Incongruence

James Fleming - j.f.fleming@nhm.uio.no

@JamesfvFleming

NHM Uio

What is Incongruence?

- Incongruence is when you have different gene trees that disagree in an irreconcilable way.
- Incongruence, and Incongruent Sequences, are defined by their **effect** not their **cause.**
- Each tree presents a hypothesis of the evolutionary history of the dataset
- For a single gene, that is a hypothesis of the evolution of that gene
- **The gene's history is not always the taxa's history**

Why does Incongruence occur?

- Sometimes, it's an error
  - Model Misspecification
  - Sampling errors, assembly errors and contamination
- However, sometimes, the gene's history and the taxa's history are not the same. And this could mean:
  - Incomplete Lineage Sorting
  - Gene Duplication and Loss (Paralogy)
  - Horizontal Gene Transfer
  - Ancestral Hybridisation
- Working out what is causing this effect can be incredibly informative! We might want to see Biological Incongruence – but we never want Methodological.

What can that look like?
- Incomplete Lineage Sorting

- Here is where you have an inkling the gene tree doesn't fit the species tree.
- This suggests that the ancestral population may have been **Polymorphic**, possessing both alleles of a gene that has become fixed in different proportions
- This can occur for a variety of reasons
  - Fixation of an allele can be random or caused by environmental stress

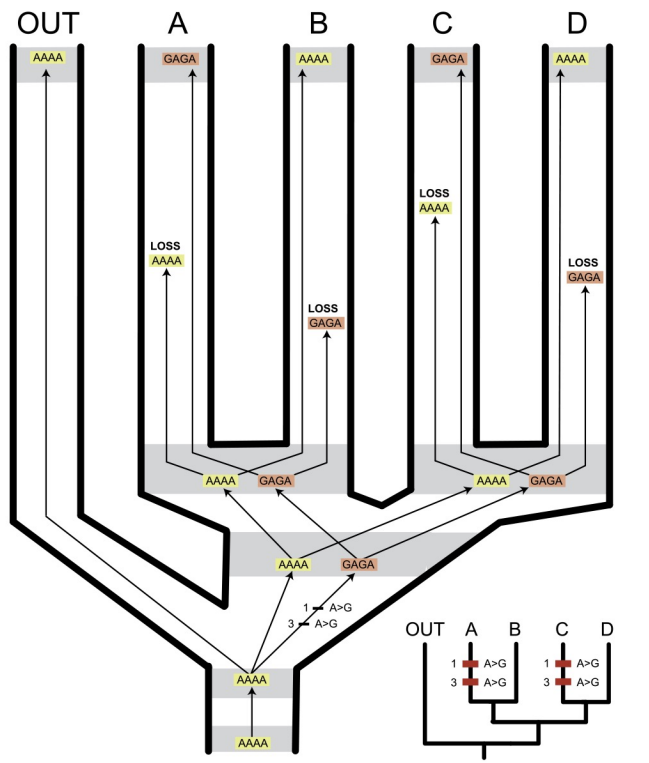
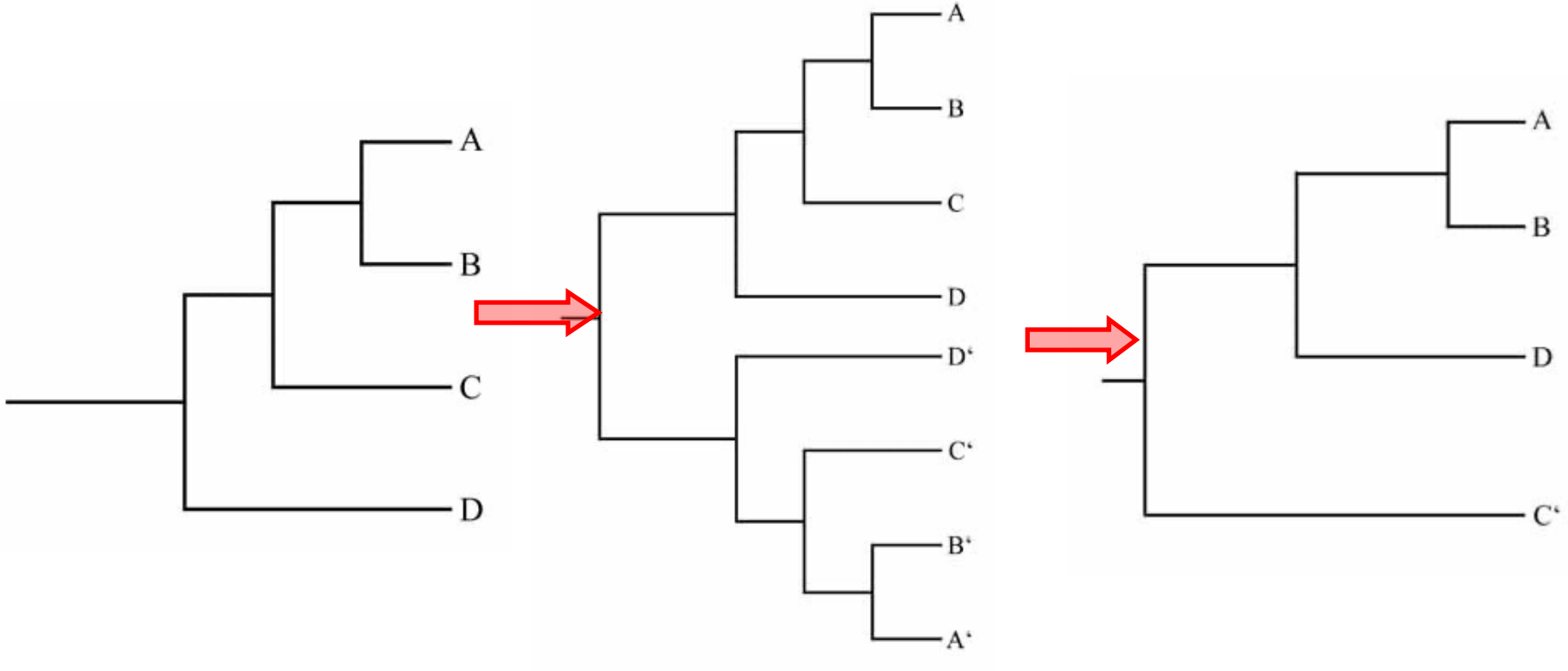


Fig. 2. A hypothetical scenario that shows mixed signal homoplasy ("hemiplasy") due to retention of ancestral polymorphism across multiple internal nodes and convergent sorting of alleles in different lineages. The true evolutionary history is shown in the large tree with four tightly spaced speciation events. A single allele is ancestral. Mutations at nucleotide positions 1 and 3 near the base of the tree yield a second allele. Alternative alleles are then fixed in each of the four terminal ingroup lineages (A–D). The small tree (lower right) shows a parsimony reconstruction of character state changes for the tree shown; note the inference of homoplasy at two sites that is due to the convergent, correlated fixation of mutations in alternative alleles, and not due to "standard" homoplasy (convergent mutation + fixation of the same state or a mutation + fixation that results in reversal to the primitive state). Note that substitution matrices commonly utilized by systematists model mutation + fixation and not just mutation (Patel et al., 2013). For example, different substitution models and branch lengths are commonly applied to different codon positions of protein-coding sequences (Ren et al., 2005). This is not because rates of mutation are thought to be very different at the three codon positions, but instead because rates of fixation are unequal due to different selective constraints.

Gatesy & Springer (2014)

What can that look like? Paralogy



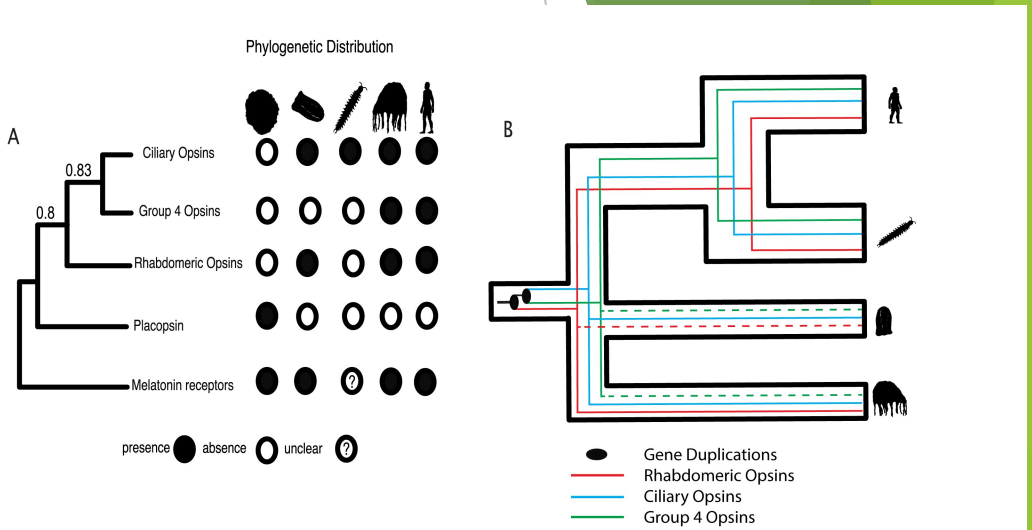
Species tree

Gene tree

Wrong species tree

What can that look like? Paralogy

- Paralogy can leave behind remnant signal that can be confusing
- Copies of a gene that have duplicated and then been lost can result in incorrect inferences if the gene tree is interpreted as a species tree.
- Be wary if your gene tree seems to require a great many independent instances of loss! It could be an indication that you have a few rogue sequences on your hands!



Fleming et al (2020)

What can that look like? Horizontal Gene Transfer

- Horizontal Gene Transfer is when genes are moved between species that are jointly extant, rather than passed to them from a shared ancestor, and can result in some exciting, and controversial gene trees.
- Again, if you find yourself predicting lots of HGT, make sure that you don't just have a situation with high contamination or sampling error!

By mapping our single-individual genome sequence reads and 35 RNA-Seq data (active and tun states of adults, first 1 d, 2 d, 3 d, 4 d, and 5 d of eggs after laying, first 1 d, 2 d, 3 d, 4 d, and 5 d after hatching, encompassing various developmental stages), we found that as many as 7,135 contigs (31.7% of all contigs), including the longest 11 contigs, are contaminated under a rather conservative estimate (less than ×1 coverage in genomic reads, or without even a single hit of RNA-Seq reads from any of the 35 datum). Only 1,771 putative HGT genes remain after removing these contigs, and their percentage within the genome, 4.47%, is in line with other eukaryotic genomes.

Arakawa (2016)

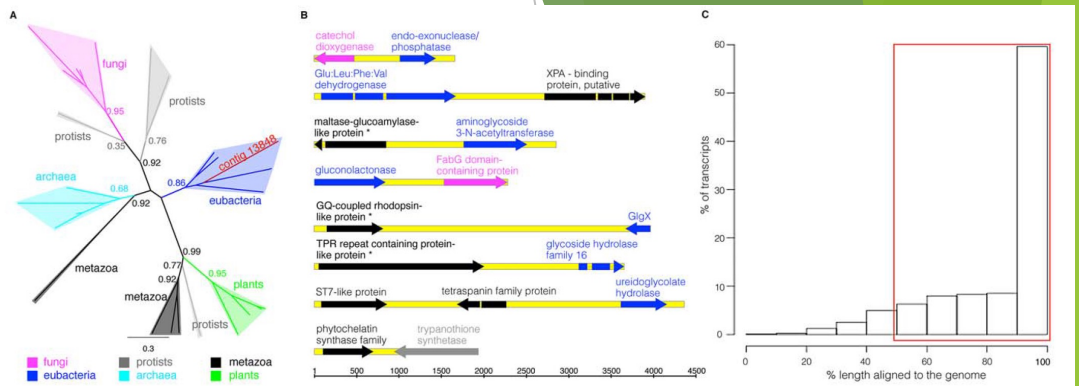
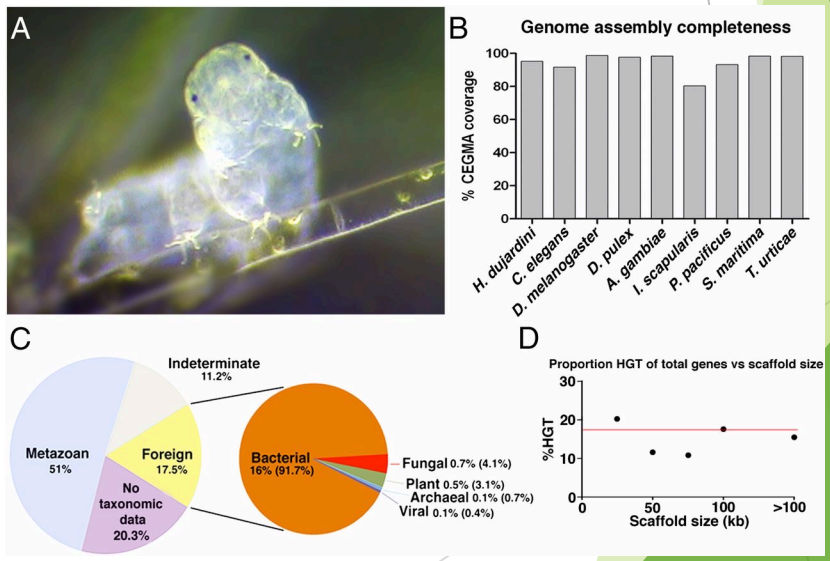


Figure 2. Foreign genes in the *A. ricciae* genome. (A) Phylogenetic tree for one exemplar bdelloid transcript (contig 13848) encoding an acetyl-CoA synthetase. Branch colours represent different taxa: metazoa, black; eubacteria, blue; archaea, light blue; fungi, pink; protists, grey; plants, green. Numbers on nodes represent aLRT support. (B) Physical linkage of foreign genes to neighbouring genes in the genome: eight different Sanger sequenced and assembled genomic regions, with arrows showing gene length and orientation (metazoa, black; eubacteria, blue; fungi, pink; protists, grey); introns are indicated as interruptions. Bdelloid genes previously identified in *A. vaga* are marked with an asterisk. In both the first and fourth genomic regions shown, the two foreign genes belong to different taxa (fungi and bacteria). Scale, bp. See also Figure S2 and Table S1. (C) Genomic coverage of *A. ricciae* foreign transcripts. Histogram of the percentage length aligned to the draft genome for all foreign transcripts. The red box indicates all foreign transcripts which align to the draft genome along greater than 50% of their length.
doi:10.1371/journal.pgen.1003035.g002

Boschetti et al. (2012)



Boothby et al (2015)

What can that look like? Ancestral Hybridisation

- Ancestral Hybridisation will regularly produce two conflicting hypotheses across multiple genes
  - Operational Genes (for metabolic processes) favour Eukaryotes+Eubacteria
  - Informational Genes (for transcription and translation) favour Eukaryotes+Archaea
- This implies that hybridization events occurred close to the Last Universal Common Ancestor between these groups, causing this confusion.

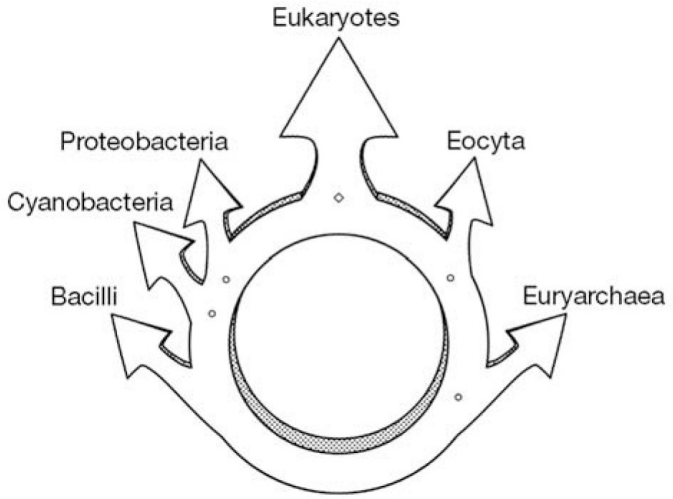


Figure 3 A schematic diagram of the ring of life. The eukaryotes plus the two eukaryotic root organisms (the operational and informational ancestors) comprise the eukaryotic realm (see Supplementary Discussion). Ancestors defining major groups in the prokaryotic realm are indicated by small circles on the ring. The Archaea[49], shown on the bottom right, includes the Euryarchaea, the Eocyta and the informational eukaryotic ancestor. The Karyota[5], shown on the upper right of the ring, includes the Eocyta and the informational eukaryotic ancestor. The upper left circle includes the Proteobacteria[49] and the operational eukaryotic ancestor. The most basal node on the left represents the photosynthetic prokaryotes and the operational eukaryotic ancestor.

Rivera & Lake (2004)

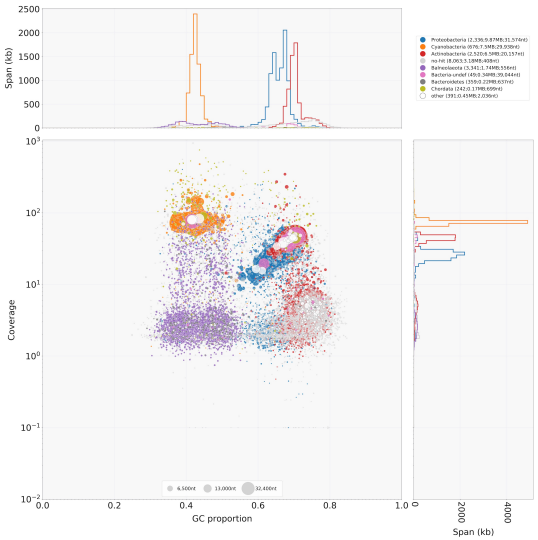
How can we find out what kind of incongruence we have?

- To detect incongruence, we can use a variety of approaches. Some are better at detecting certain types of incongruence than others.

1) BLAST-based approaches
2) Site-likelihood approaches
3) Tree-based approaches
4) Nodal support-based approaches

BLAST

- You can detect paralogy by BLASTing against a large dataset to find homologs not in your phylogeny
- You can detect Horizontal Gene Transfer and contamination by finding where the genes most similar to these genes occur
- With a tool like blobtools we can BLAST everything in our genome to determine the likely origin of each sequence in the assembly, and visualize it in a taxon plot!



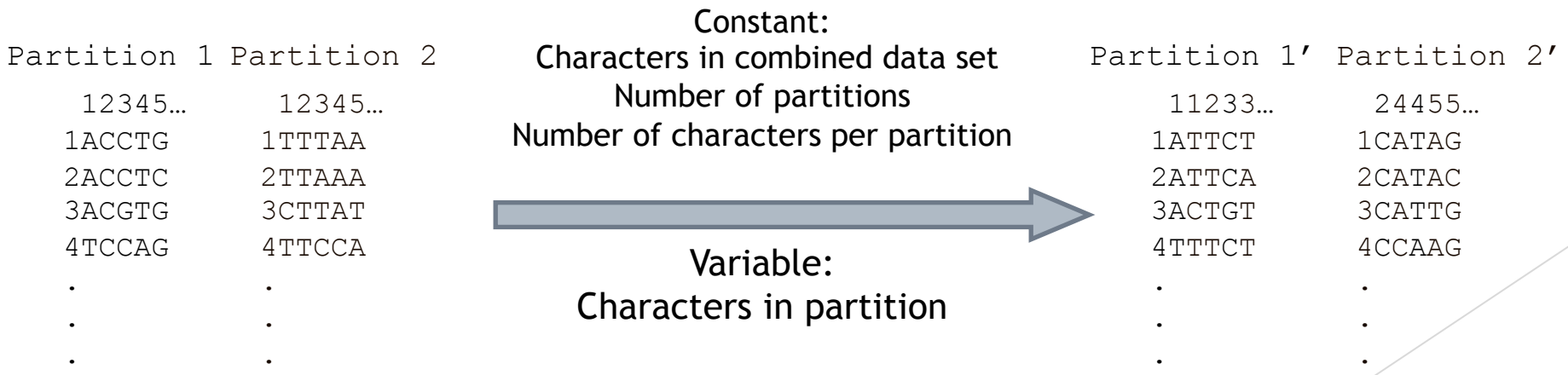
Laetsch and Blaxter (2017)

BLAST

- BLAST is the basic standard for gene search and comparison strategies, so it is useful to remember it!
- BLAST initially looks for short matches between your sequence and sequences in the target database called ‘words’ – by default 3 amino acids in protein datasets, for example.
- It then searches for whether the words exist in the same region (neighbourhood) as other words, and starts to join them together
- It then ranks these final matches using an ‘e-value’ – a value representing how likely it is that a match of that size and similarity occurred by chance.

A Site-Likelihood Approach: Permutation Test

- The Permutation test asks a simple question
  - “Given the frequencies of the character states observed, is the signal different from a neutral baseline?”
- It shuffles the character states within your dataset to do this, attempting to assess if there is structure in the data. If there is structure, and it is conflicting between datasets, that implies the existence of incongruence!

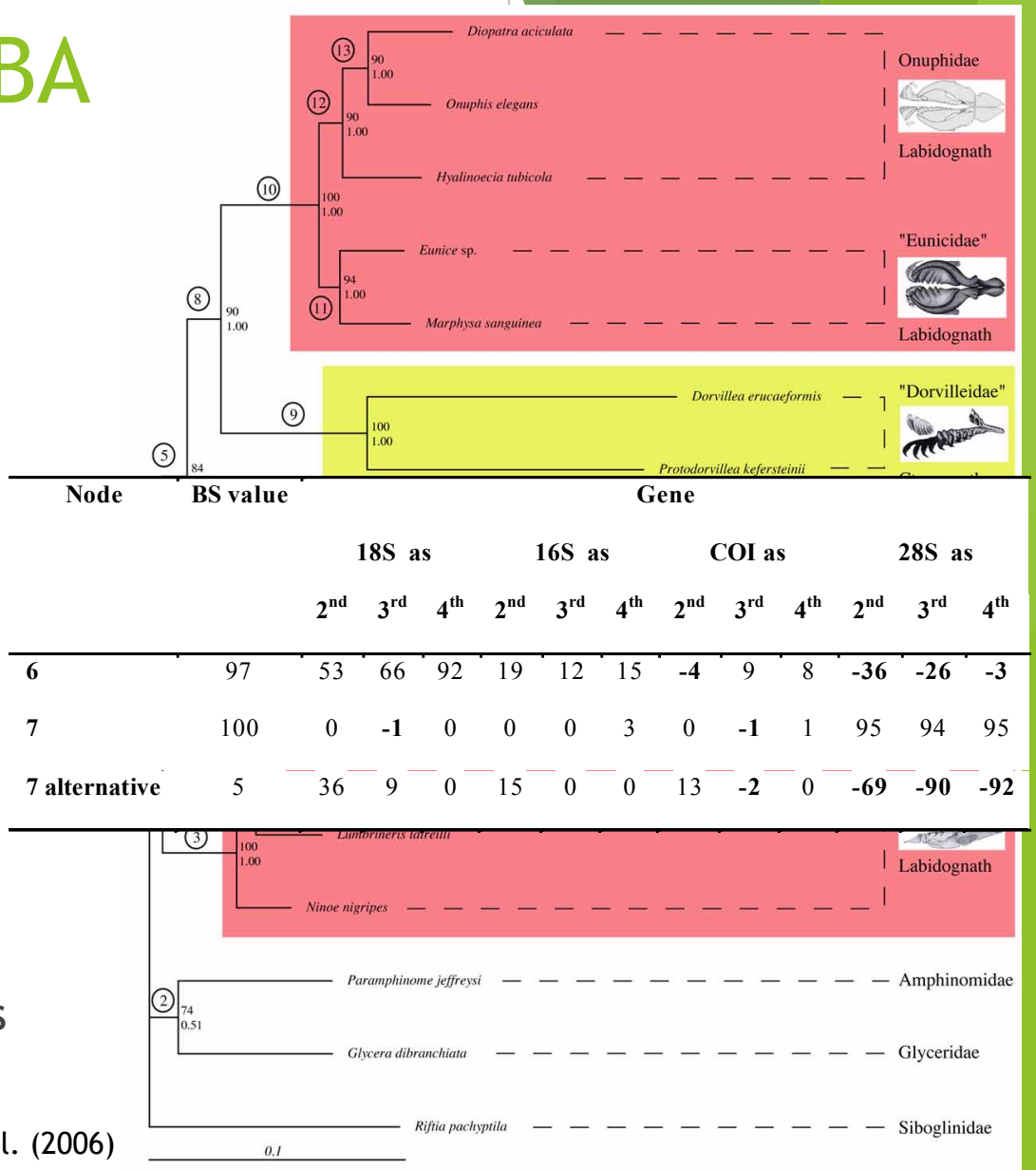


A Tree-Based Approach: Incongruence Length Difference

- The Incongruence Length Difference or “Partition Homogeneity” test developed by Farris et al (1994) is a simple tree-based approach
- It fixes the topology of the tree, and then optimizes each partition in your dataset independently on it.
- This allows you to determine whether the uncertainty in a large dataset is due to incongruity (between partitions) or noise (within partitions)
- Is the length difference between the fixed tree and the optimized tree due to the partitioning of the data?
- It expresses this as a p-value detecting the presence of incongruity in the dataset respective to a particular partition.

A Nodal Support Approach: PABA

- PABA is short for Partition Addition Bootstrap Alteration.
- Fundamentally, it again simplifies the complex question of Incongruence to a single question:
  - Where does incongruence occur, and which datasets significantly support which topologies.
- This means it is useful for detecting Paralogy and Incomplete Lineage Sorting.
- Does the addition of more data partitions affect the bootstrap support values significantly. Where, and in which direction?
- PABA then expresses this as a table of changing BS values



Node	BS value	Gene											
		18S as			16S as			COI as			28S as		
		2nd	3rd	4th	2nd	3rd	4th	2nd	3rd	4th	2nd	3rd	4th
6	97	53	66	92	19	12	15	-4	9	8	-36	-26	-3
7	100	0	-1	0	0	0	3	0	-1	1	95	94	95
7 alternative	5	36	9	0	15	0	0	13	-2	0	-69	-90	-92

Struck et al. (2006)

Summary

- Incongruence is an issue that can arise biologically or methodologically.
- In order to detect and separate true incongruence from data artifacts, we can employ a wide array of tools. This is the goal of the whole first week of this course.
- Biological incongruence can manifest as
  - Incomplete Lineage Sorting
  - Gene Duplication and Loss (Paralogy)
  - Horizontal Gene Transfer
  - Ancestral Hybridisation
- BLAST, Permutation tests, Incongruence Length Difference tests and PABA are all methods we can use to detect incongruence.