Gene and species tree reconciliation
(β-globin and SERA genes)

Statistics 246 Spring 2006

Week 7 Lecture 1

Homologs

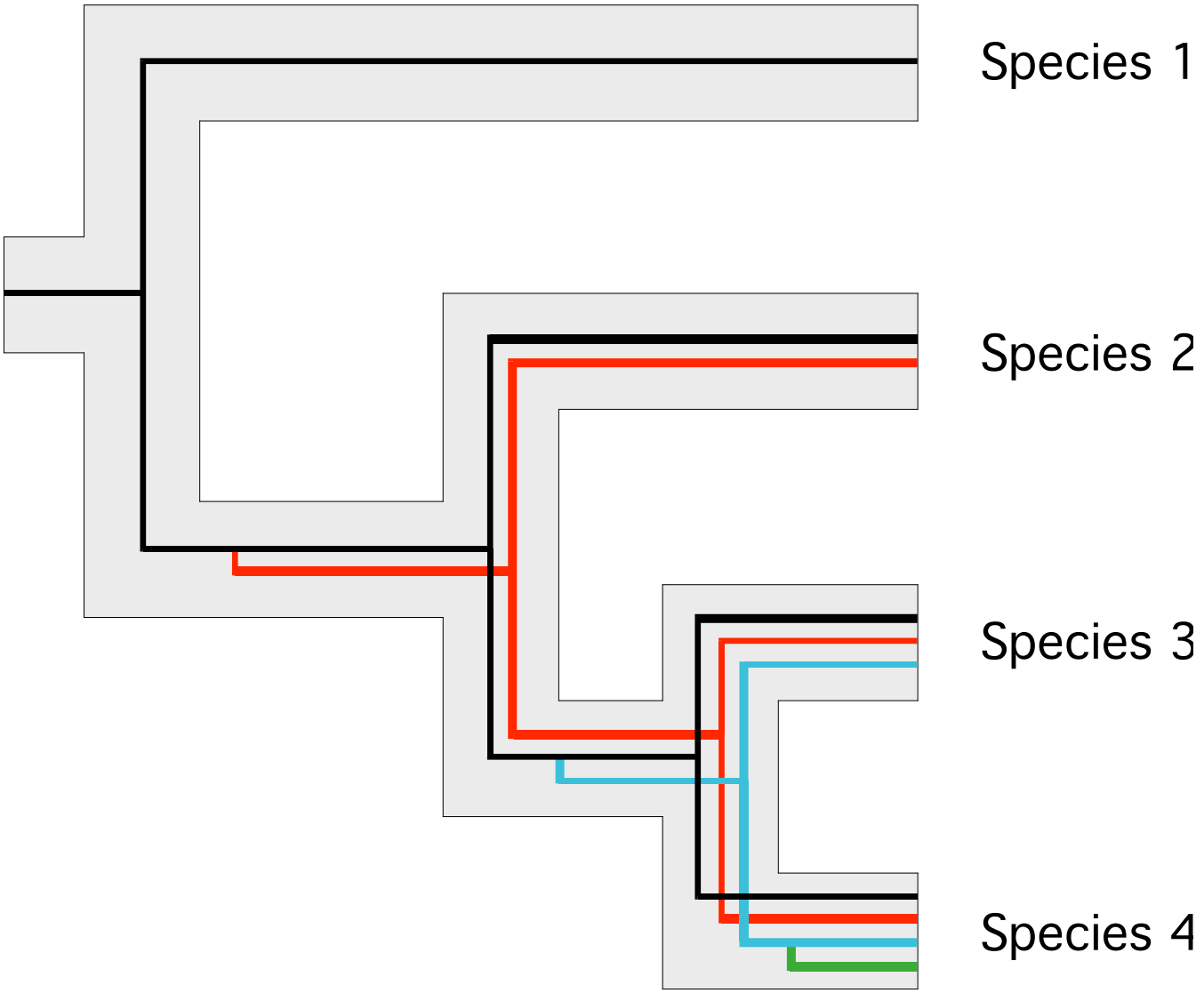
Recall that homologous genes or proteins result from

Speciation (orthologs): when separate lineages diverge from a common ancestor and experience different evolutionary pressure, or

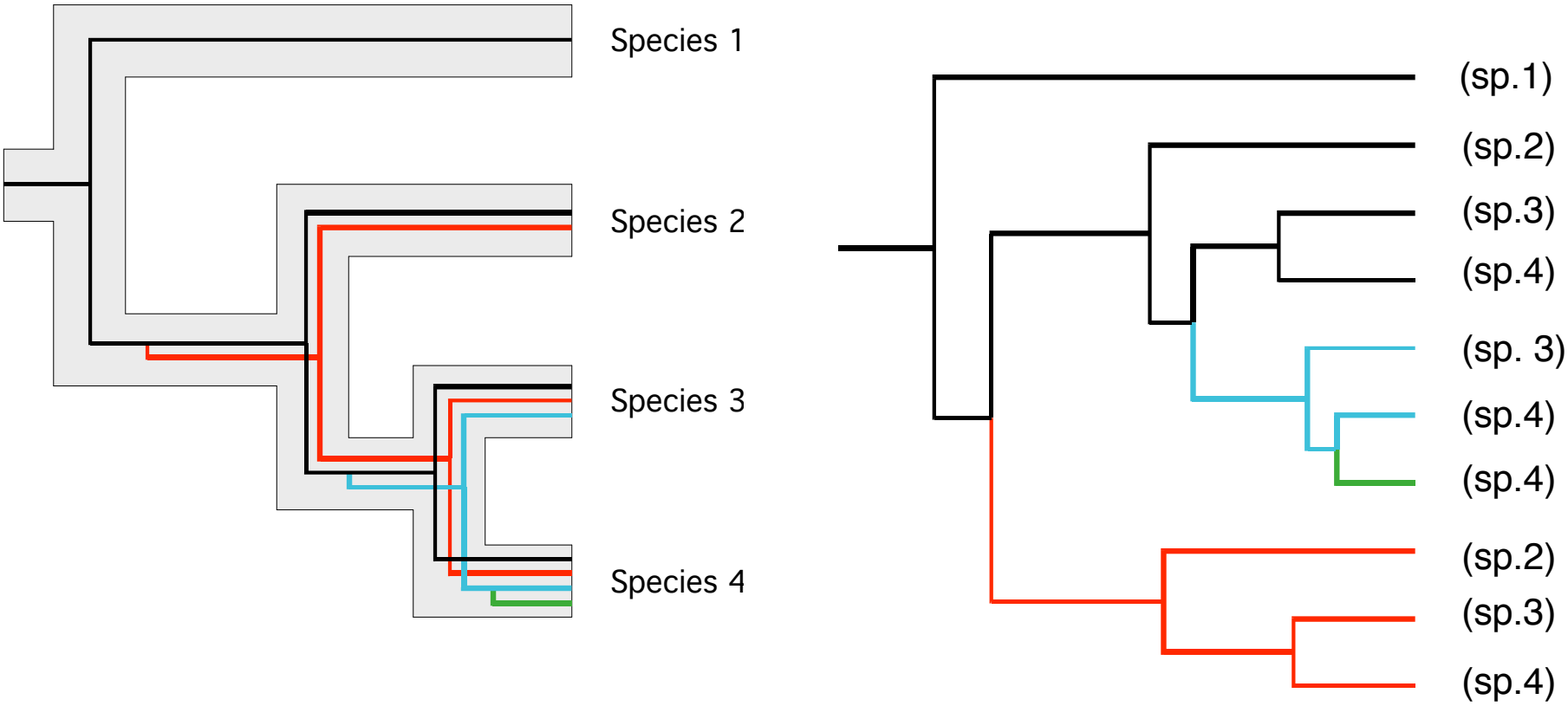
Duplication (paralogs): when part of a gene, a full gene, or a group of genes are duplicated within a species and the duplication becomes fixed in the population. Subsequent evolution of the new copy or copies may differ from that of the original, e.g. one copy may take on a new or more specialized function.

In our discussion so far, we have seen trees with *all* orthologs (β-globins) or *all* paralogs (globins generally). Now let's see them together, and consider the question of determining which is which.

How we envision it happening

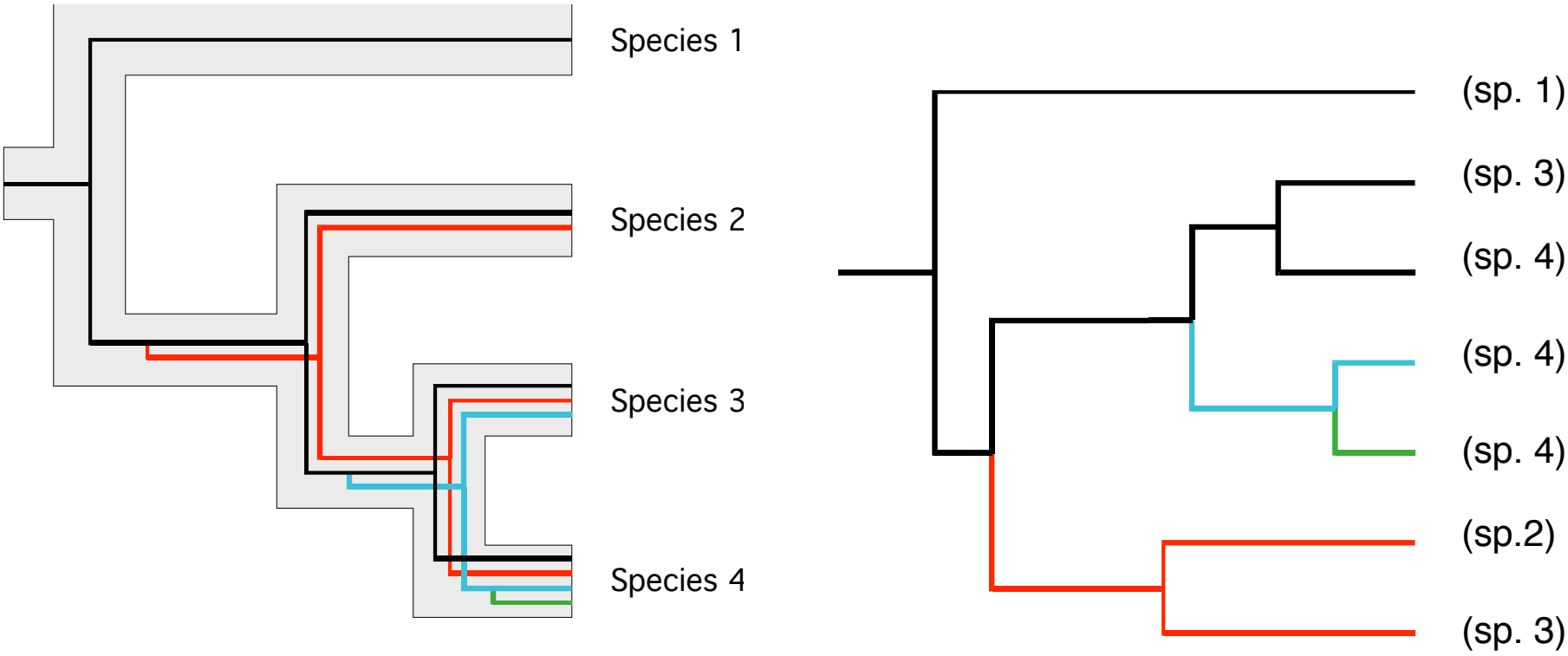


Reduction to a gene tree



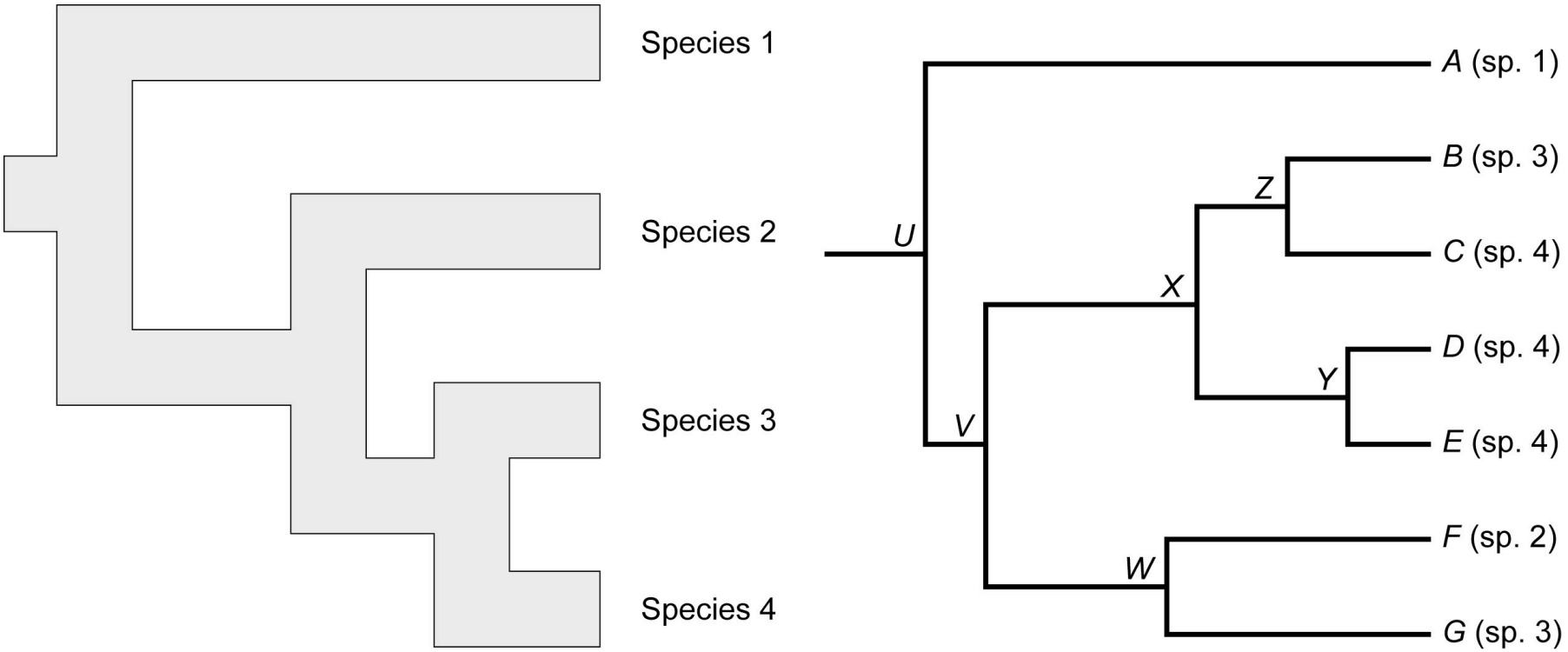
Note that when no losses occur, we have copies of different parts of the species tree within the gene tree.

Homologs: unrecognized paralogy



Suppose that certain homologs are lost or not yet found, and that sp. 2 only appears once, as above. We will think red 2 and red 3 are orthologs.

Reconciliation



A reconciliation is a map between a gene tree and a species tree with gene duplications and losses being postulated to explain any incongruence between the trees.

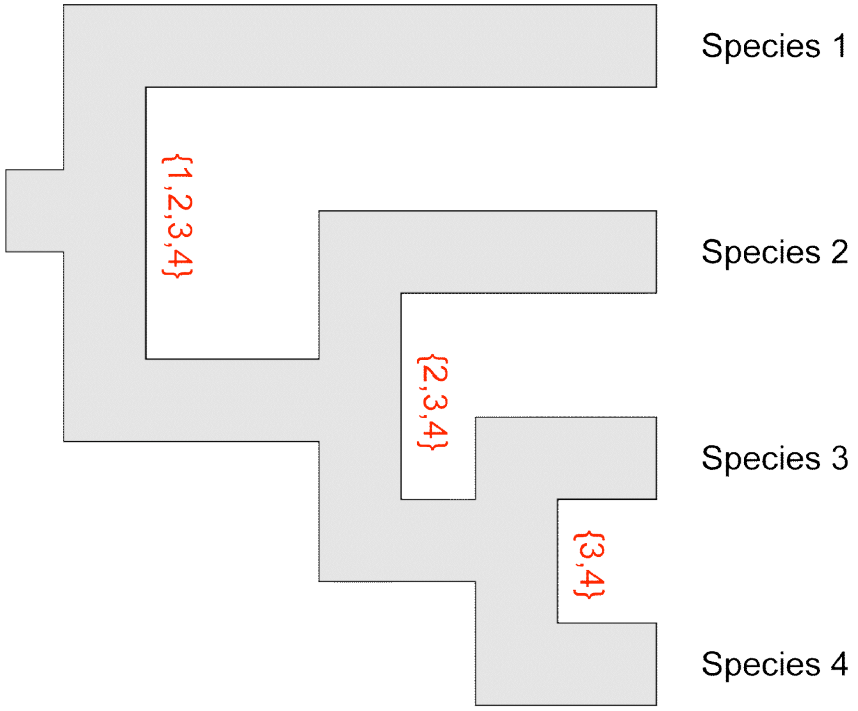
Algorithm

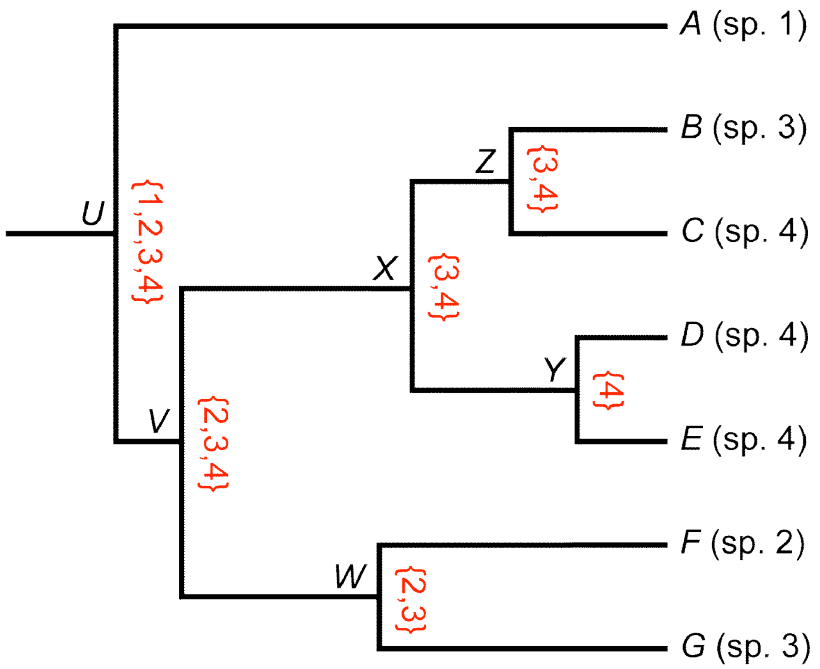
Let S and G denote the set of nodes of the species tree and gene tree. (Both trees are assumed to be rooted and binary.) For $g \in G$, define $\sigma(g)$ to be the set of species contained in the subtree that begins at node g. For $s \in S$ define $\sigma(s)$ similarly.

A *map* from G to S: for every $g \in G$, let $M(g)$ be the lowest (most recent) $s \in S$ for which $\sigma(g) \subseteq \sigma(s)$.

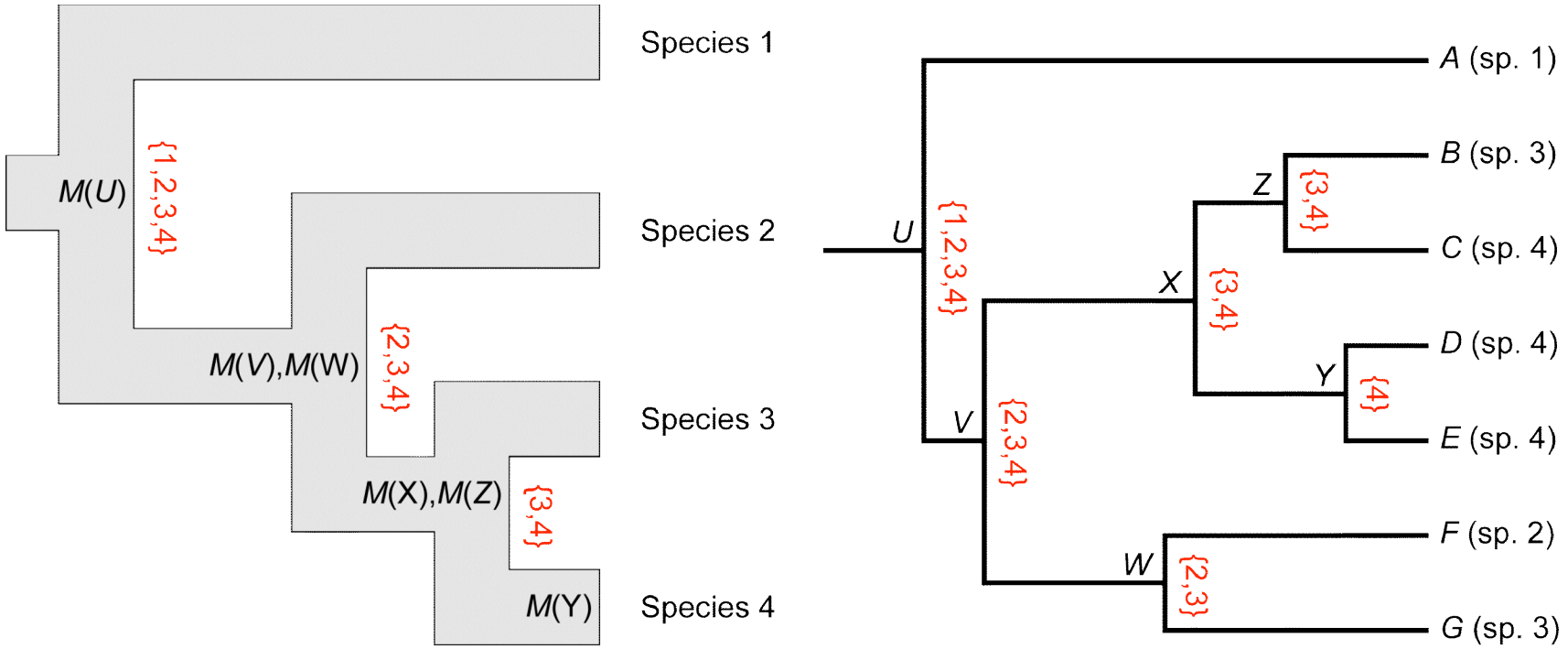
For any internal $g \in G$, with child nodes g_1 and g_2, we infer that g represents a duplication event if and only if $M(g)$ is equal to either $M(g_1)$ or $M(g_2)$ i.e. if the node g maps to the same position in the species tree as one (or both) of its children.

$\{\sigma(s):s \in S\}$ and $\{\sigma(g):g \in G\}$

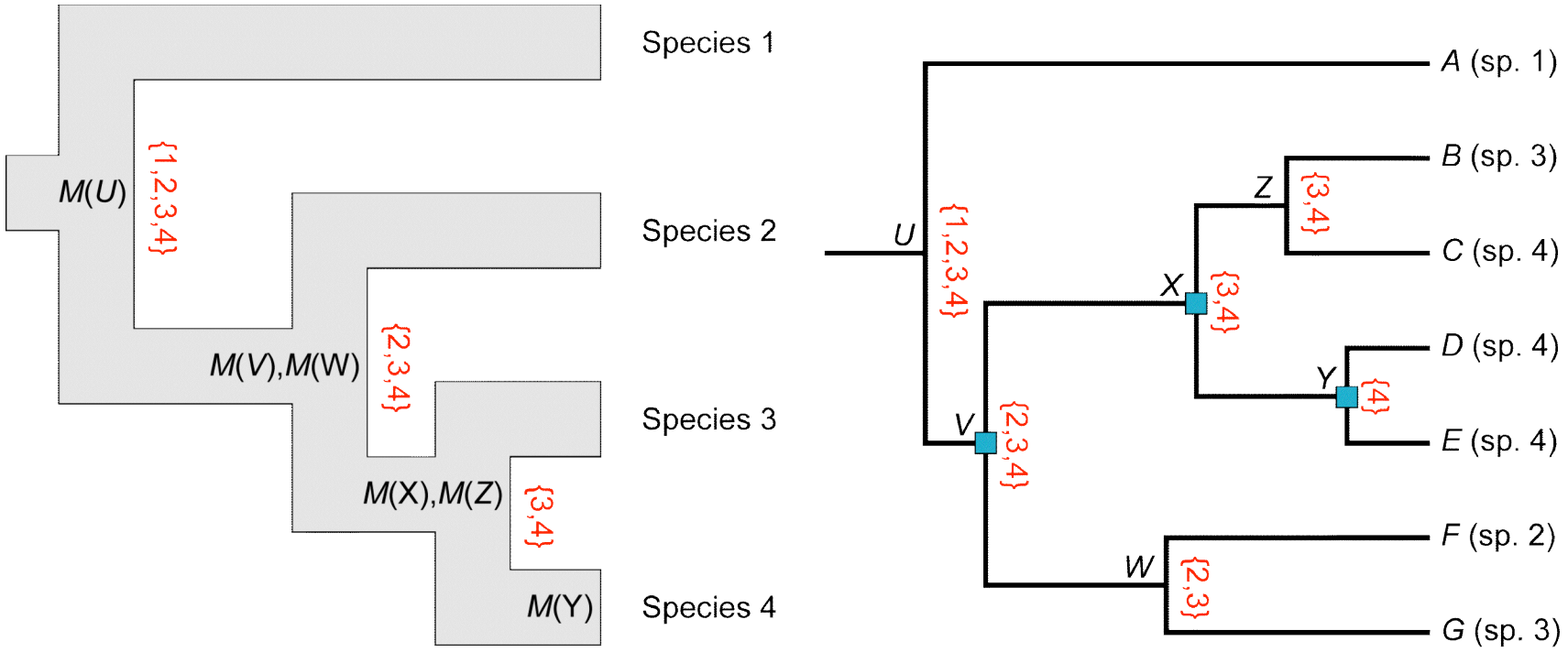




Reconciliation mapping *M*:*G* → *S*

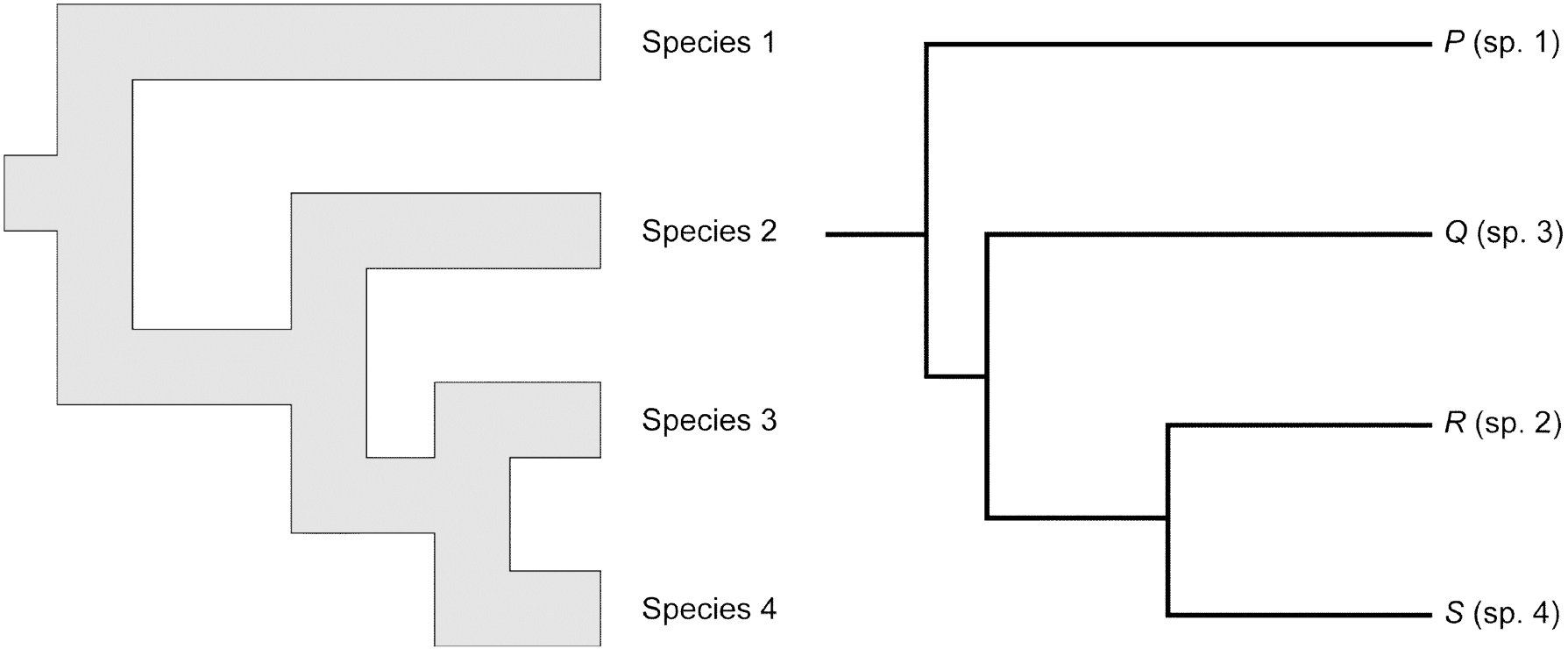


Inferred duplications (boxes)



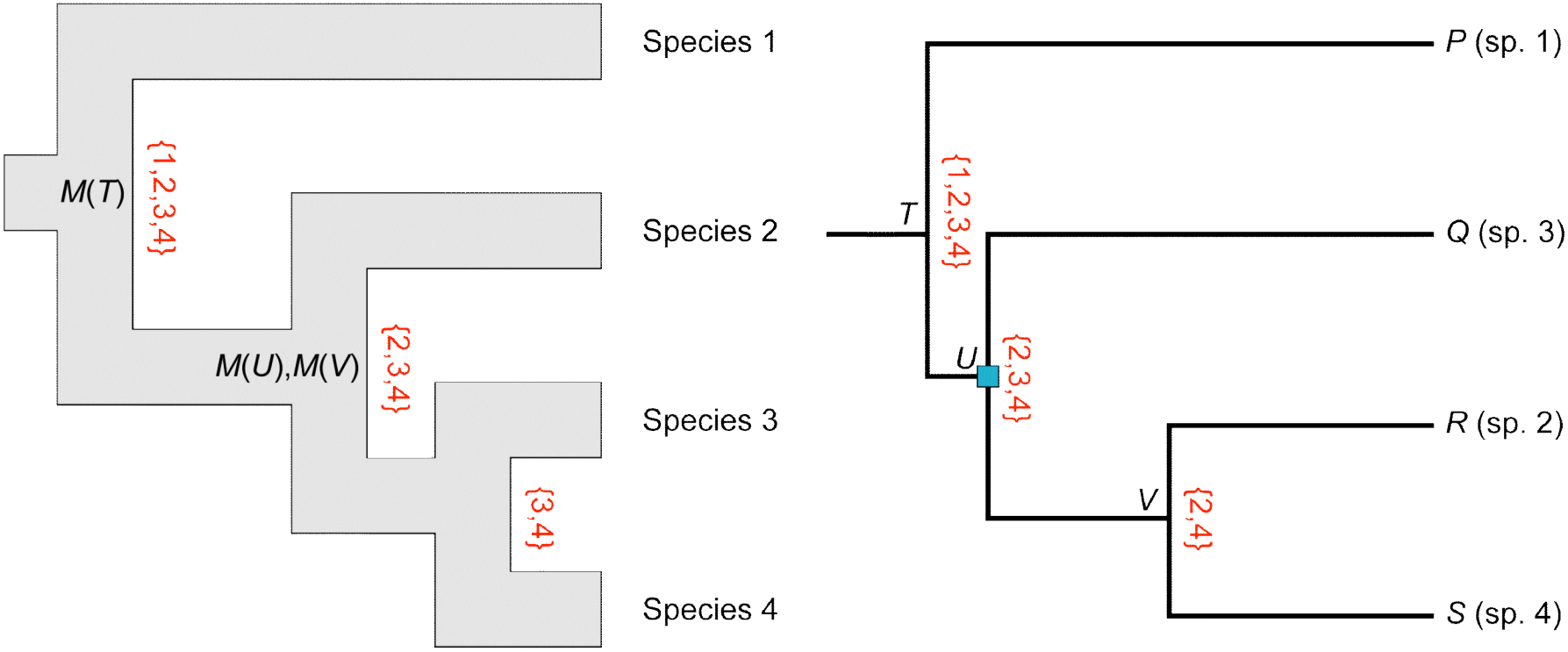
We can now go on to infer genes lost or not yet found, completing the part of the species tree remaining at each duplication, see next slides.

Inconsistencies between *S* and *G*

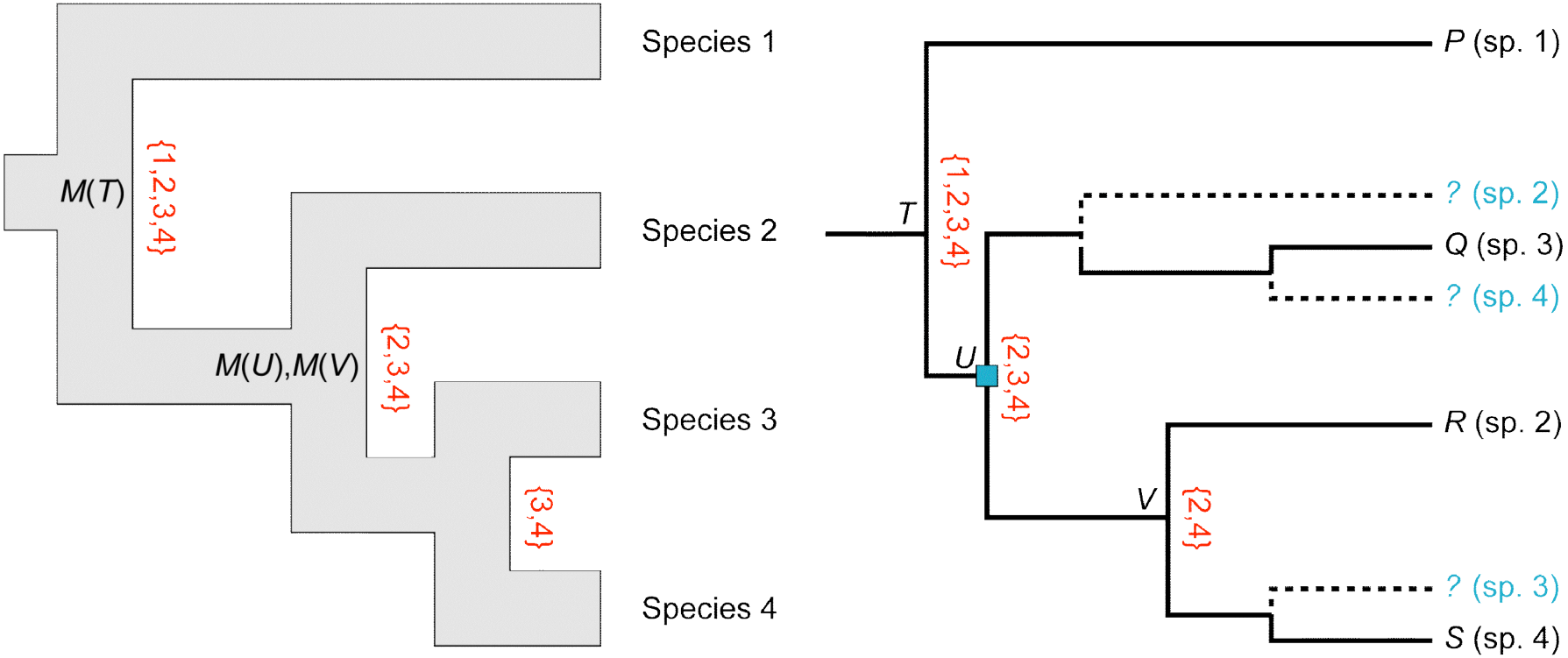


can be resolved in the same way.

Inconsistencies between *S* and *G*



Inconsistencies between *S* and *G*



Inserting lost or not yet found genes reconciles the two trees.

Vertebrate β-like globin genes (with a focus on mammals)

The human β-like globin cluster has 5 active globin genes: embryonic ε-globin, two fetal γ-globins, the adult δ-globin and the abundant adult β-globin.

Chickens also have a β-like cluster, with ρ-globin and ε-globin at the ends, and two β-globin in between, one β expressed at hatching and the other in adulthood.

Given that the avian ρ-β-ε and the eutherian ε-γ-δ-β gene clusters were the only known clusters in these taxa, it was natural to suppose that they were orthologous. That turned out to be wrong!

Vertebrate β-like globins: cont.

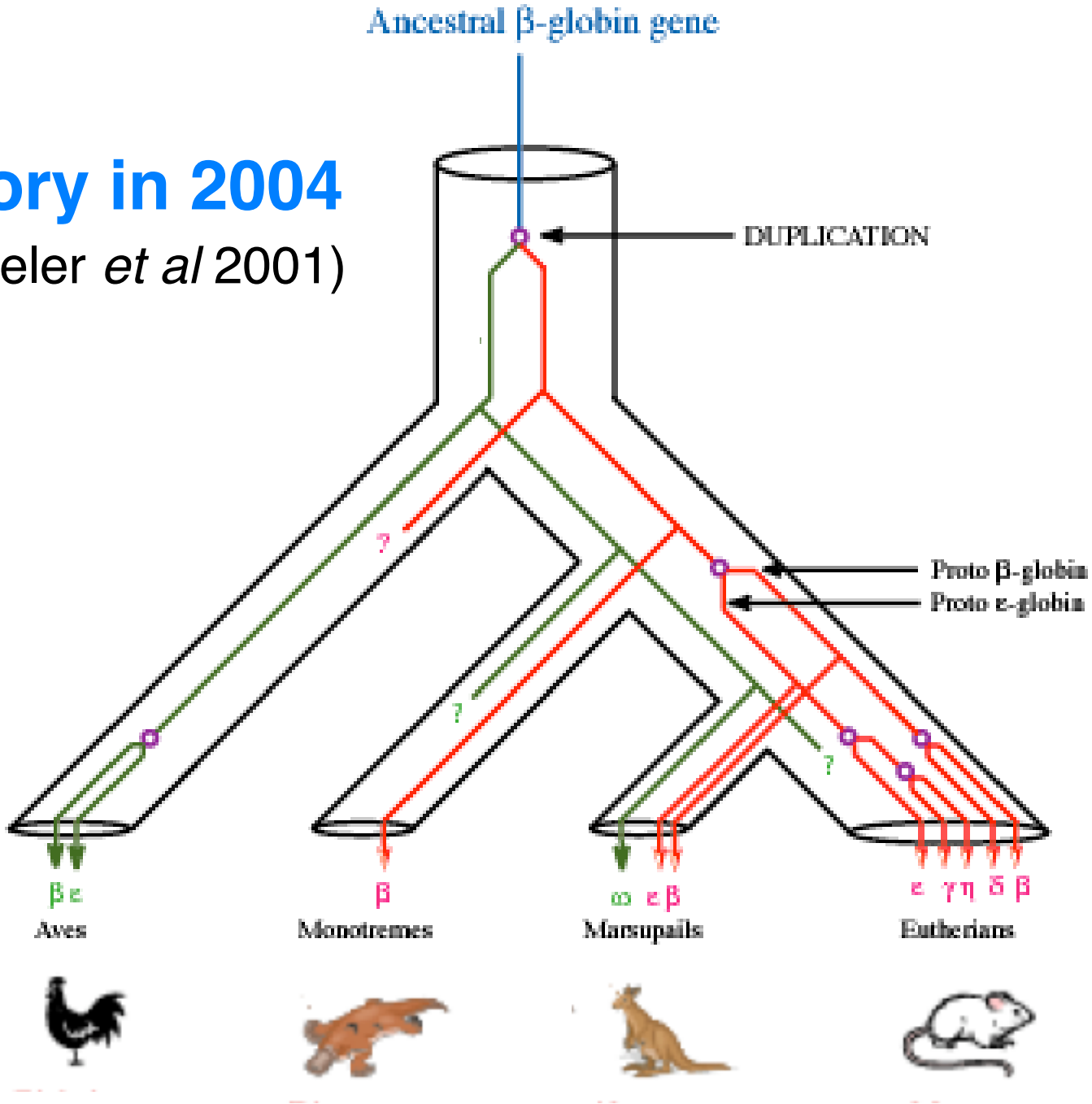
In 2001 the marsupial ω-globin was discovered. This gene was a component of a novel haemoglobin found in the blood of neonatal tammar wallabies, expressed just before and after the birth of the joey.

The figure which follows describes the view that the ω is orthologous to the avian ρ-β-ε, and that its orthologue in eutherian mammals has been lost. The slides after that summarize the evidence for this conclusion.

We present the picture from 2001-2004, and then present a discovery from 2005.

The story in 2004

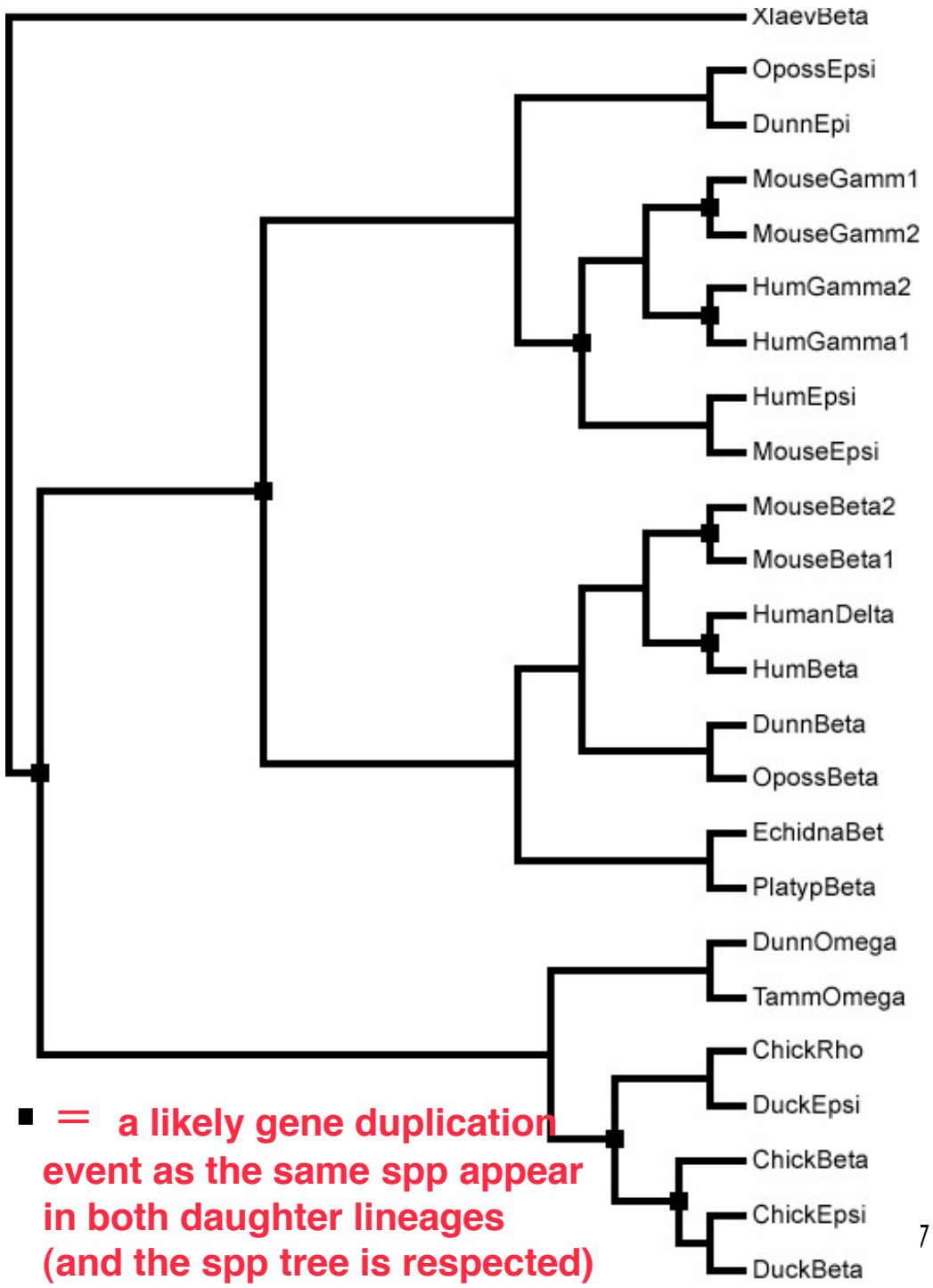
(after Wheeler *et al* 2001)



(Edited) gene tree for many β-globins

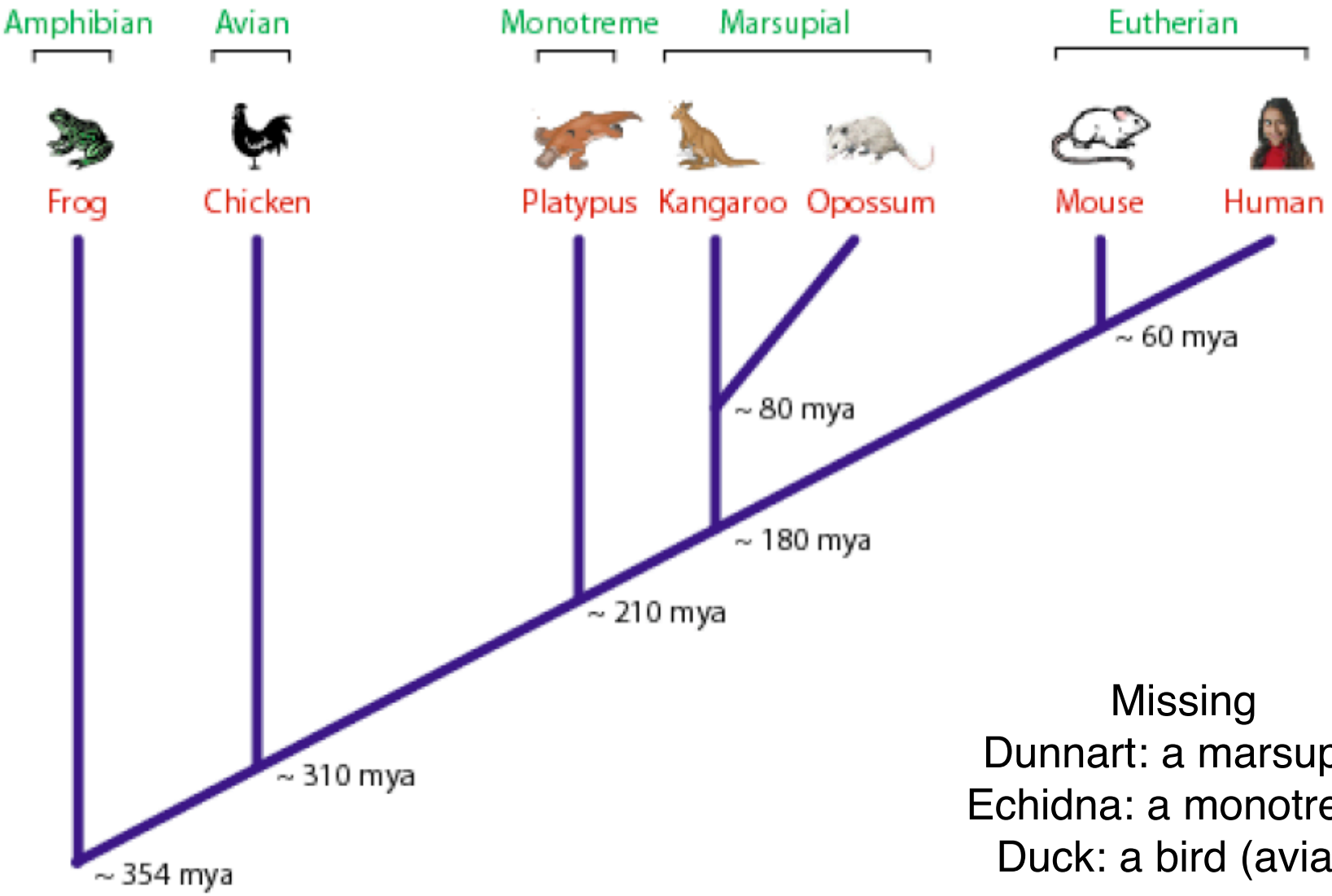
Species

Xlaev = a frog (outgroup)
Oposs = grey short-tailed opossum
Dunn = fat-tailed dunnart
Mouse
Hum = human
Echidna
Platyp = platypus
Tamm = tammar wallaby
Chick = chicken
Duck



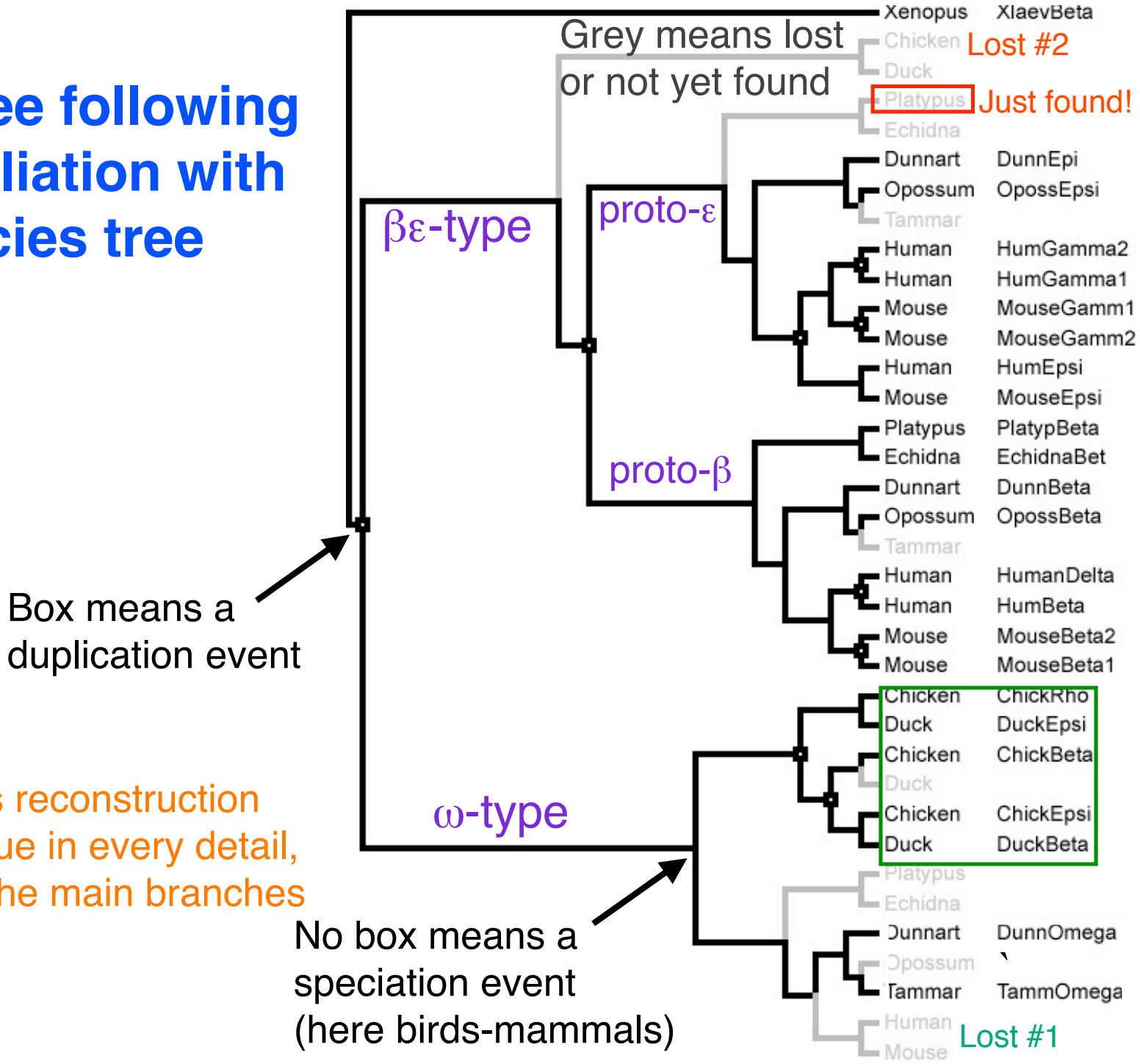
■ = a likely gene duplication event as the same spp appear in both daughter lineages (and the spp tree is respected)

Species tree used



Missing
Dunnart: a marsupial
Echidna: a monotreme
Duck: a bird (avian)

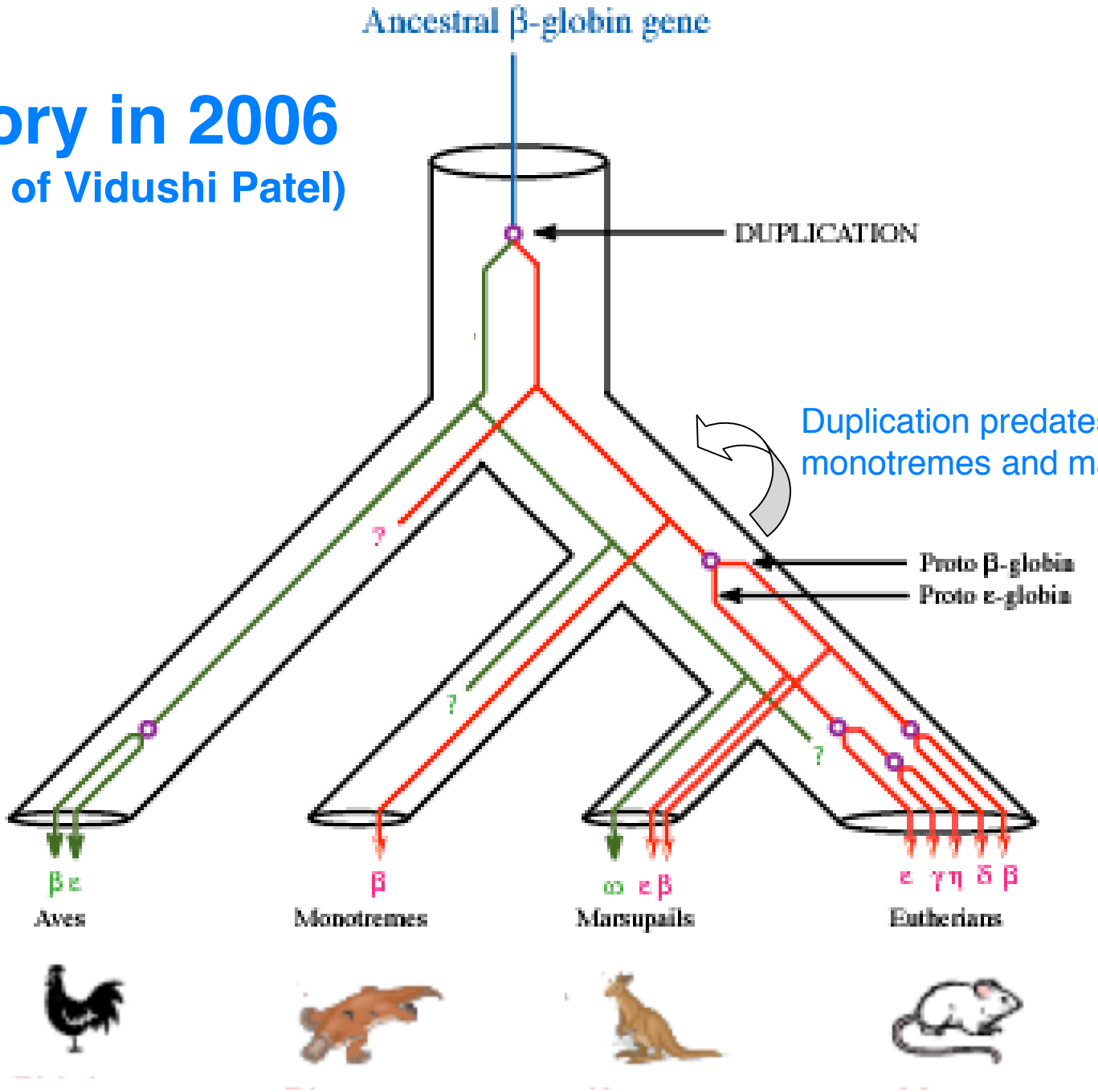
Gene tree following reconciliation with species tree



Warning: this reconstruction may not be true in every detail, but we think the main branches are correct.

The story in 2006

(after work of Vidushi Patel)



Acknowledgements

Tracey Wilkinson, Florey/WEHI

Vidushi Patel, ANU

Toby Sargeant, WEHI

Richard Bourgon, UCB

Selected references

Molecular evolution

M. Nei (1987) **Molecular evolutionary genetics.** Col.UP

W-H Li (1997) **Molecular evolution.** Sinauer

R.D.M. Page (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. **Bioinformatics** 14 819-820.

J. Felsenstein (2004) **Inferring phylogenies**. Sinauer

Globins in general

RE Dickerson and I Geis (1983) **Hemoglobin: Structure, Function, Evolution, and Pathology**. Benjamin/Cummings.

R Hardison (1998) Hemoglobins from bacteria to man: Evolution of different patterns of gene expression. **J Exp Biol** 201:1099-1117

Recent work on globins, see

Wheeler *et al,* **J Molecular Evolution** (2004) 58:642-52

Cooper *et al*, **J Molecular Evolution** (2005) 60:653-64…..and references therein.

Evolution of the SERA gene family in *Plasmodium*

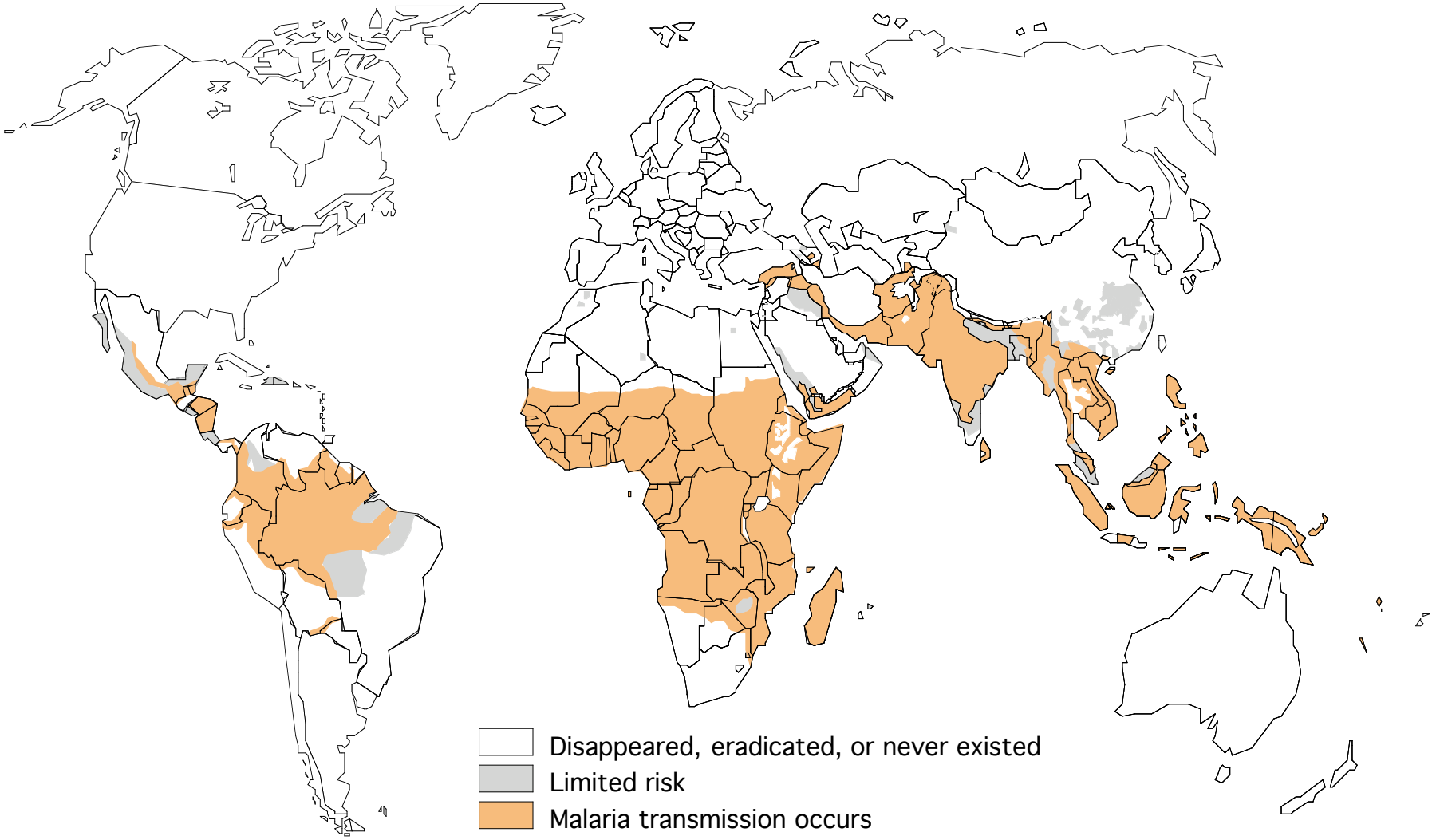
Overview

- Introduction to malaria, *Plasmodium*, and SERA
- Gene tree inference
- GC content
- Homology: paralogs vs. orthologs
- Reconciling gene and species trees
- Modifications suggested by reconciliation
- Predictions

Malaria

- Approximately 40% of the world's population is at risk of malaria. It is found throughout the tropical and sub-tropical regions of the world.
- Malaria causes more than 300 million acute illnesses and at least one million deaths annually. Ninety per cent of deaths due to malaria occur in Africa, south of the Sahara — mostly among young children.
- The disease was once more widespread but it was successfully eliminated from many countries with temperate climates during the mid 20th century.

Malaria's scope, 1999



The *Plasmodium* parasite

Four species of the *Plasmodium* parasite are responsible for malaria in humans:

- *P. vivax*
- *P. malariae*
- *P. ovale*
- *P. falciparum*.

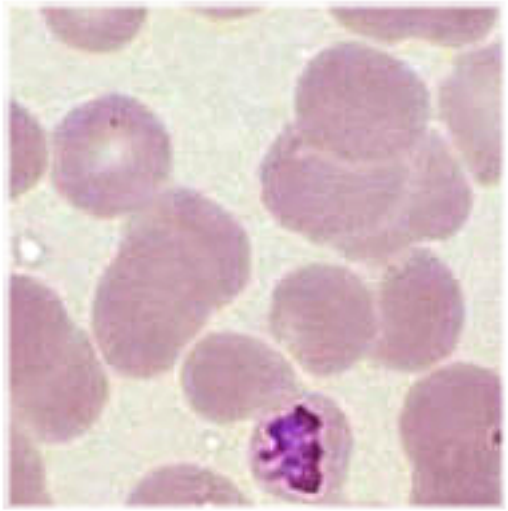
P. vivax and *P. falciparum* are the most common, and *P. falciparum* causes the most deadly type of malaria infection.

Plasmodium and the *Anopheles* mosquito

The *Plasmodium* parasite enters the human host from the saliva of an infected *Anopheles* mosquito.

Once in the human bloodstream, it undergoes a series of changes as part of its complex life-cycle.

The various stages allow the plasmodia to evade the immune system, infect the liver and red blood cells, and finally develop into a form able to infect another mosquito when it bites the infected person.



P. ovale in a red blood cell

Attempts at malaria control

Efforts at malaria control have run into two major problems:

- The plasmodium parasites have become resistant to one drug after another.
- Many insecticides that were once effective in controlling the mosquito vector are no longer useful against it.

Years of vaccine research have produced few hopeful candidates; an effective vaccine is at best years away.

The SERA5 gene in *P. falciparum*

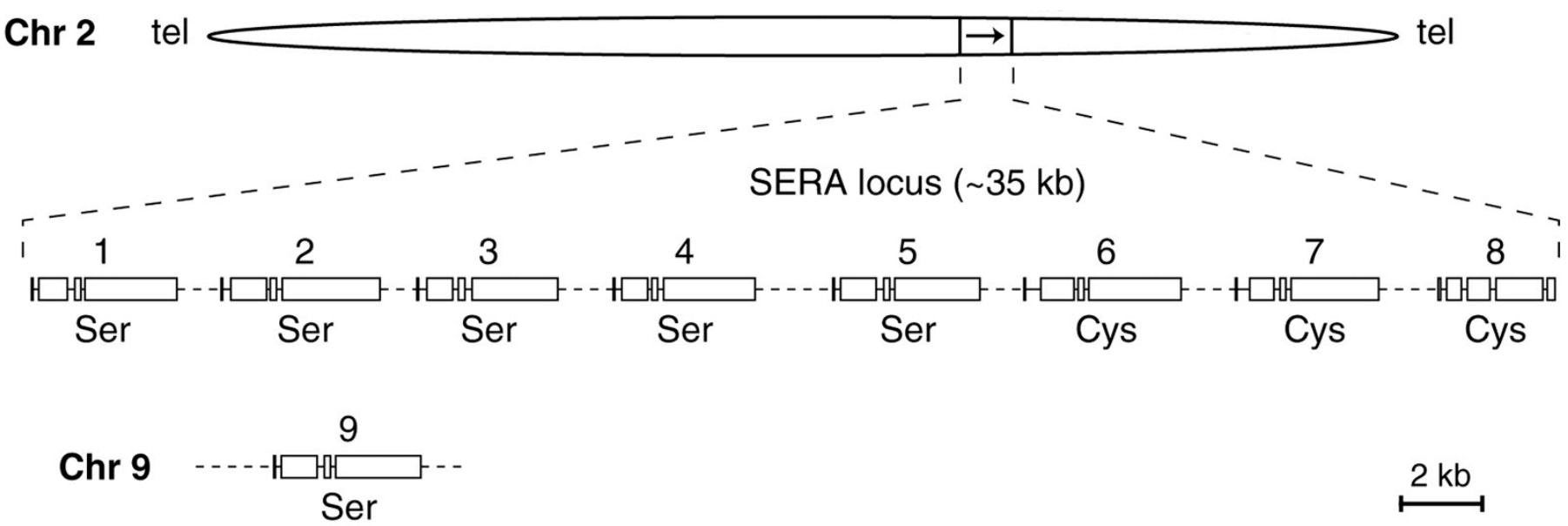
SK Miller, et al. 2002. A subset of *Plasmodium falciparum* SERA genes are expressed and appear to play an important role in the erythrocytic cycle. *Journal of Biological Chemistry* 277(49):47524-47532.

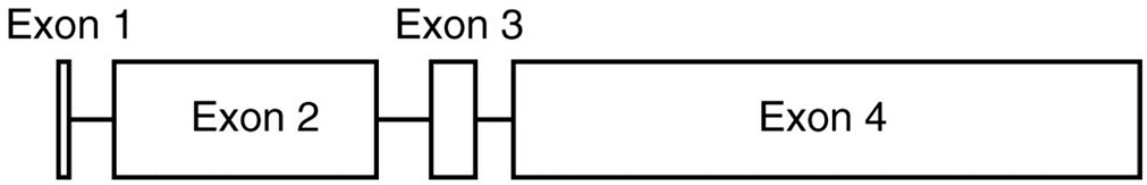
"The *P. falciparum* serine repeat antigen (SERA5) has shown considerable promise as a blood stage vaccine… Whereas the biological role remains unknown, the protein possesses a papain-like protease domain that may provide an attractive target for therapeutic intervention."

The SERA gene family in *P. falciparum*

- SERA5 is the fifth gene in a cluster of eight SERA homologs present on chromosome 2 of *P. falciparum*. A ninth homolog is found on chromosome 9.
- Recombinant proteins expressing the N-terminal SERA5 domain are highly immunogenic and elicit antibodies that inhibit erythrocyte invasion and parasite replication *in vitro*. *In vivo* studies using rodents and primates have shown that this domain confers significant protection from parasite challenge.

The SERA gene family in *P. falciparum*





The {Ser,Cys} and {His, Leu, Met} sites

All SERA proteins possess a central domain that shows homology to the papain family of cysteine proteases. Some exhibit an unusual cysteine-to-serine substitution at the active site cysteine residue.

There is another functionally important site in these genes, principally occupied by histidine. However, two further subsets of the serine family also feature a second active site mutation: histidine to methionine in one case, and to leucine in the other.

Thus we have four subfamilies of SERA proteins: Cys-His, Ser-His, Ser-Met and Ser-Leu.

The SERA gene family

- BLASTP and TBLASTN searches against Plasmodium DNA databases (Genbank, NCBI, PlasmoDB, Sanger Centre, TIGR) turned up presumptive SERA homologs in other *Plasmodium* species.
- The presence of relatively large SERA multi-gene families in divergent *Plasmodium* species suggests an important biological role for the proteins encoded by these genes.
- SERA 8 was left out as it appeared to be a pseudo-gene.

P. vivax (human)	v1, v2, v3, v4, v5a, v5b, v6
P. knowlesi (rhesus monkey)	k1, k2, k3, k4, k5
P. falciparum (human)	f1, f2, f3, f4 f5, f6, f7, f9
P. reichenowi (chimpanzee)	r2, r4, r5, r7, r9
P. yoelli (rodent)	y1, y2, y3, y4
P. chabaudi (rodent)	c1
P.vinckei (rodent)	vi1, vi2, vi3

Gene tree inference: overview

1. Align SERA sequences and extract a region or regions for which the alignment is unambiguous and requires few gaps.
2. Compute pairwise distance estimates (expected number of substitutions per site) using an appropriate substitution model.
3. Use Neighbor Joining to construct the inferred gene tree (i.e., its topology and estimated edge lengths).
4. Resample columns of the alignment and repeat steps 1 to 3 to obtain “bootstrap” confidence scores for configuration around tree edges.

Alignment

- Amino acid sequence data was obtained for 33 (putative) genes of the SERA family, representing the 7 different species of *Plasmodium*.
- Initial CLUSTALW alignment revealed highly variable regions, five moderately conserved blocks, and one well-conserved protease domain.
- Approximately 266 amino acids (up to 6 gaps per sequence) from the readily alignable protease domain were used for inference.
- Pair-wise similarity for this subset of the sequence data ranged from 49% to 99%, with an average of 61%.

Pair-wise distance matrix

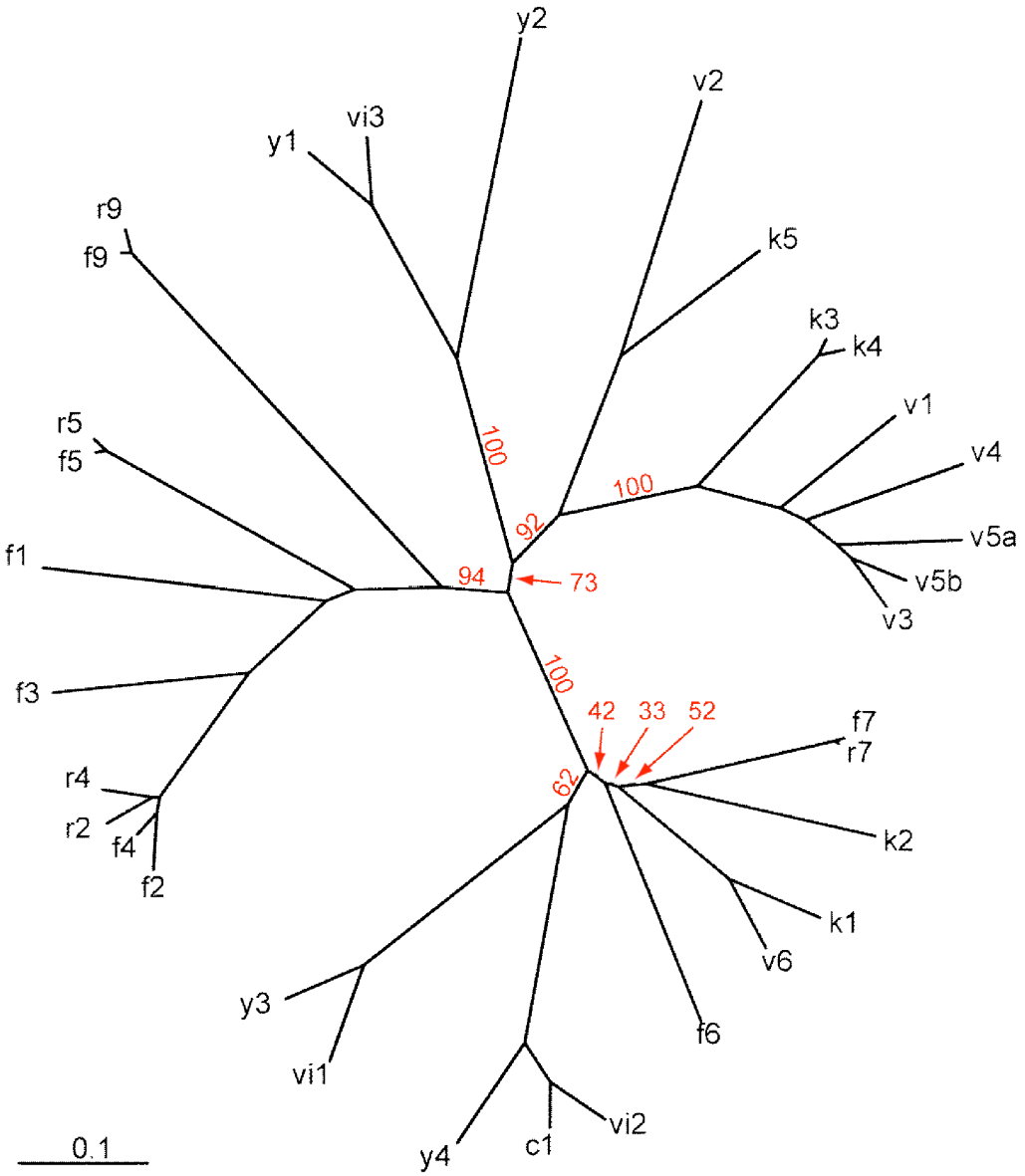
A MLE under the Jones, Taylor, and Thornton (JTT) substitution model for amino acids was used. The JTT model is similar to that underlying PAM matrices

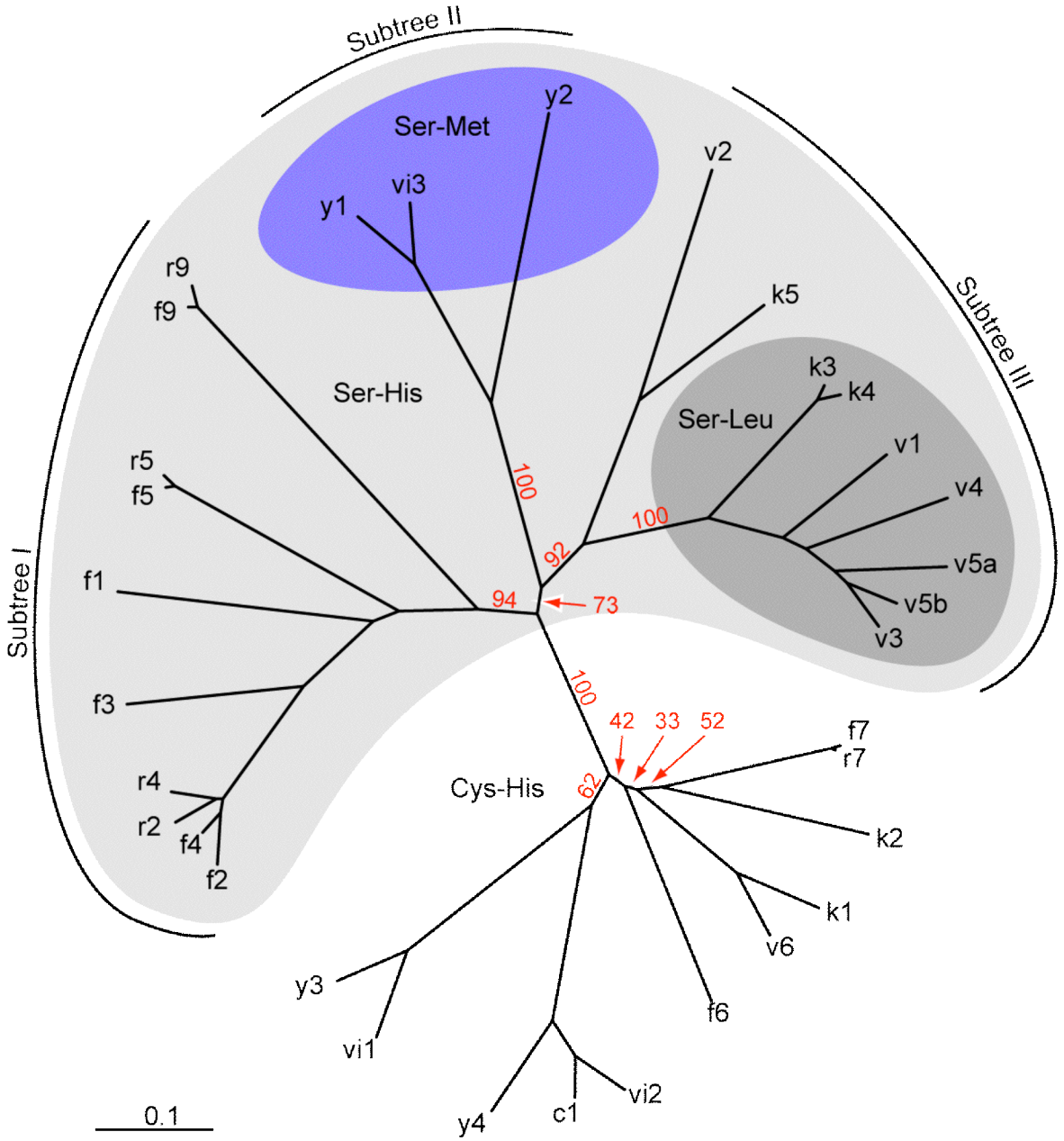
- …is based on a substitution matrix obtained empirically from a large number of closely related proteins.
- …uses evolutionary rather than chronological "time."
- …takes the chemical and physical properties of amino acid residues into account.
- …corrects for multiple substitutions at the same site.

Pairwise distance matrix

	v3	v5b	v5a	v4	v1	k3	k4	v2	k5	f2	f4	f3	f1	f5	f9	k1	v6	f7	k2	f6	vi1	y3	vi2	c1	y4	y1	vi3	y2	r2	r4	r5	r7	r9
v3	0																																
v5b	.093	0																															
v5a	.185	.130	0																														
v4	.227	.240	.245	0																													
v1	.214	.253	.279	.256	0																												
k3	.284	.304	.320	.407	.345	0																											
k4	.283	.302	.312	.392	.347	.033	0																										
v2	.552	.580	.615	.605	.545	.633	.655	0																									
k5	.555	.556	.581	.620	.562	.555	.560	.342	0																								
f2	.709	.698	.750	.760	.763	.712	.708	.836	.721	0																							
f4	.684	.673	.751	.717	.738	.660	.662	.818	.718	.067	0																						
f3	.655	.669	.679	.728	.676	.619	.622	.771	.687	.334	.323	0																					
f1	.689	.683	.740	.702	.705	.651	.663	.829	.722	.444	.420	.498	0																				
f5	.705	.683	.726	.753	.716	.687	.679	.770	.696	.494	.496	.488	.480	0																			
f9	.774	.815	.871	.833	.773	.770	.772	.814	.765	.687	.668	.687	.692	.666	0																		
k1	.738	.754	.792	.740	.725	.701	.713	.764	.666	.812	.759	.785	.735	.689	.857	0																	
v6	.734	.758	.798	.744	.717	.680	.692	.732	.711	.758	.719	.727	.685	.704	.803	.138	0																
f7	.746	.736	.785	.757	.702	.659	.675	.742	.688	.711	.704	.710	.614	.674	.714	.391	.351	0															
k2	.787	.791	.817	.787	.769	.771	.788	.762	.670	.828	.827	.811	.798	.767	.793	.352	.397	.339	0														
f6	.795	.771	.803	.831	.744	.745	.760	.794	.706	.779	.748	.806	.757	.722	.756	.399	.367	.395	.424	0													
vi1	.810	.806	.830	.824	.810	.724	.748	.875	.766	.825	.808	.855	.816	.826	.870	.537	.493	.549	.569	.493	0												
y3	.791	.782	.836	.830	.787	.700	.724	.891	.746	.810	.780	.835	.817	.831	.838	.525	.497	.537	.522	.500	.145	0											
vi2	.907	.909	.907	.925	.838	.823	.843	.902	.881	.898	.900	.866	.861	.752	.867	.564	.522	.498	.478	.513	.564	.577	0										
c1	.864	.856	.860	.875	.810	.802	.816	.898	.871	.894	.900	.823	.861	.747	.869	.549	.495	.477	.482	.504	.548	.549	.087	0									
y4	.896	.898	.909	.916	.825	.823	.837	.903	.880	.912	.924	.890	.860	.780	.878	.569	.531	.483	.512	.573	.532	.554	.168	.180	0								
y1	.691	.665	.688	.708	.722	.644	.642	.765	.694	.780	.763	.691	.729	.745	.821	.733	.710	.745	.764	.747	.763	.742	.894	.854	.863	0							
vi3	.649	.620	.653	.677	.685	.613	.616	.757	.687	.763	.749	.701	.704	.786	.812	.712	.706	.729	.753	.747	.797	.771	.911	.903	.877	.116	0						
y2	.747	.719	.786	.783	.767	.690	.699	.851	.777	.835	.784	.724	.765	.760	.864	.823	.801	.815	.842	.814	.870	.861	.906	.916	.936	.451	.442	0					
r2	.717	.713	.757	.740	.750	.680	.682	.838	.720	.089	.084	.288	.412	.478	.672	.794	.748	.688	.834	.766	.832	.813	.889	.889	.920	.792	.780	.795	0				
r4	.684	.686	.725	.728	.745	.646	.657	.814	.726	.106	.085	.322	.418	.497	.695	.792	.775	.704	.832	.797	.833	.811	.882	.888	.939	.797	.785	.784	.080	0			
r5	.718	.706	.735	.766	.730	.705	.702	.786	.711	.486	.488	.481	.490	.021	.661	.689	.709	.680	.754	.724	.829	.825	.750	.746	.777	.753	.794	.766	.469	.489	0		
r7	.736	.726	.774	.746	.692	.649	.665	.731	.667	.707	.699	.710	.621	.674	.712	.392	.349	.013	.339	.398	.554	.539	.496	.473	.476	.745	.729	.809	.686	.703	.680	0	
r9	.777	.818	.875	.835	.776	.779	.781	.826	.785	.705	.686	.705	.701	.658	.026	.840	.816	.733	.814	.759	.889	.856	.894	.897	.906	.834	.812	.862	.691	.713	.654	.731	0

The inferred gene tree




Subtree II
Subtree III
Subtree I
Ser-Met
Ser-His
Ser-Leu
Cys-His
y2
vi3
y1
v2
r9
f9
k5
k3
k4
v1
v4
v5a
v5b
v3
r5
f5
f1
f3
r4
r2
f4
f2
f7
r7
k2
k1
v6
f6
y3
vi1
y4
c1
vi2
100
92
100
94
73
100
62
42
33
52
0.1

Note on the bootstrap in this context

The numbers on the internal branches in the preceding tree are "bootstrap support" or "bootstrap confidence" values.

They are obtained in the following manner,

Columns of the alignment are sampled with replacement, in the standard bootstrap manner.

For each bootstrap sample - 200 were taken here - distances are calculated and a NJ tree constructed.

The numbers written on certain edges of the originally constructed tree - bootstrap confidences - are the the percentages of the sampled trees in which that edge is recreated. (Each edge splits the leaves into 2 sets.)

You might wonder what theory there is concerning these percentages, and in particular, whether there is any strict interpretation of them, perhaps asymptotic. Here I refer you to the book by Felsenstein, simply commenting that they are a useful rough guide to the reliability of the topology.