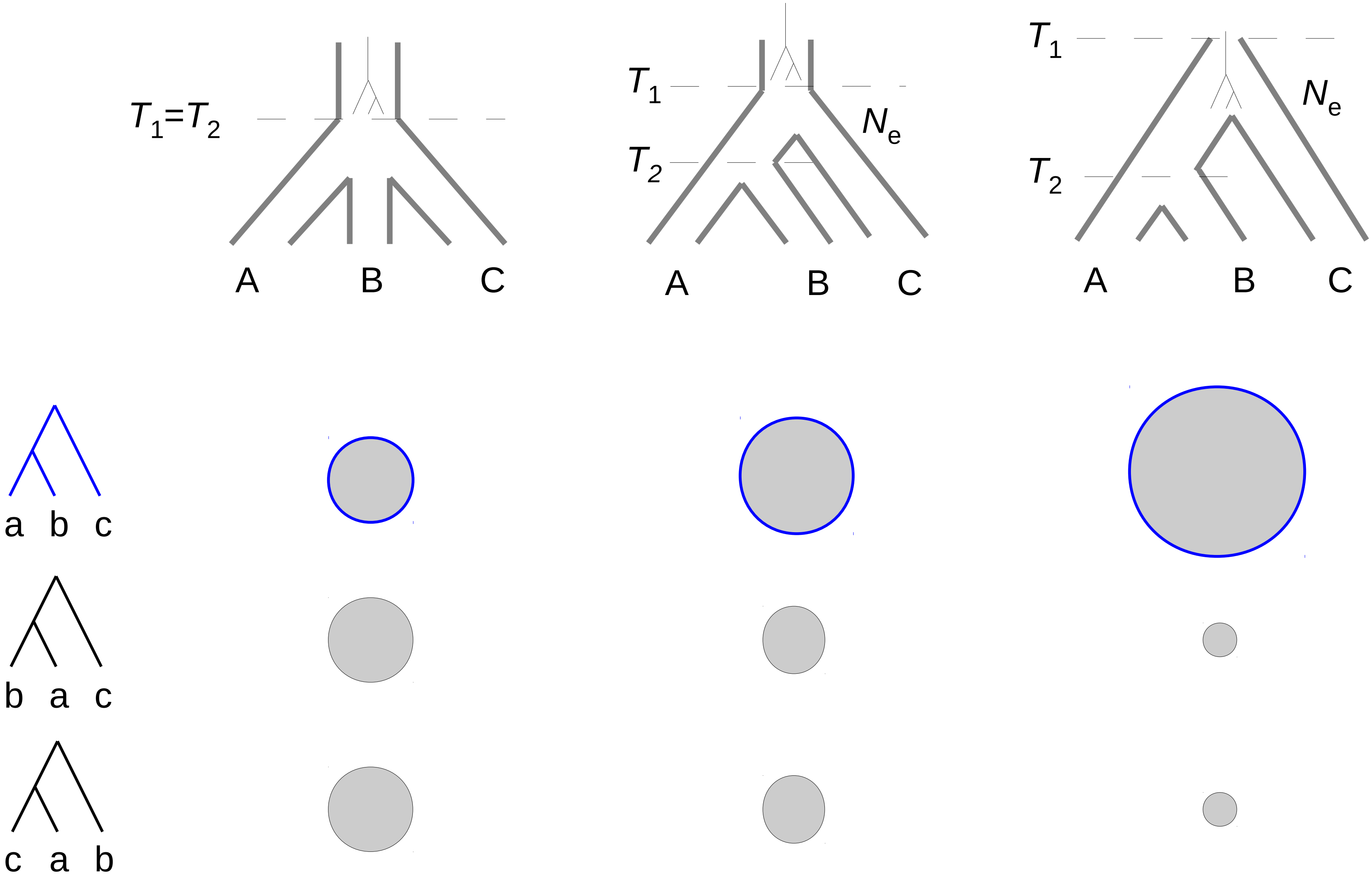
Incomplete Lineage Sorting (ILS)

GENE TREES TRULY CONFLICTING

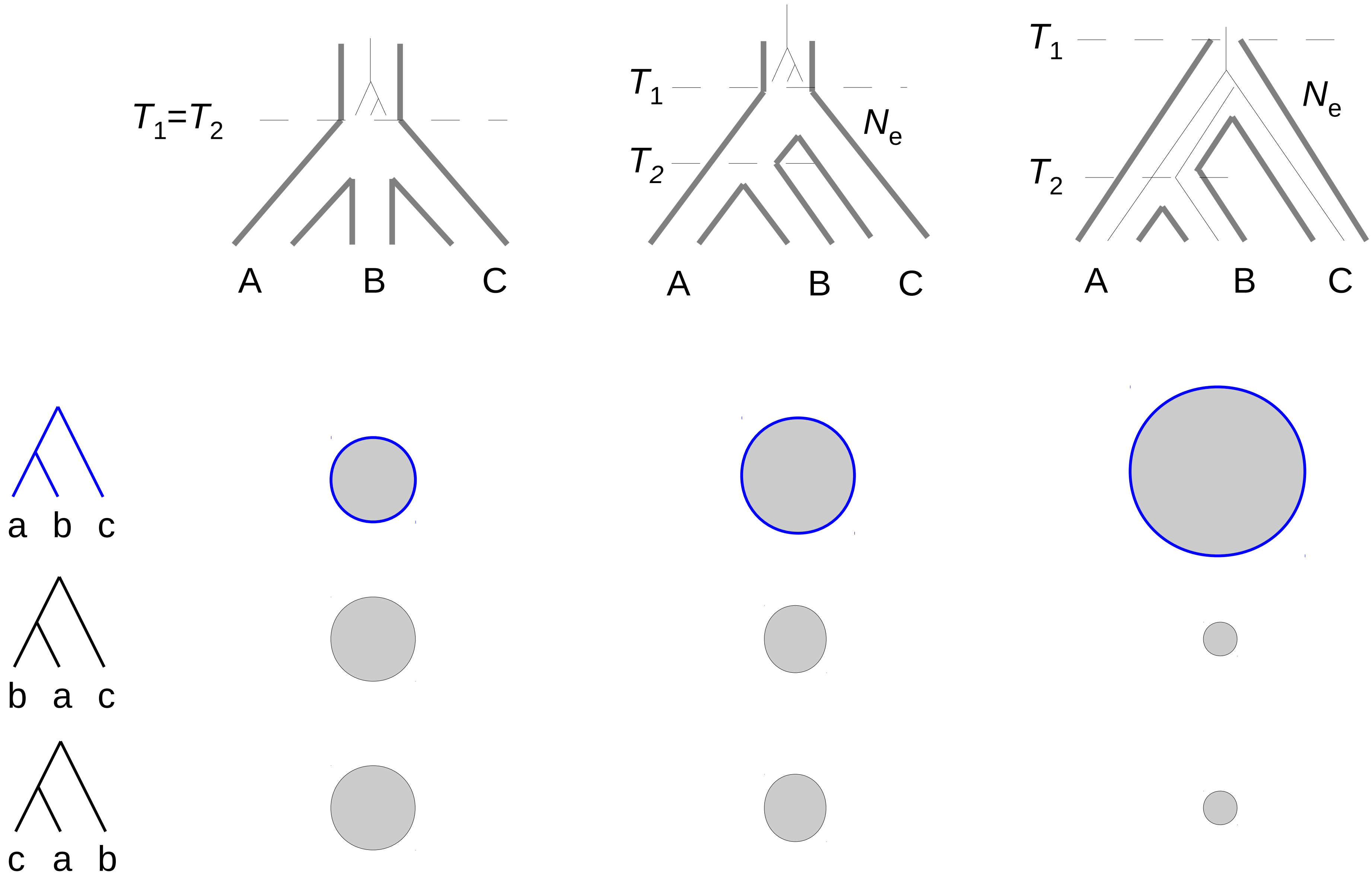
Incomplete lineage sorting and the probability of gene trees



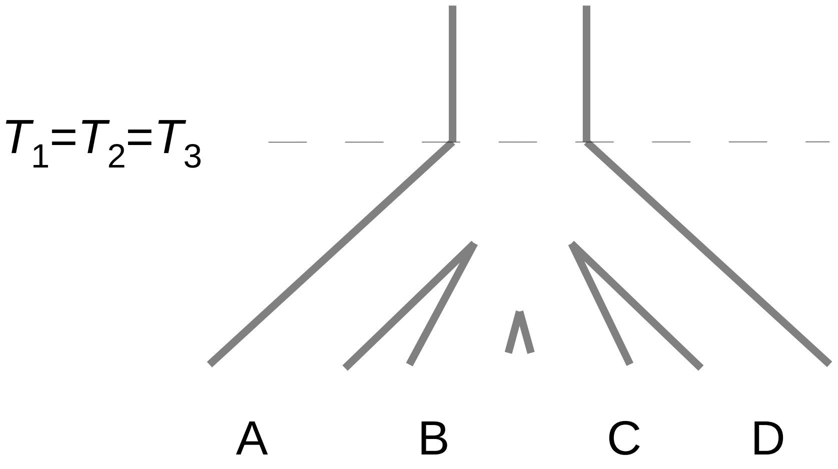
Incomplete lineage sorting and the probability of gene trees

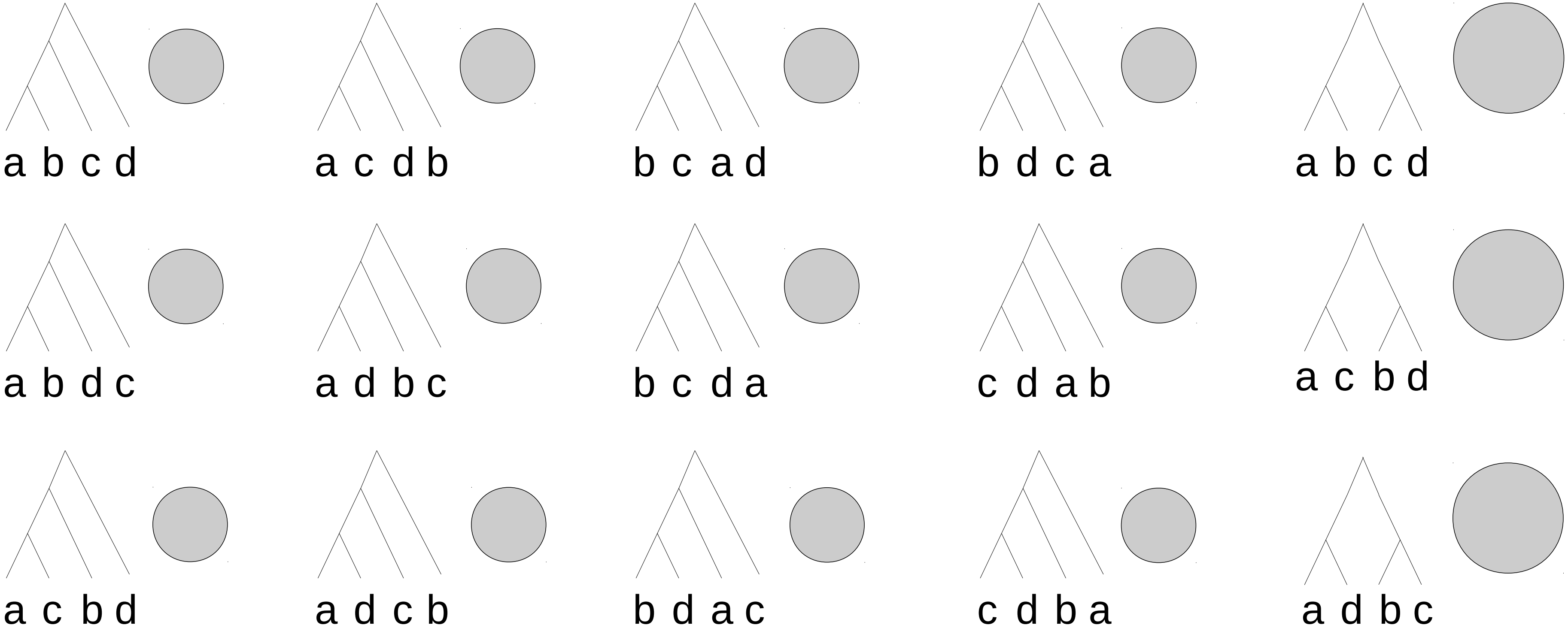


Incomplete lineage sorting and the probability of gene trees

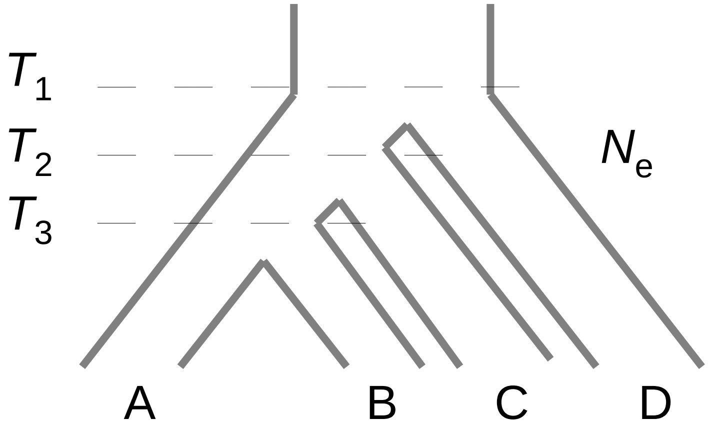


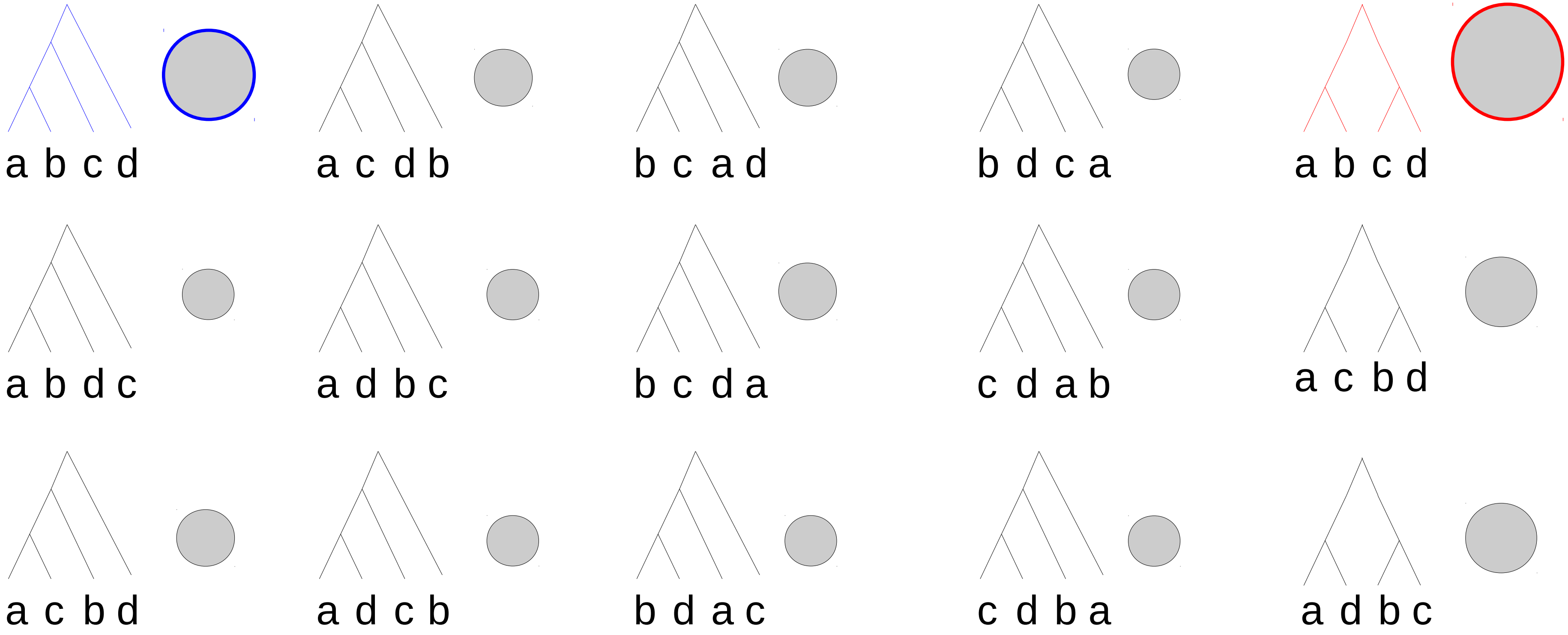
Incomplete lineage sorting and the probability of gene trees: the anomaly zone





Incomplete lineage sorting and the probability of gene trees: the anomaly zone





- Standard phylogenetic methods can be **inconsistent under ILS**

- **New methods** have been developed to cope for this bias, and often applied to gene trees estimated from **protein coding sequences**

A hot topic:

Gatesy & Springer 2013
Gatesy & Springer 2014
Springer & Gatesy 2014
Springer & Gatesy 2016

Wu, Song, Liu & Edwards 2013
Liu, Xi, Wu, Davis, Edwards 2015
Edwards et al. 2016

- Standard phylogenetic methods can be **inconsistent under ILS**

2. Is this problematic?

- **New methods** have been developed to cope for this bias, and often applied to gene trees estimated from **protein coding sequences**

1. Is this justified?

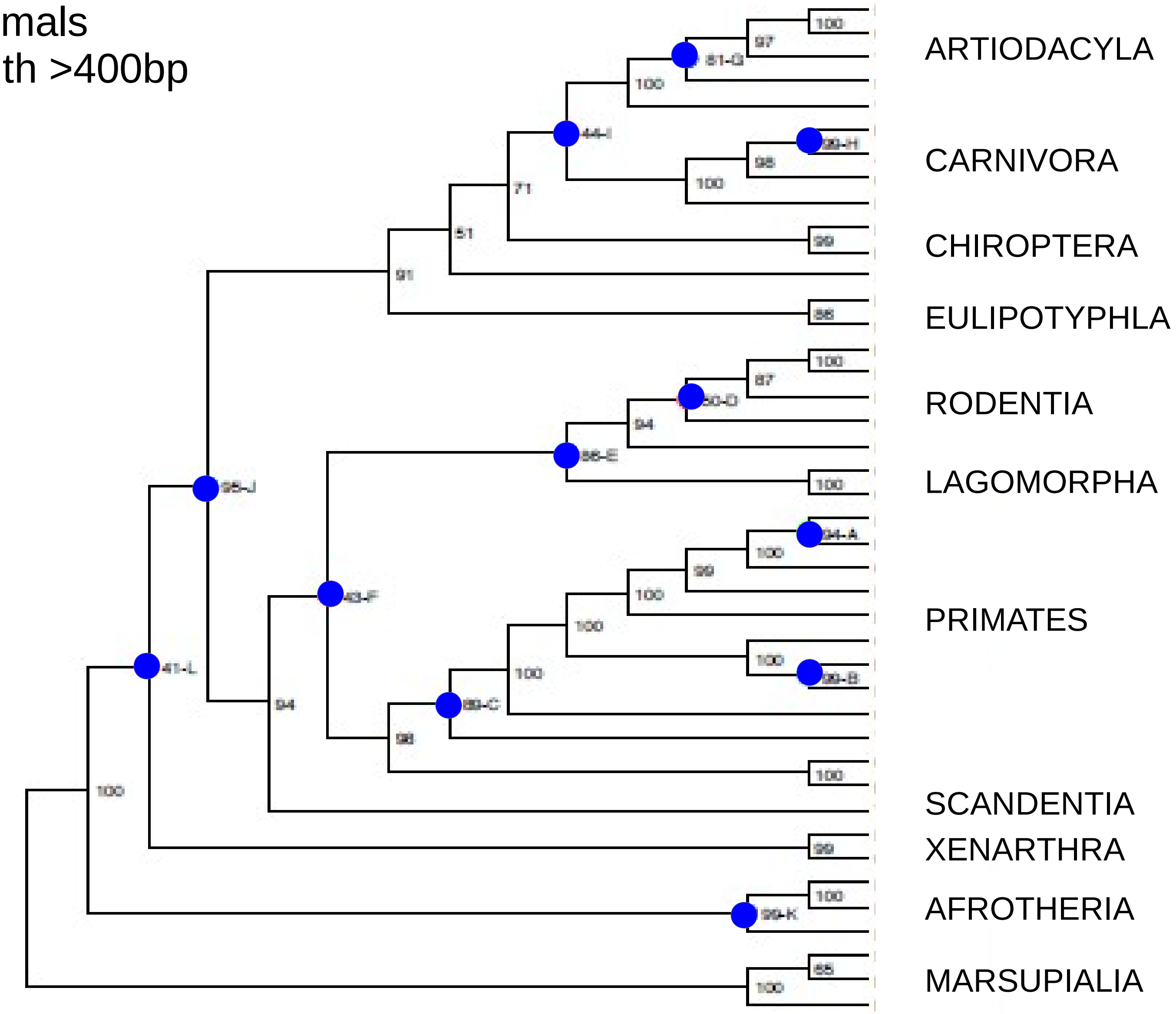
A hot topic:

Gatesy & Springer 2013
Gatesy & Springer 2014
Springer & Gatesy 2014
Springer & Gatesy 2016

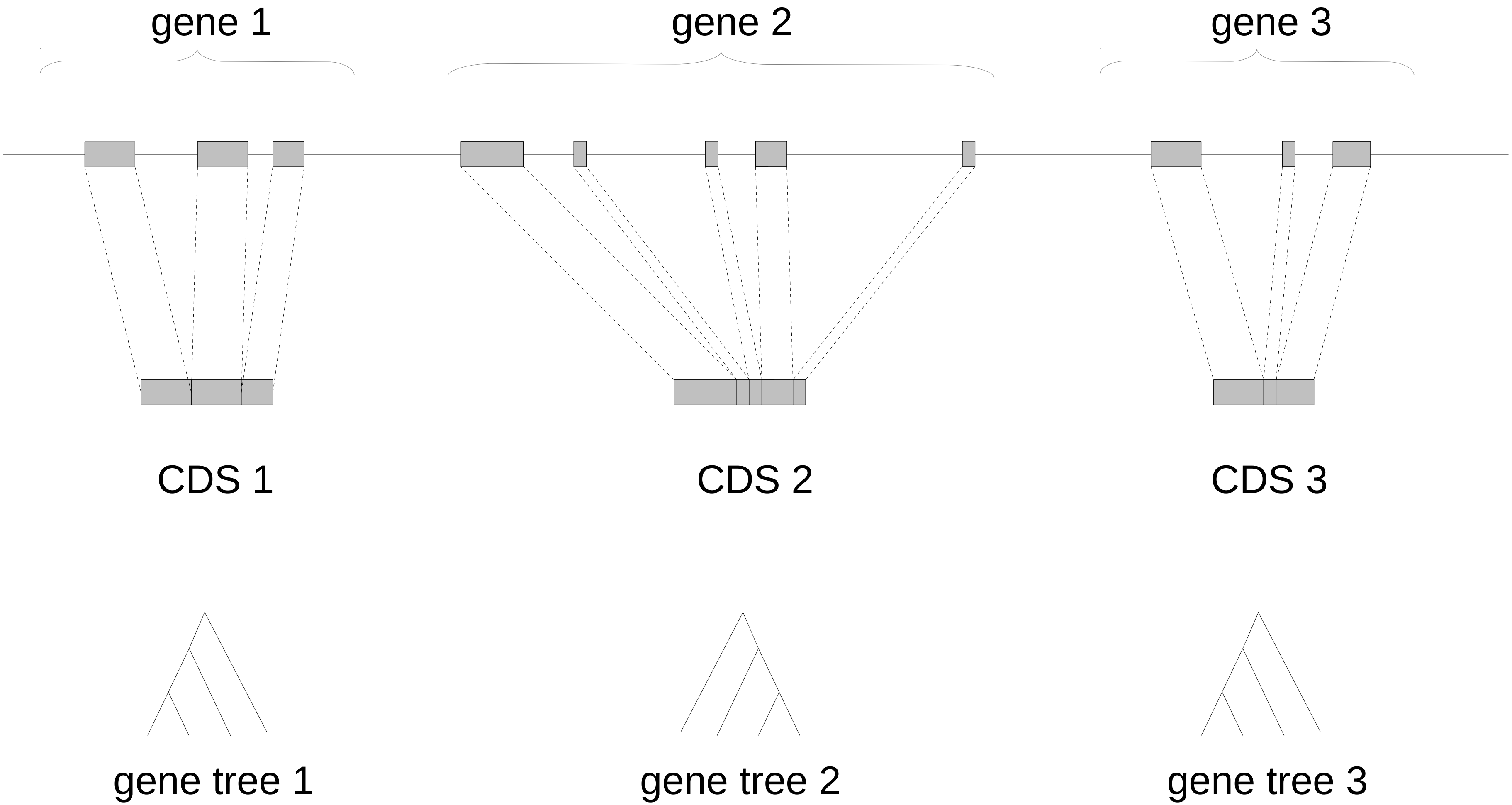
Wu, Song, Liu & Edwards 2013
Liu, Xi, Wu, Davis, Edwards 2015
Edwards et al. 2016

Data set: OrthoMaM v9

43 species of mammals
7349 exons of length >400bp



CDS-based, ILS-aware gene tree analysis



gene 1

gene 2

gene 3

CDS 1

CDS 2

CDS 3

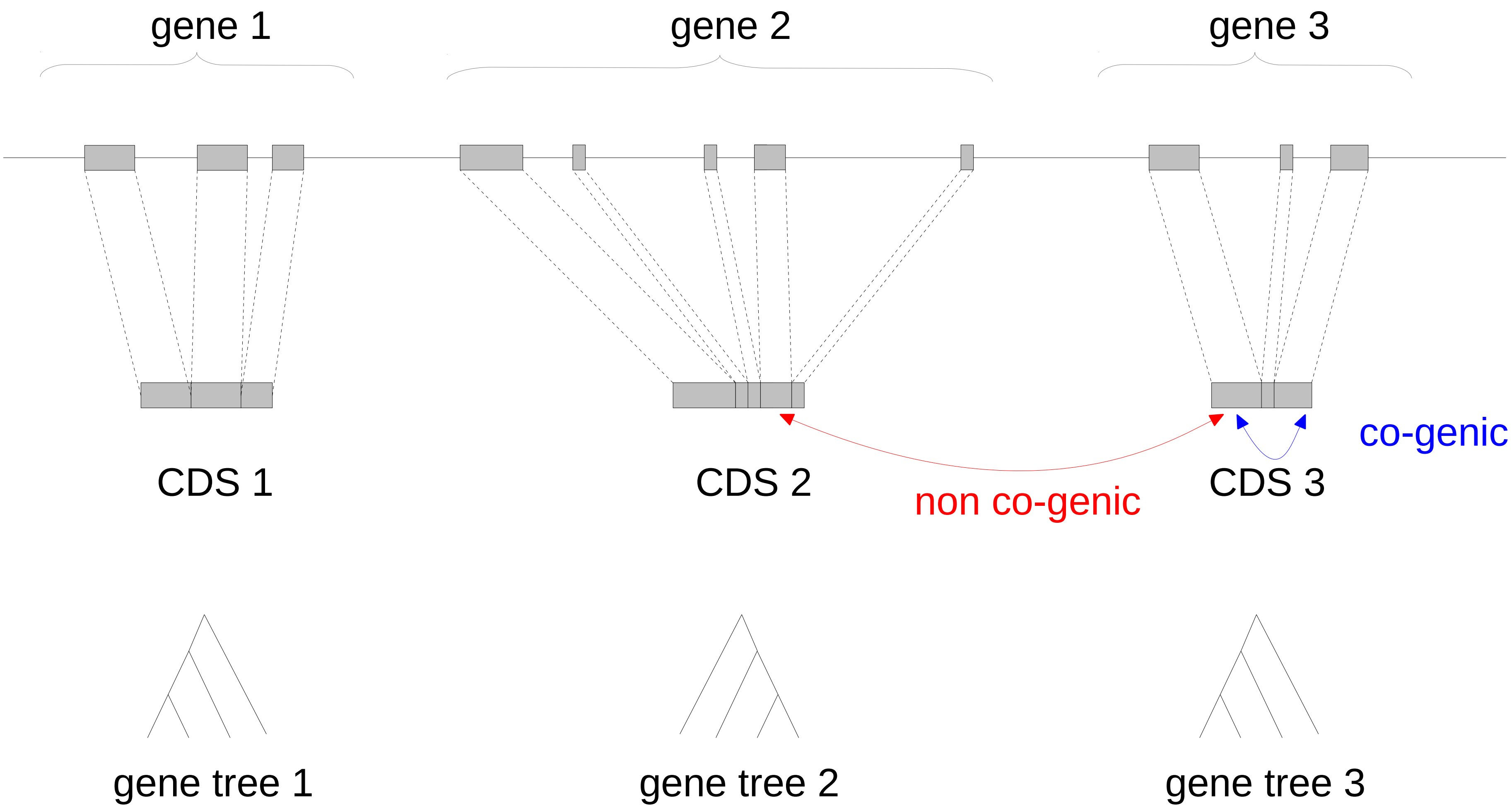
gene tree 1

gene tree 2

gene tree 3

Implicit assumption: the distinct exons of a gene share a common genealogy

CDS-based, ILS-aware gene tree analysis



gene 1

gene 2

gene 3

CDS 1

CDS 2

CDS 3

co-genic

non co-genic

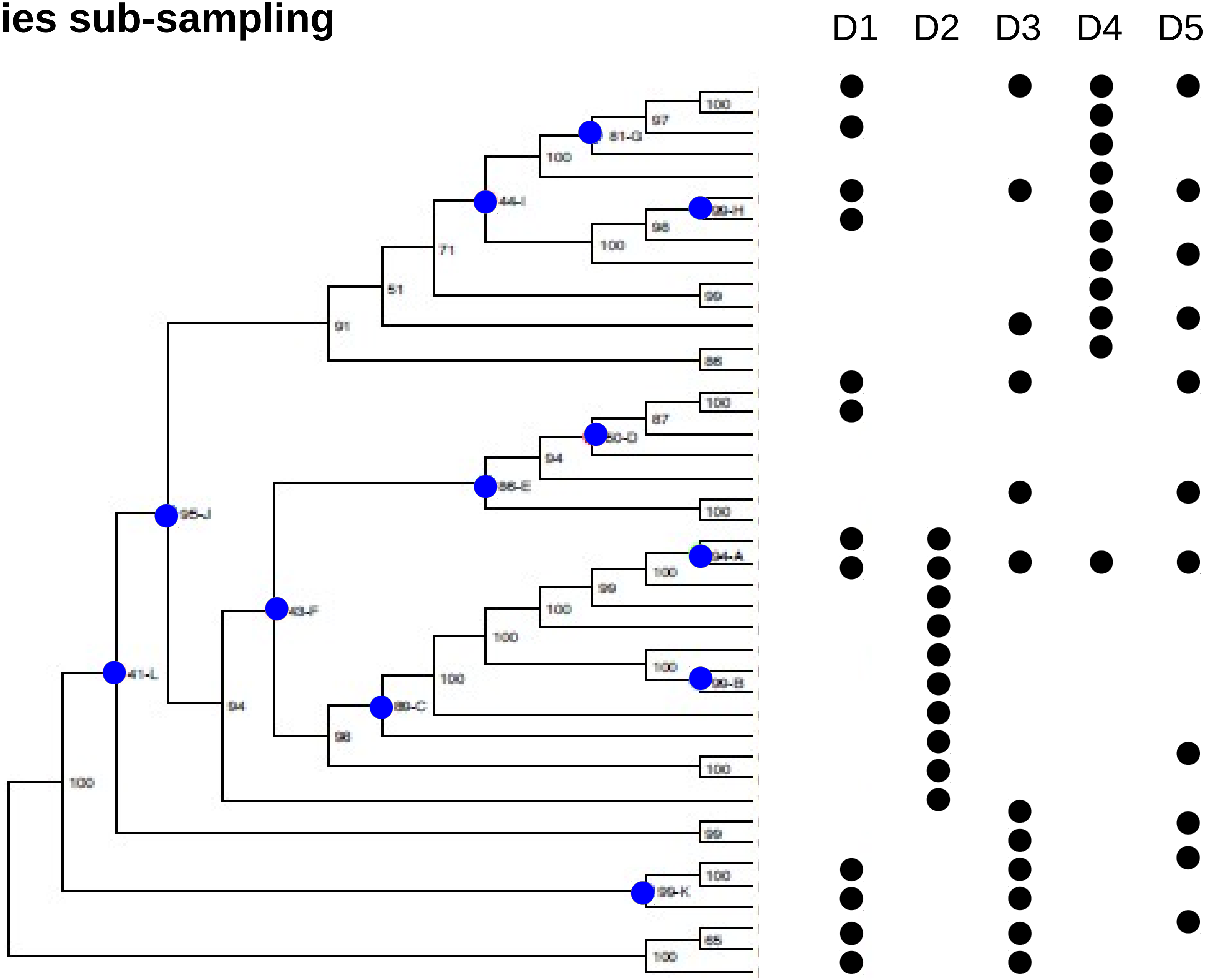
gene tree 1

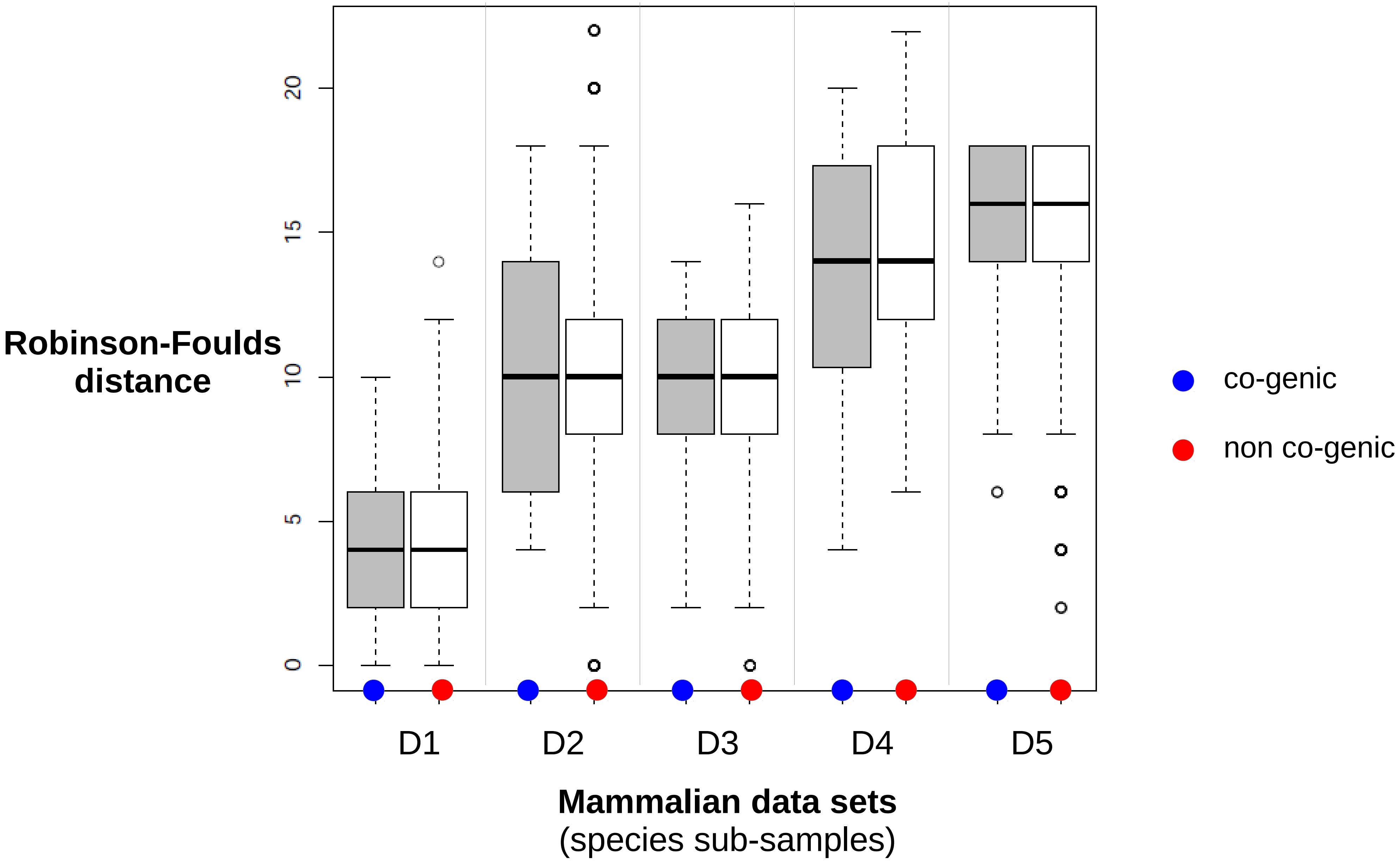
gene tree 2

gene tree 3

Implicit assumption: the distinct exons of a gene share a common genealogy

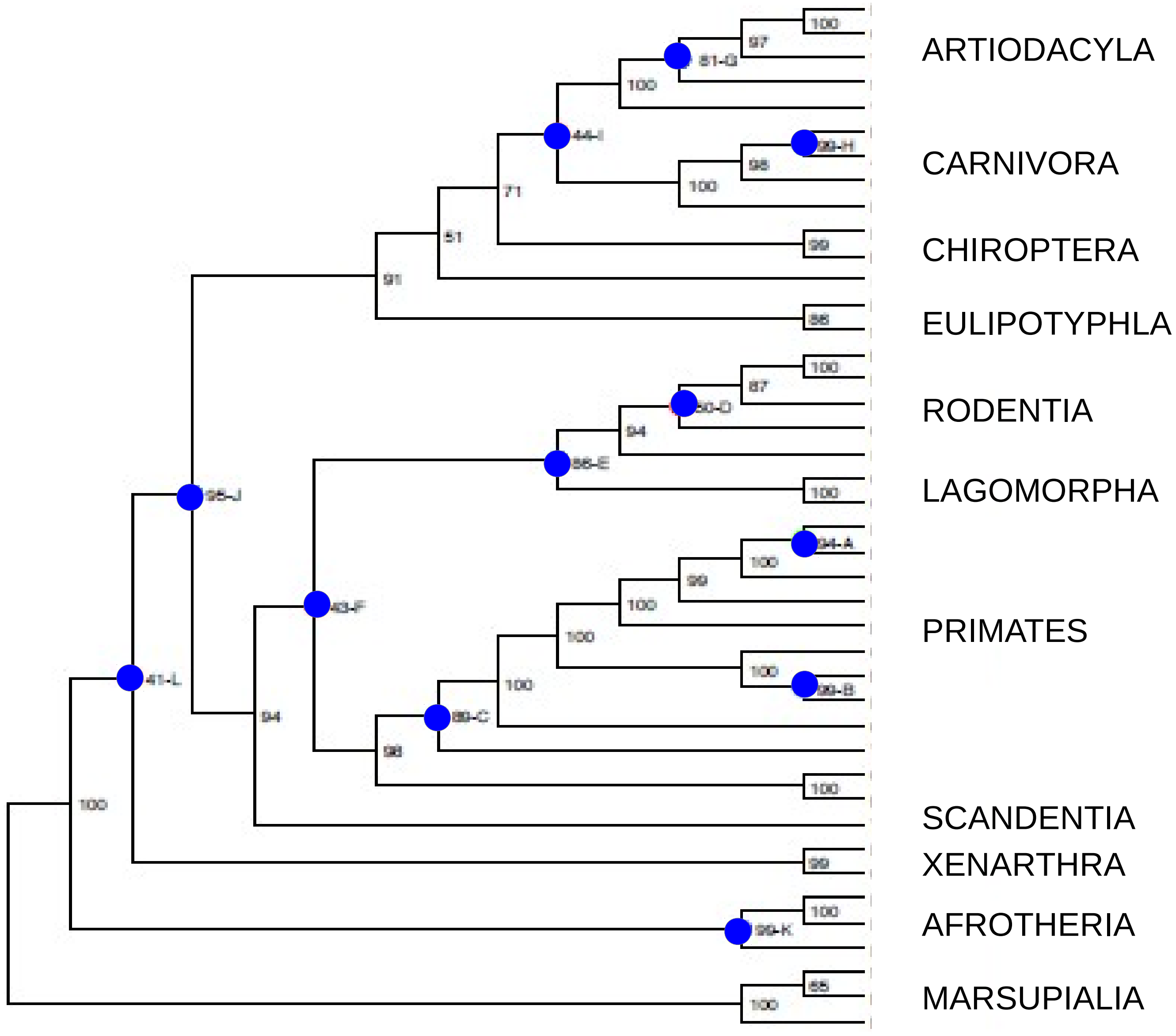
Species sub-sampling



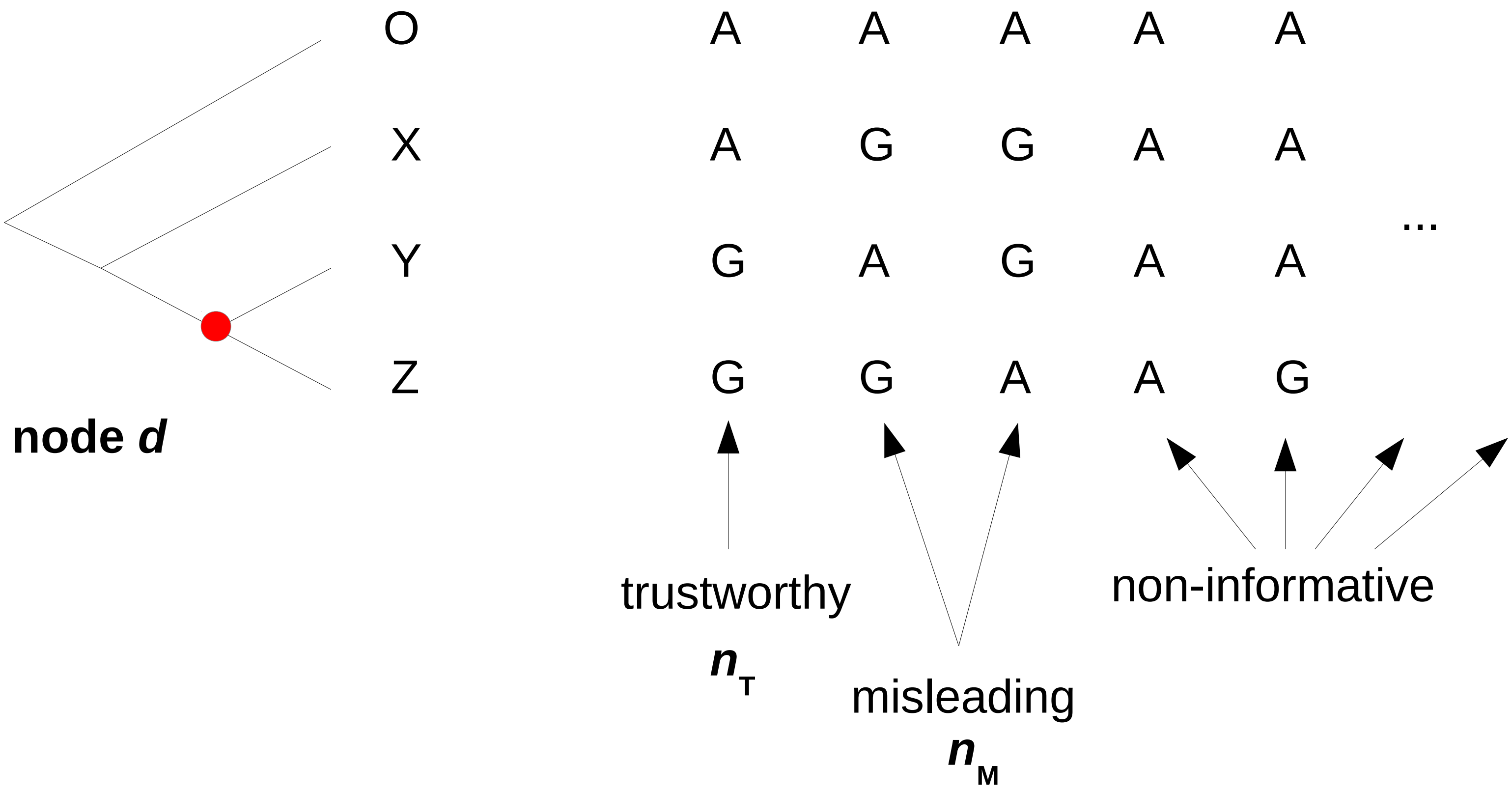


Co-genic exons do not yield gene trees more similar than exons from distinct genes

Reference mammalian phylogeny, focal nodes



alignment exon *e*



O	A	A	A	A	A	
X	A	G	G	A	A	
Y	G	A	G	A	A	…
Z	G	G	A	A	G	

node *d*

trustworthy

$\boldsymbol{n}_{\mathbf{T}}$

misleading

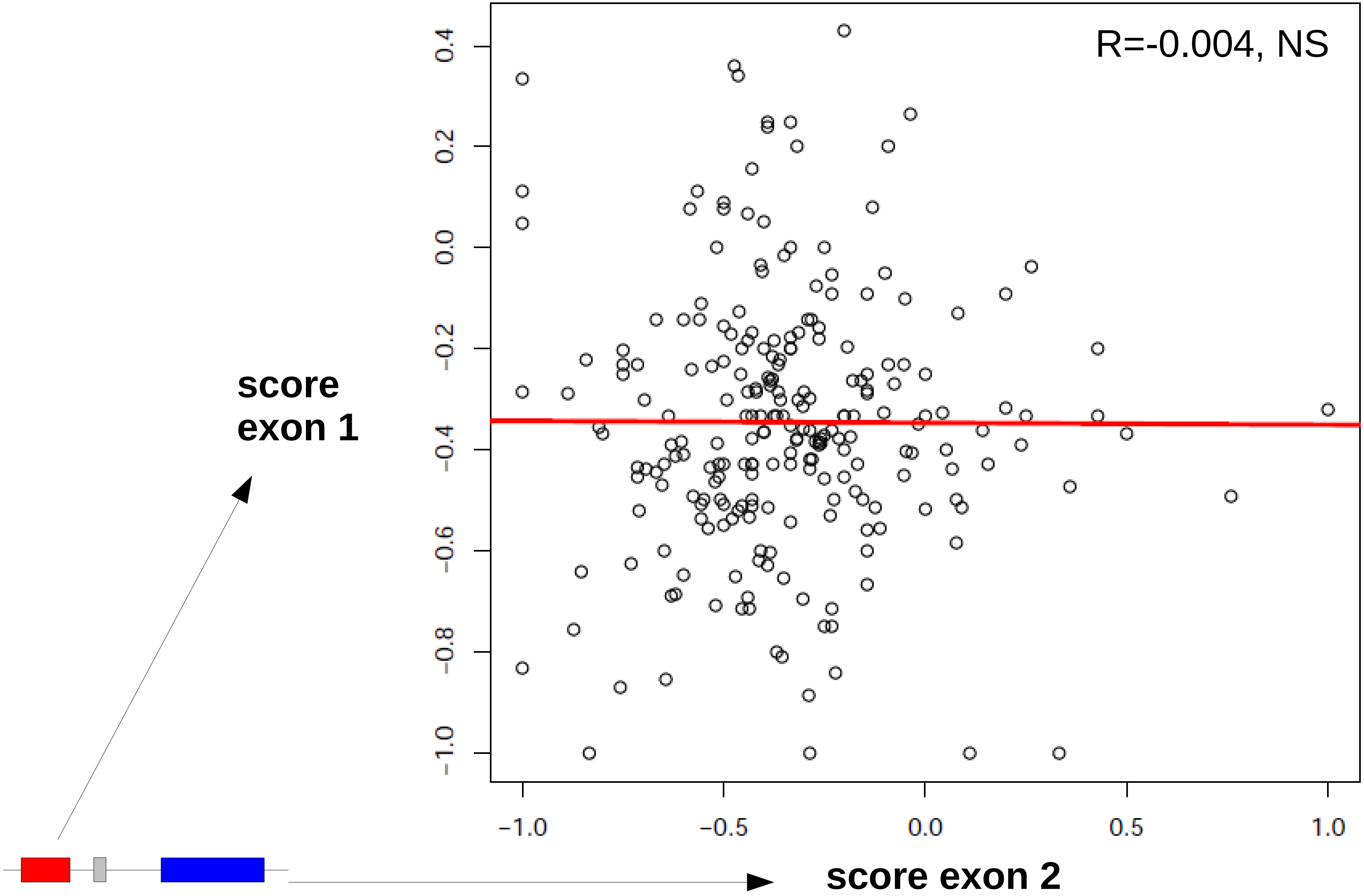
$\boldsymbol{n}_{\mathbf{M}}$

non-informative

Score of exon *e* at node *d*: $q(e,d) = \dfrac{n_T - n_M}{n_T + n_M}$

Node I

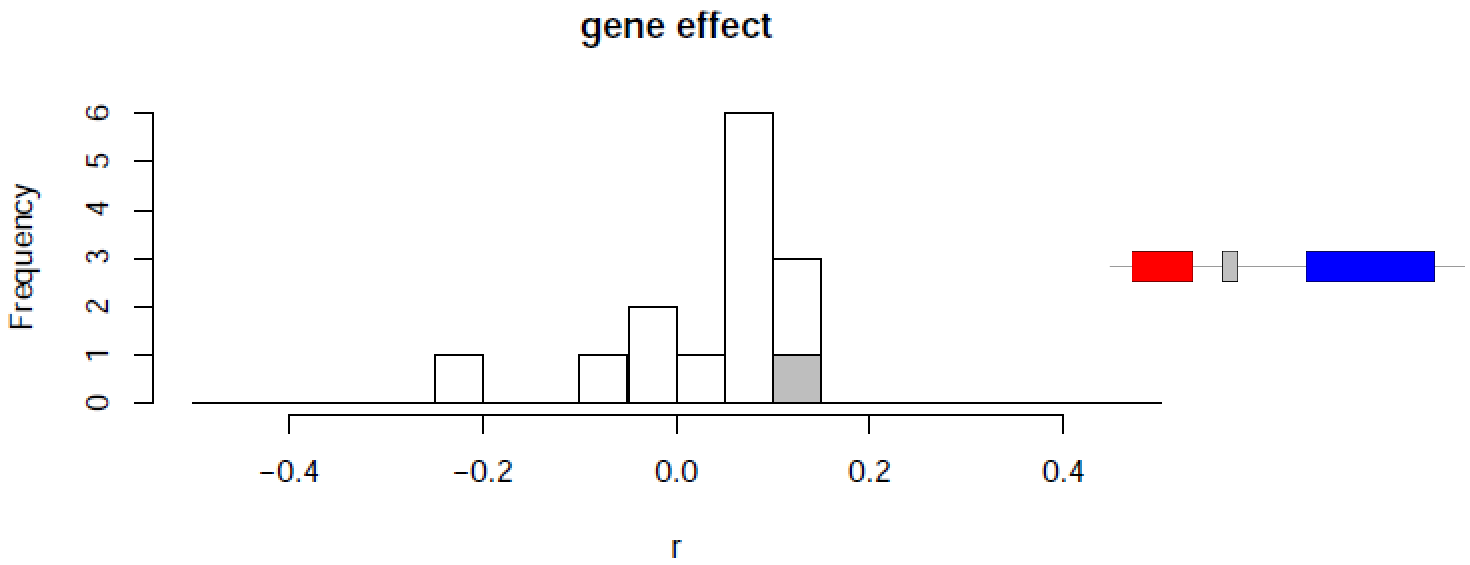
O=Eulipotyphla; X=Chiroptera; Y=Carnivora; Z=Cetartiodactyla



score exon 1

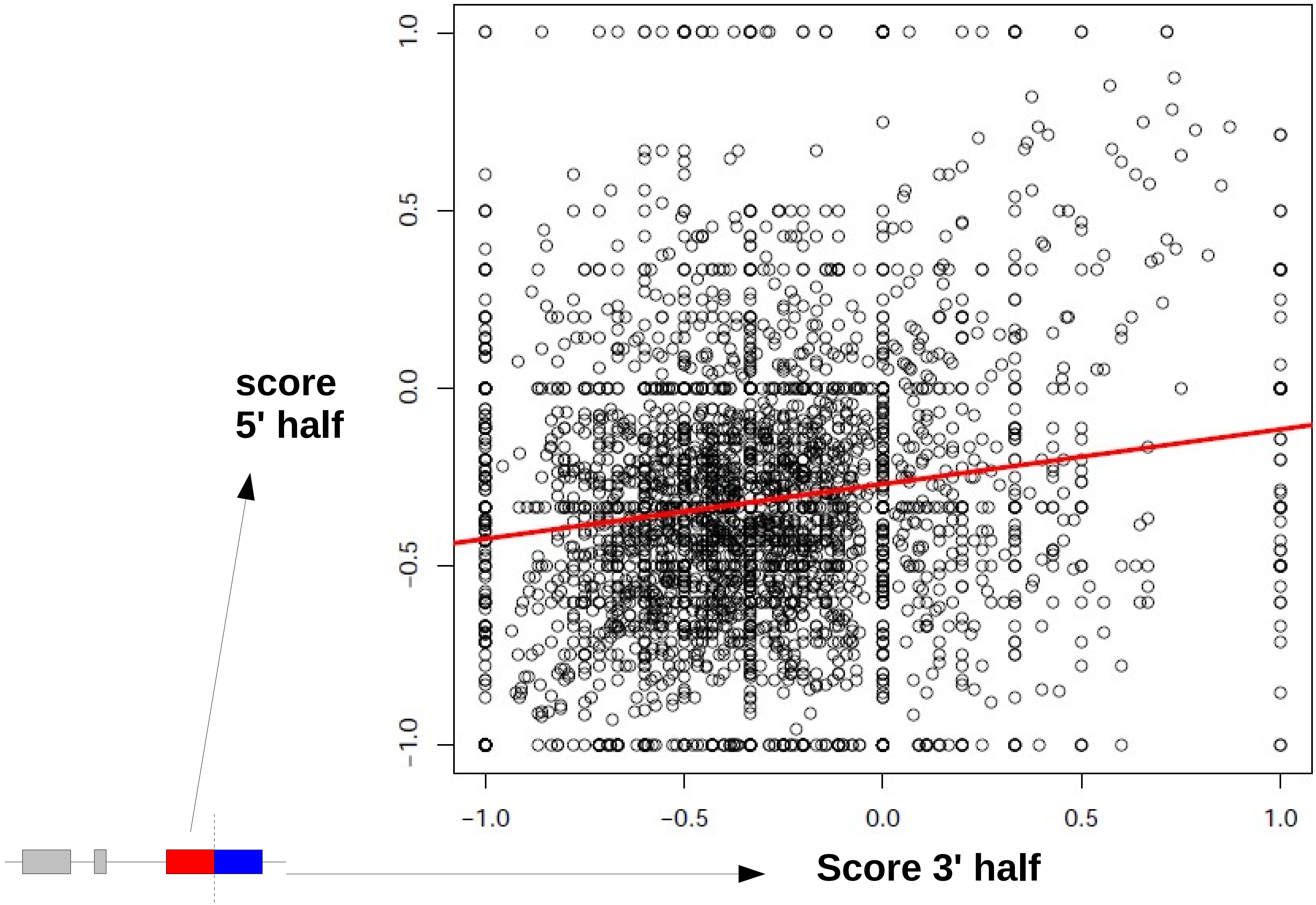
score exon 2

co-genic exons


gene effect
Frequency
0
1
2
3
4
5
6
−0.4
−0.2
0.0
0.2
0.4
r

Node I

O=Eulipotyphla; X=Chiroptera; Y=Carnivora; Z=Cetartiodactyla



R=0.15

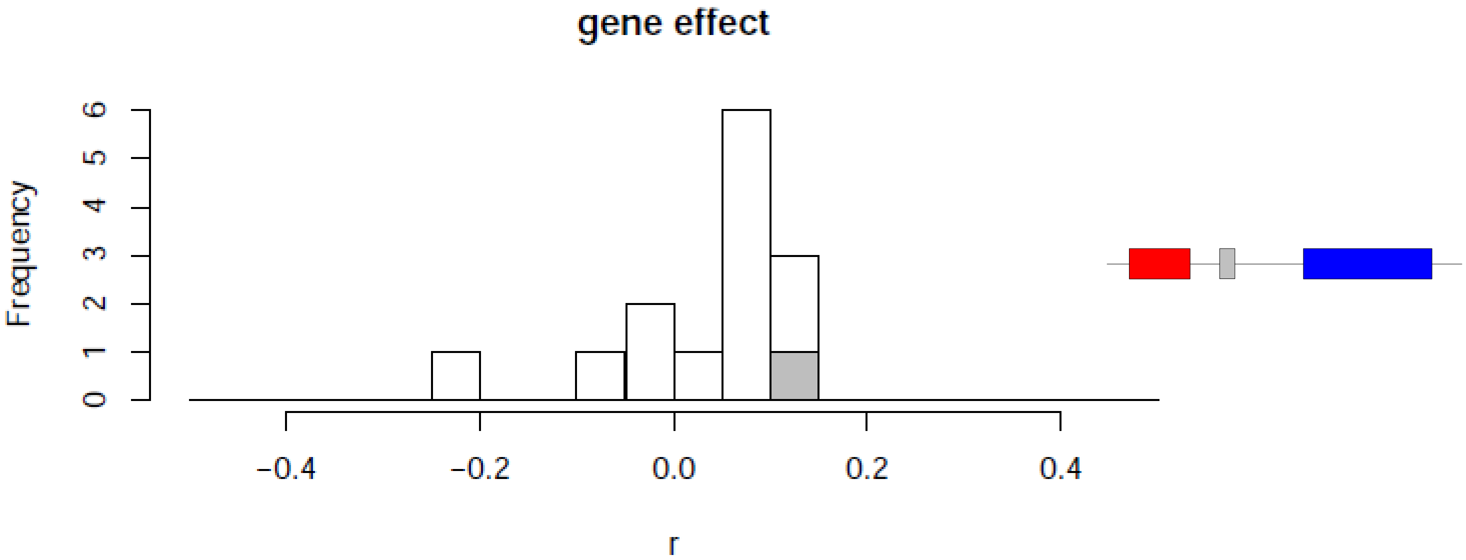
p-val<10^{-15}

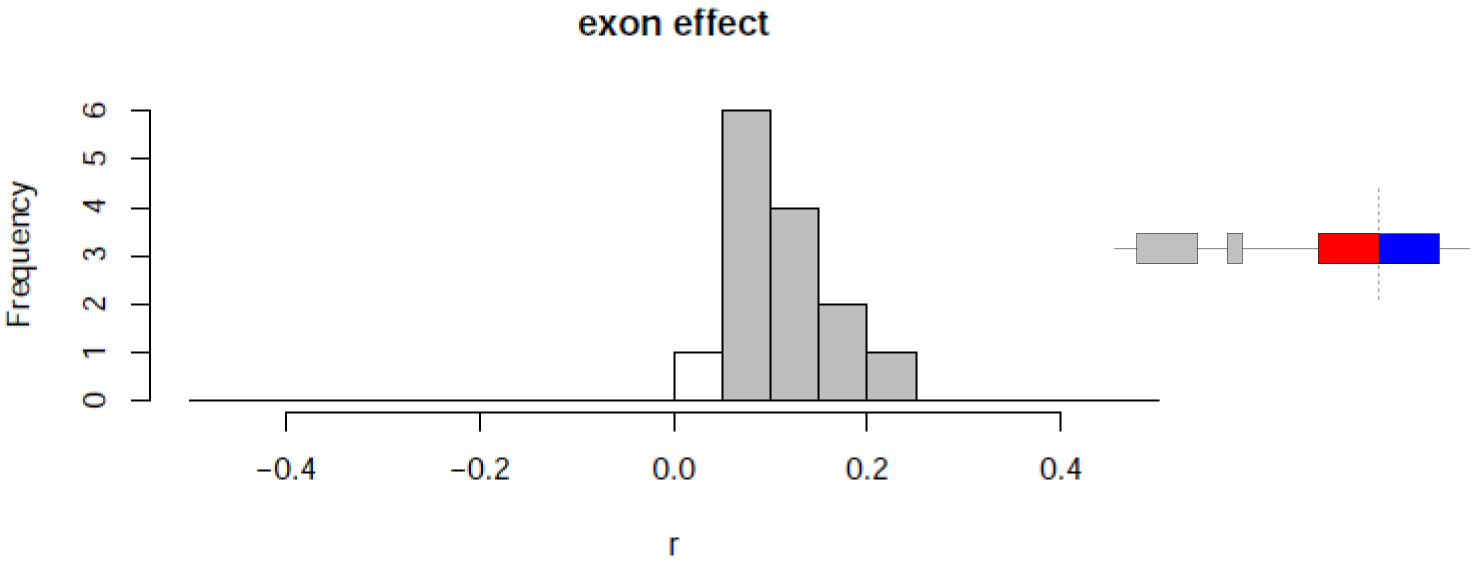
score 5' half

Score 3' half

half-exons

Exons, not CDS, are plausible phylogenomic units in mammals


gene effect


exon effect

- Standard phylogenetic methods can be **inconsistent under ILS**

2. Is this problematic?

- **New methods** have been developed to cope for this bias, and often applied to gene trees estimated from **protein coding sequences**

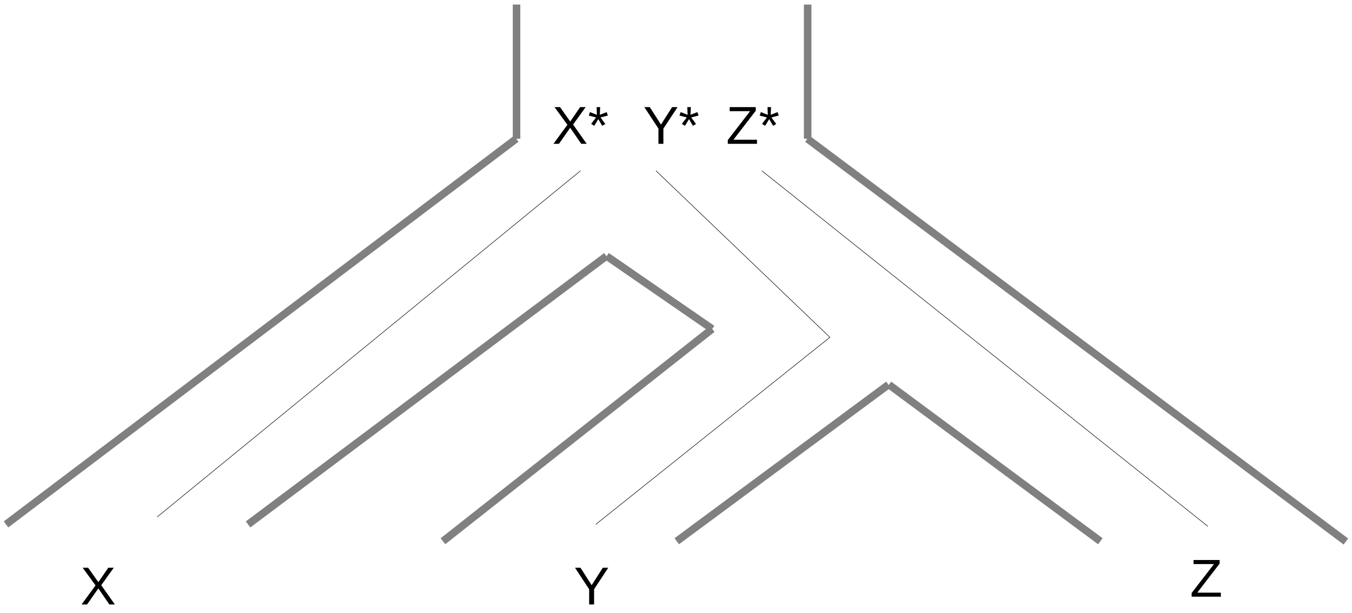
1. Is this justified? NO

A hot topic:

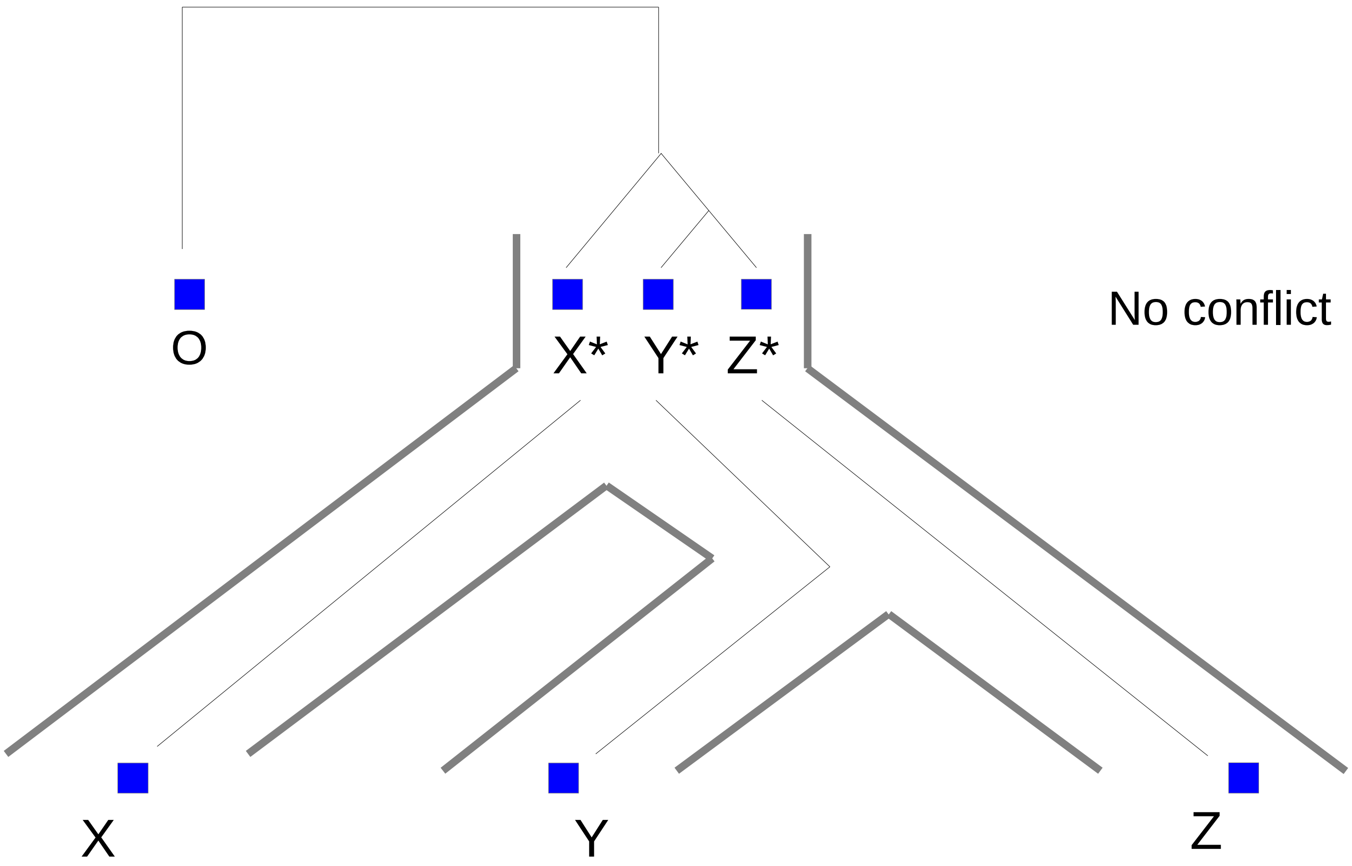
Gatesy & Springer 2013
Gatesy & Springer 2014
Springer & Gatesy 2014
Springer & Gatesy 2016

Wu, Song, Liu & Edwards 2013
Liu, Xi, Wu, Davis, Edwards 2015
Edwards et al. 2016

The prevalence of ILS cannot exceed the level of ancestral of polymorphism

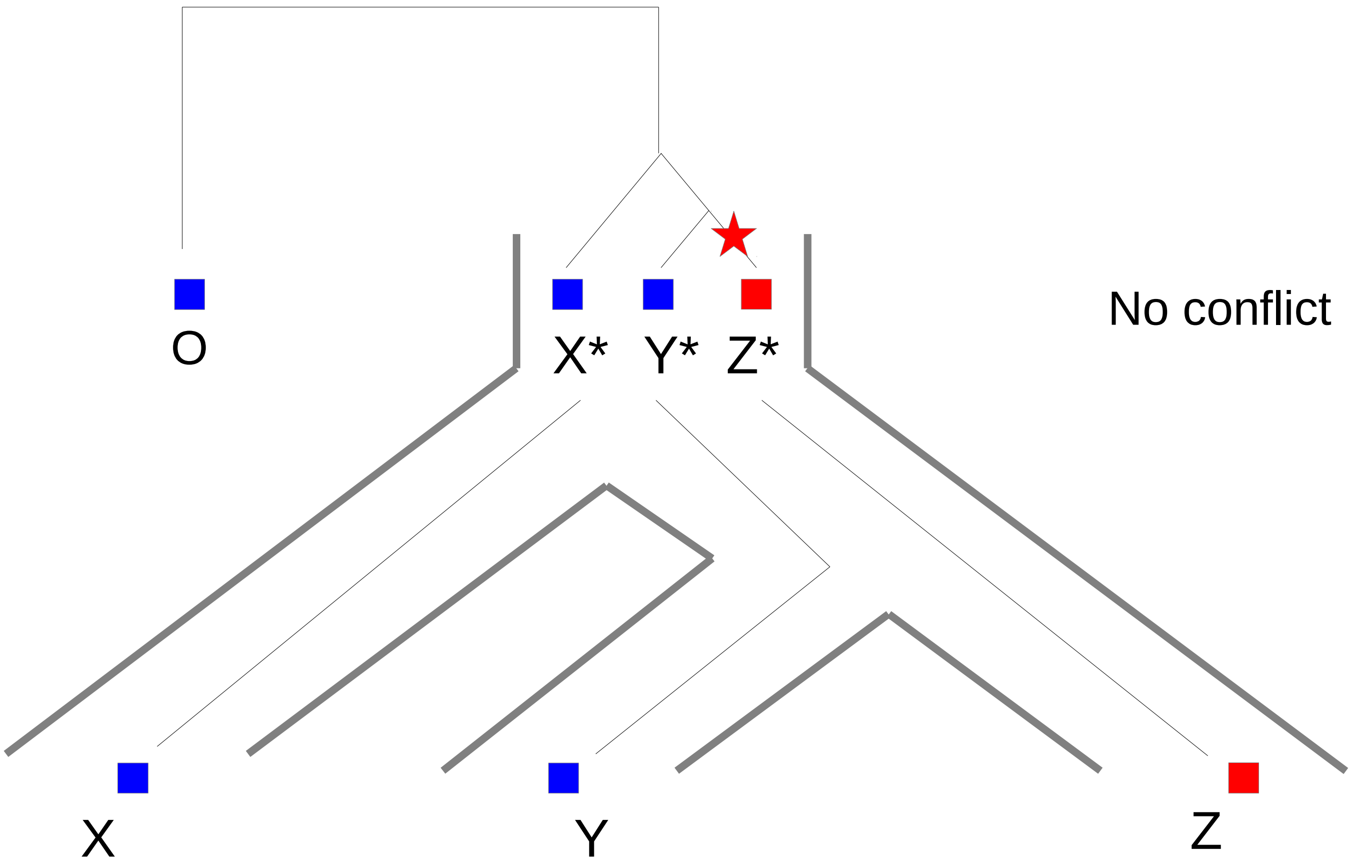


The prevalence of ILS cannot exceed the level of ancestral of polymorphism



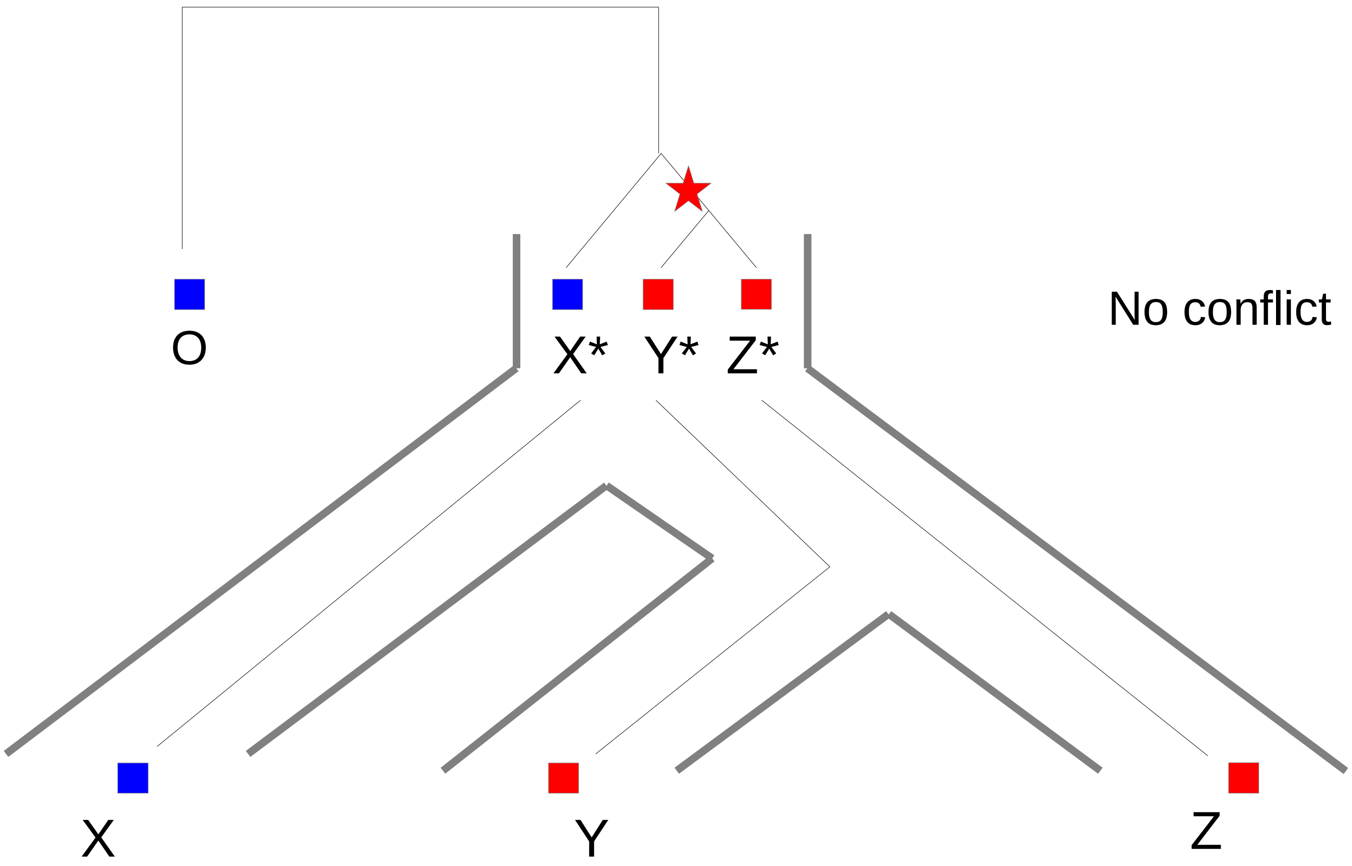
No conflict

The prevalence of ILS cannot exceed the level of ancestral of polymorphism



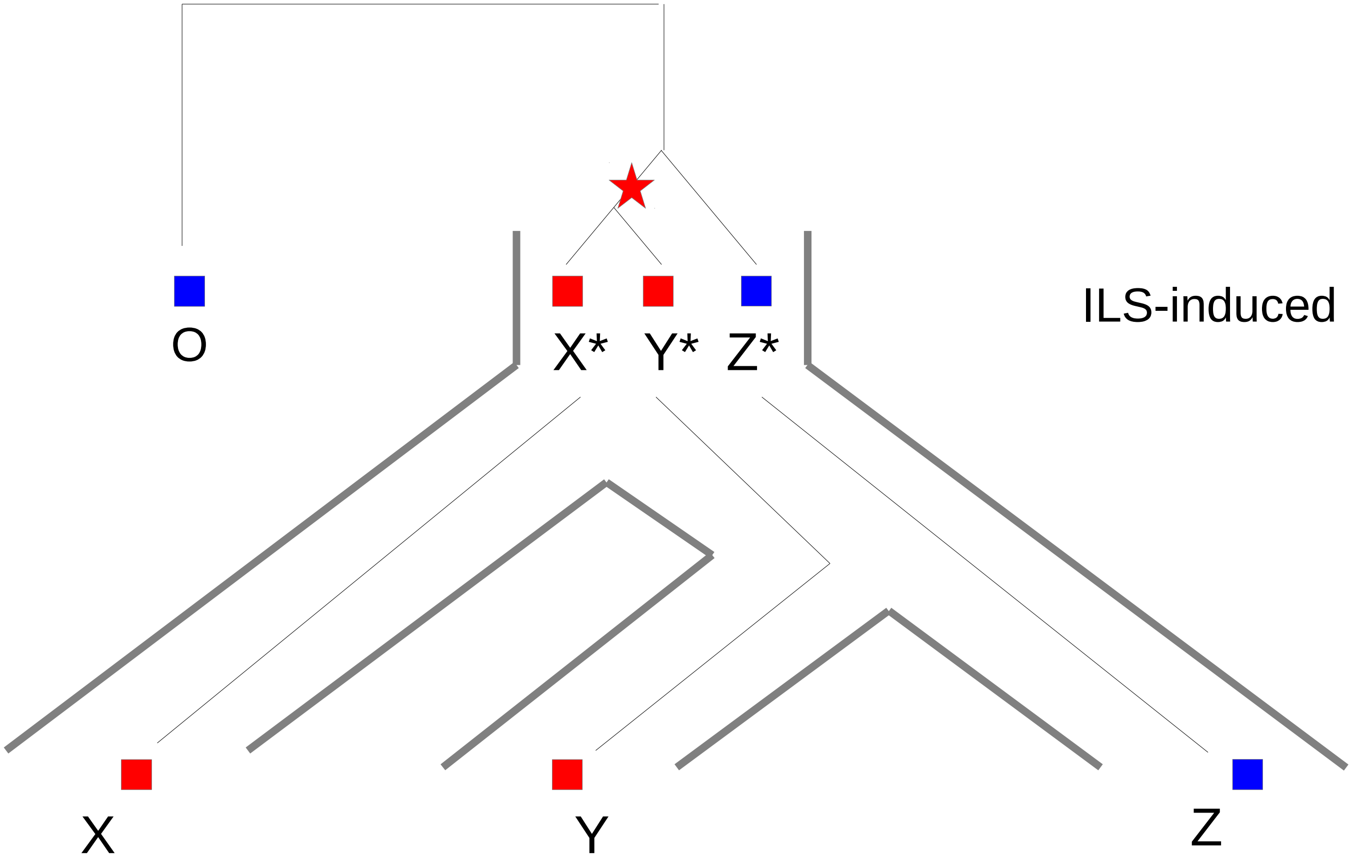
No conflict

The prevalence of ILS cannot exceed the level of ancestral of polymorphism



No conflict

The prevalence of ILS cannot exceed the level of ancestral of polymorphism

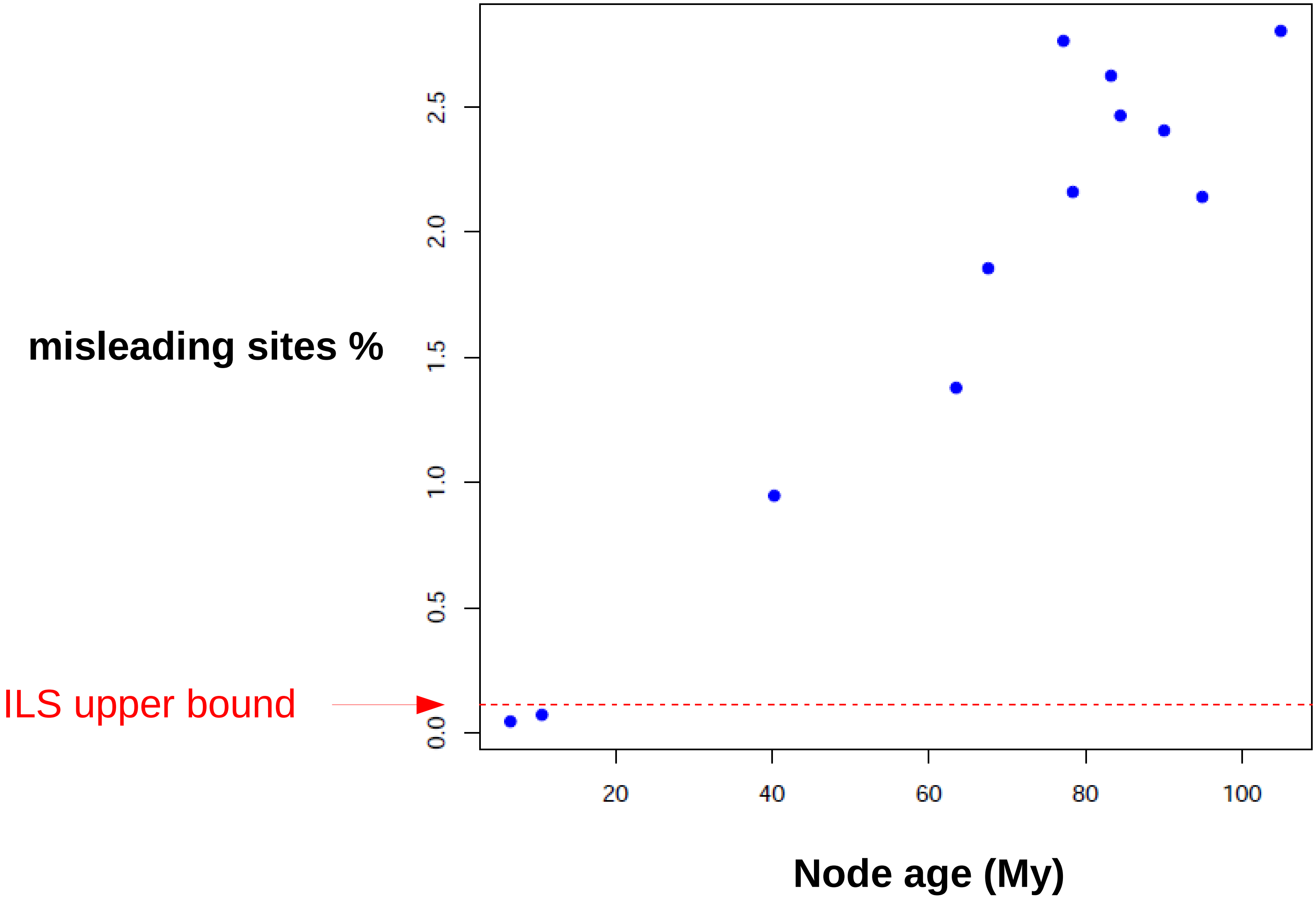


ILS-induced conflict

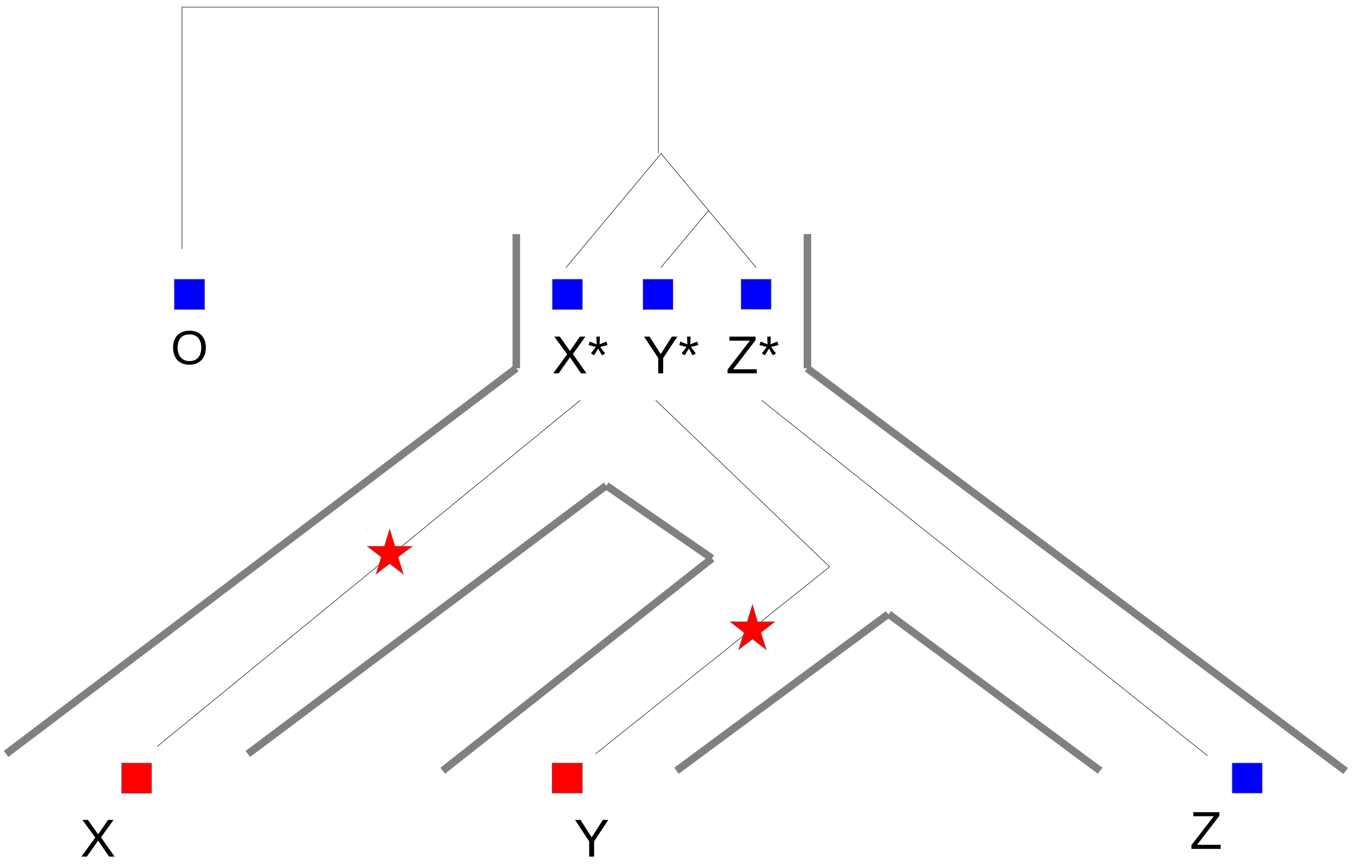
Maximal prevalence of ILS-induced conflicts = $\mathbf{4N_e\mu/3 = \theta/3}$

Estimated θ in extant mammalian exomes: 0.0003 - 0.0035

Most of the phylogenetic conflict in mammals results from processes other than ILS



Most of the phylogenetic conflict in mammals results from processes other than ILS



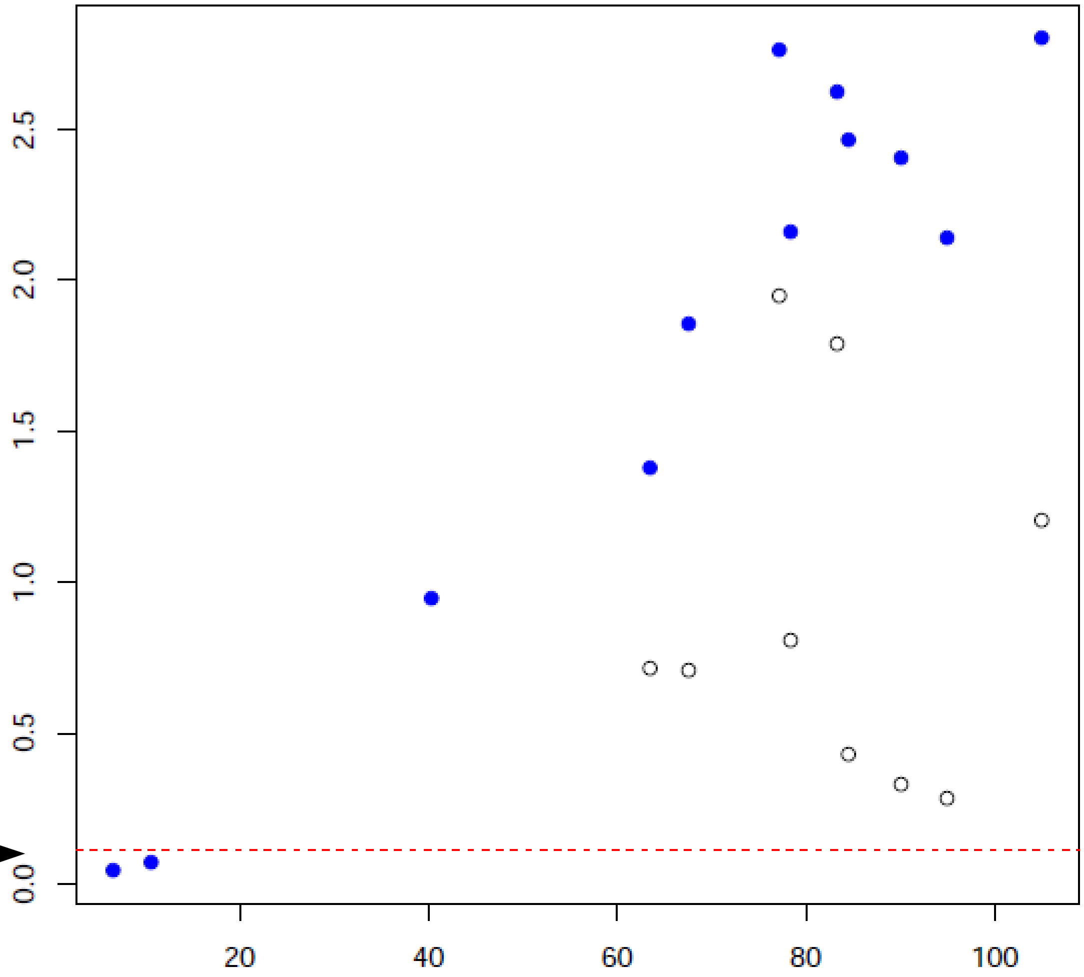
Applied to a data set of 5299 exon trees in 43 species of mammals, ILS-aware **ASTRAL** and ILS-unaware **SuperTriplets** yielded the **exact same trees**.

Conclusions

- Exons, not CDS, are plausible phylogenomic units in mammals
- Phylogenetic conflict in OrthoMaM v9 mainly results from multiple substitutions, not ILS
- ILS-aware, gene tree-based methods are not obviously superior to more classical supertree methods

Thanks to:

Most of the phylogenetic conflict in mammals results from processes other than ILS



misleading sites %

ILS upper bound

Node age (My)