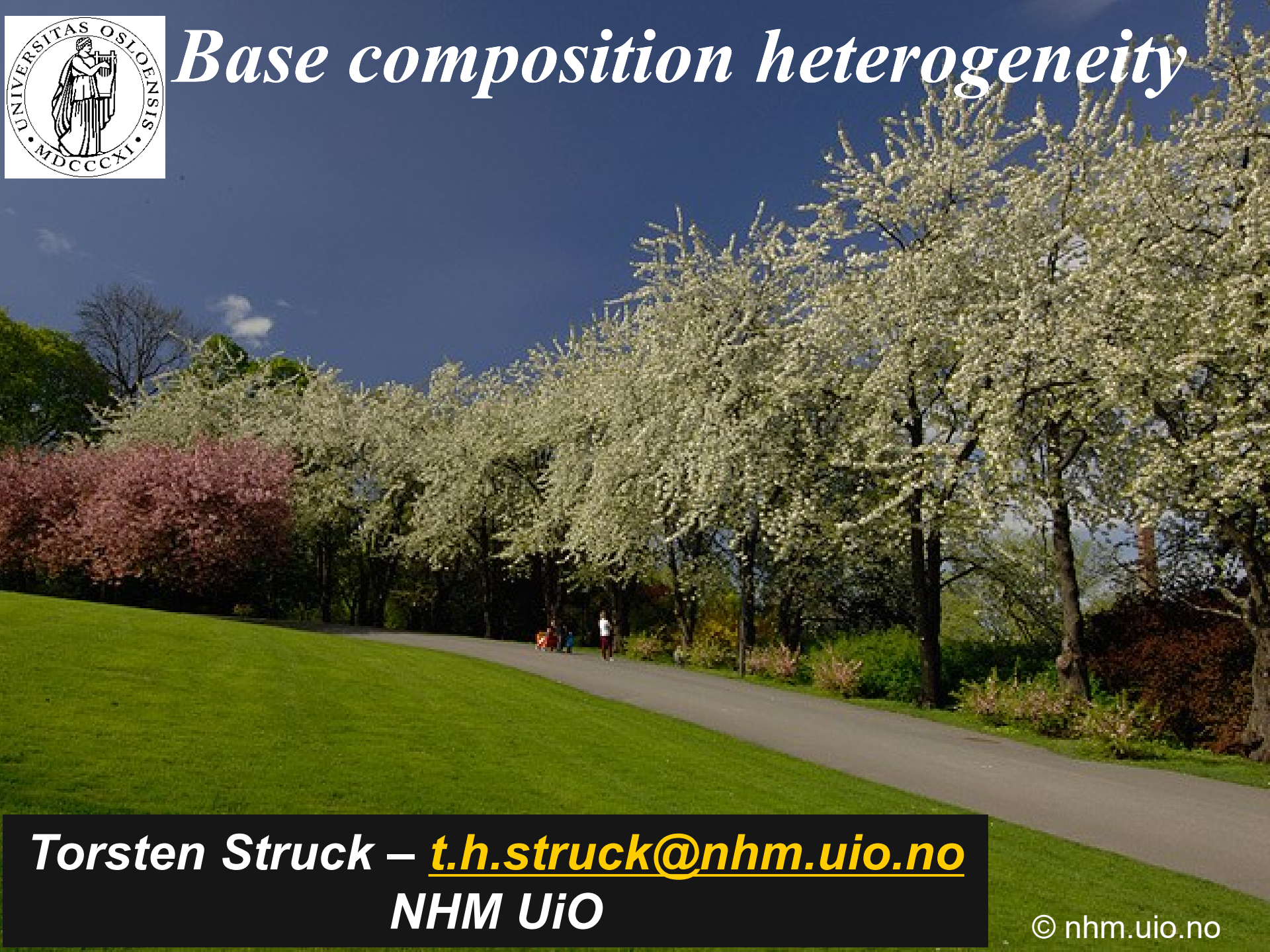
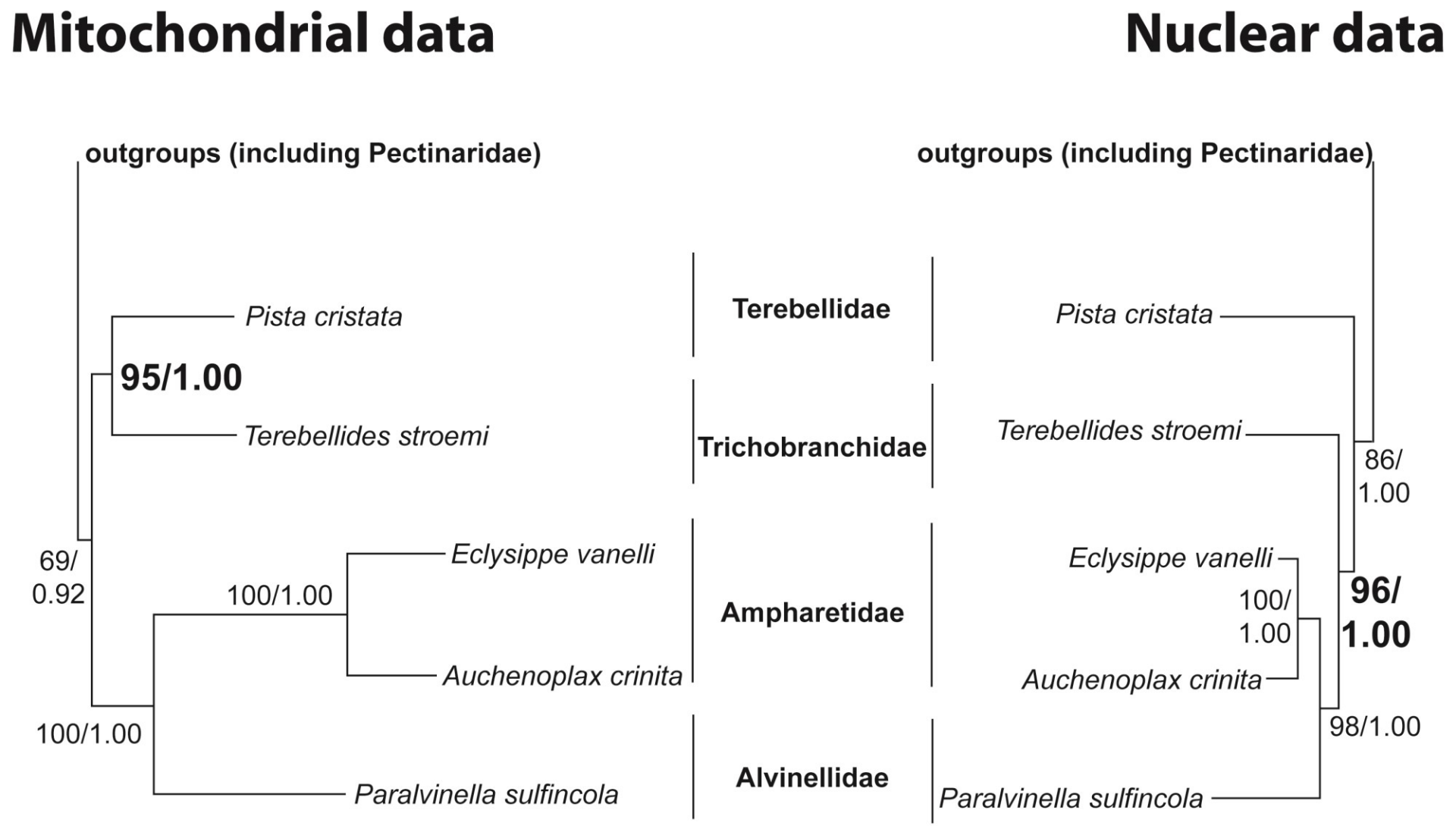


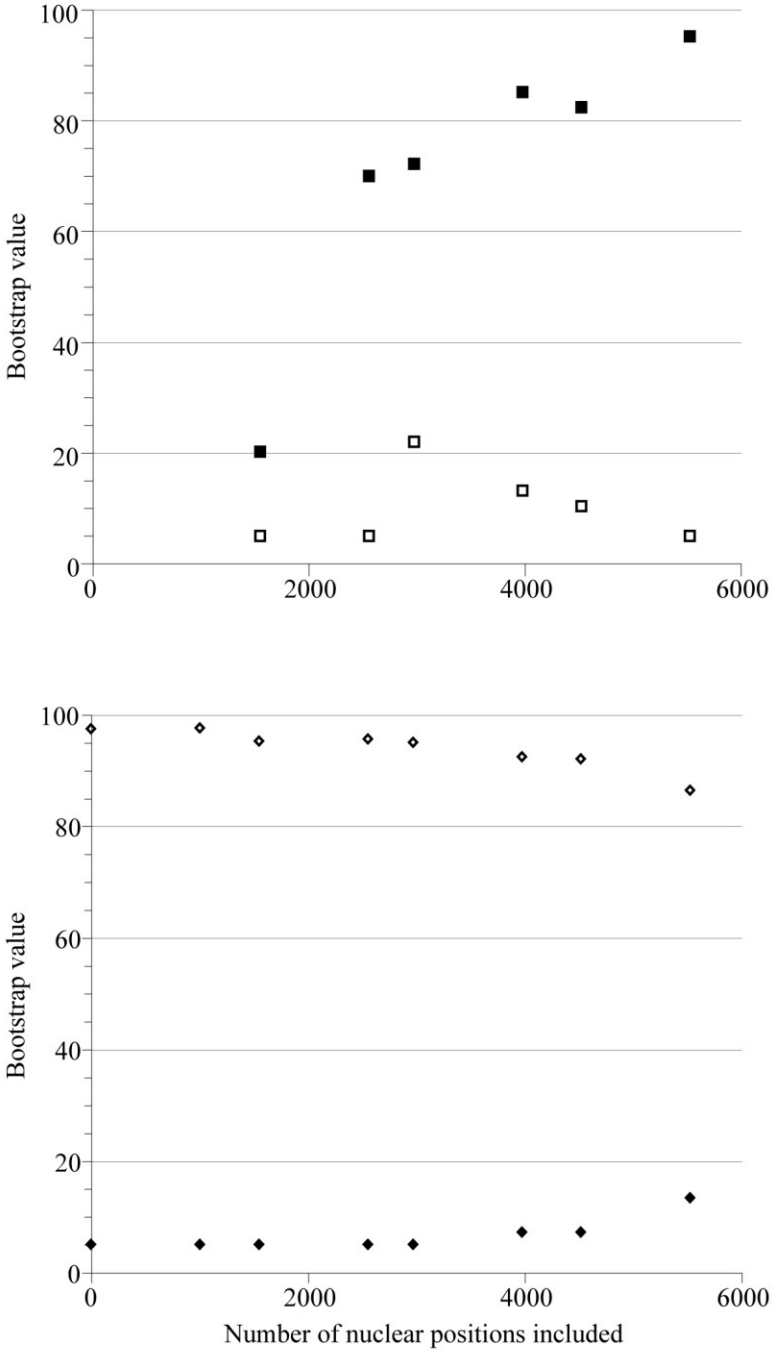
Base composition heterogeneity

Torsten Struck – t.h.struck@nhm.uio.no
NHM UiO

An example

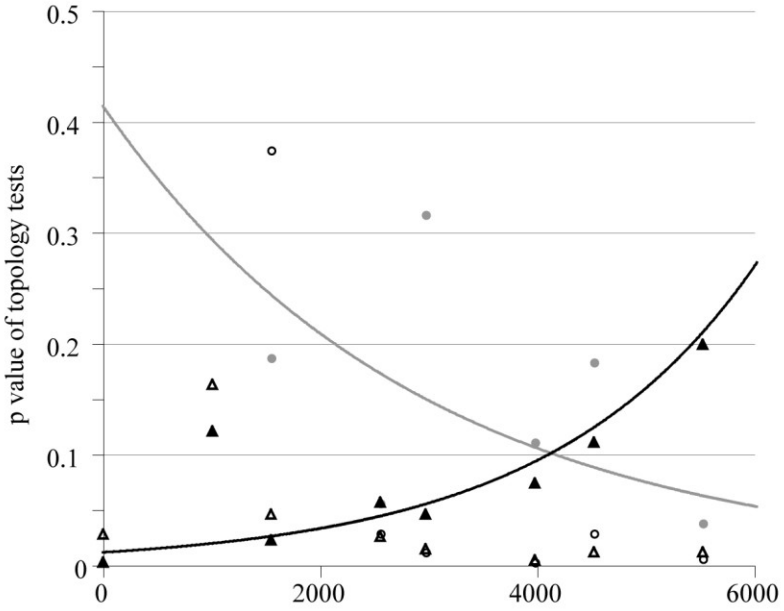


Zhong et al. (2011)



Bootstrap-Support (nuclear data without mitochondrial data)

Topology-Tests



Black symbols indicate TriAA, grey symbols the TriTer and open symbols the TerAA hypothesis. Circles stand for all possible combinations of only the nuclear partitions and triangles for mitochondrial data plus all possible combinations of the nuclear partitions.

Bootstrap-Support (nuclear data with mitochondrial data)

Rooting problem

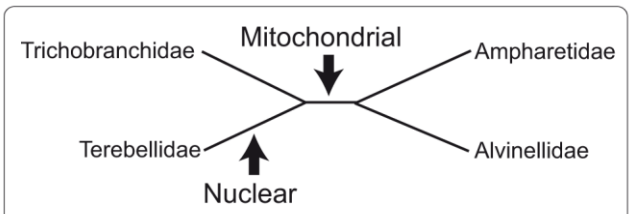


Figure 5 The unrooted subtree of Trichobranchidae, Terebellidae, Alvinellida and Ampharetidae. Arrows indicate the position of the root by either nuclear or mitochondrial data.

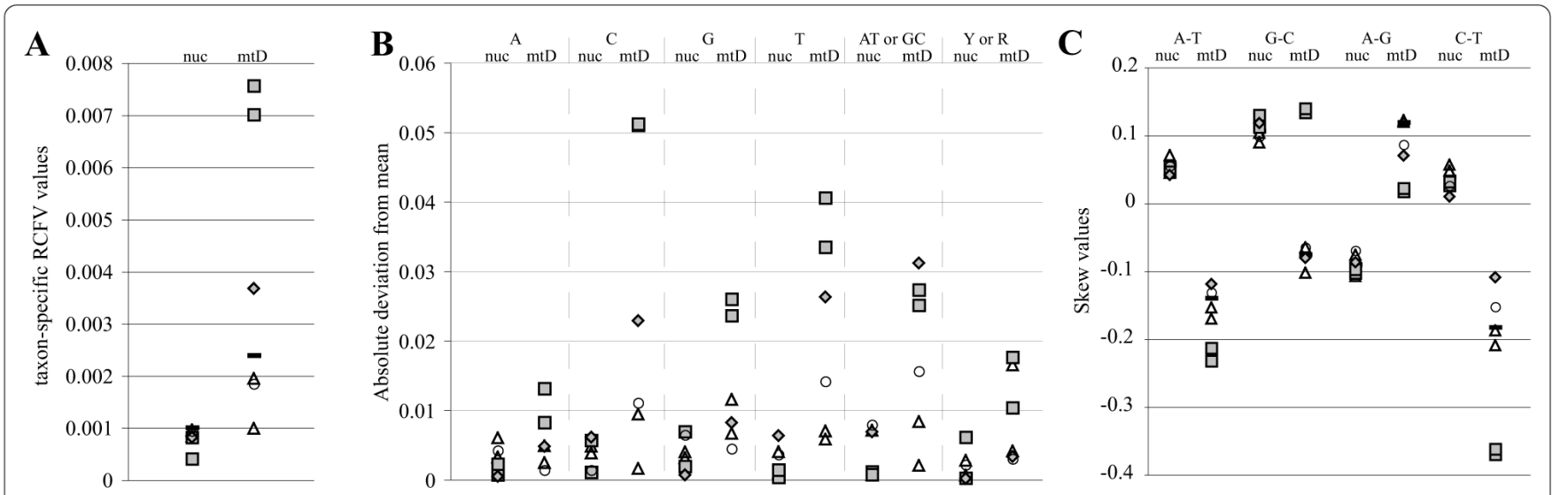


Figure 6 Analyses of compositional heterogeneity in nuclear and mitochondrial datasets in the mitochondrial protein-coding genes. (A) Taxon-specific relative composition frequency variability (RCFV). (B) Absolute deviation from mean frequency for adenine (A), cytosine (C), guanine (G), and thymine (T) as well as combinations of adenine/thymine (AT) or guanine/cytosine (GC) and of pyrimidines (Y) or purines (R). Only one absolute value is provided for AT and GC or Y and R as only two character states are now present and any change in one state has the exact opposite negative or positive value in the other. (C) Skew values within the combinations adenine/thymine (A-T), guanine/cytosine (G-C), purines (A-G) and pyrimidines (C-T). Ampharetidae (grey squares), Alvinellidae (grey diamonds), Pectinariidae (open circles), Trichobranchidae and Terebellidae (both open triangles), mean values of outgroup taxa (black bar), nuclear (nuc), mitochondrial (mtD) .

RY coding

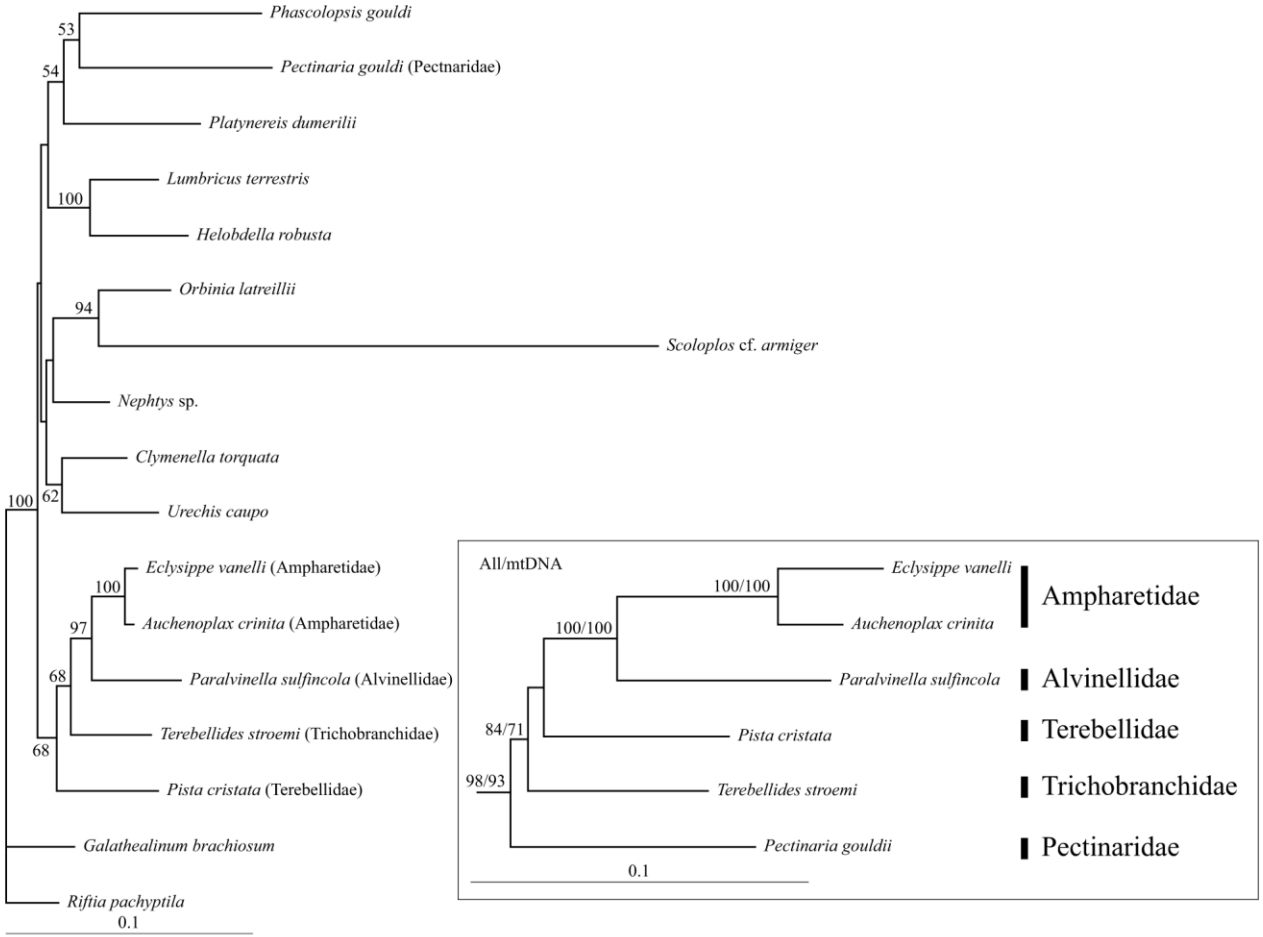


Figure 7 Phylogenetic reconstructions using nuclear, mitochondrial and combined datasets based on RY coding. Only the nuclear ML tree is completely shown. With respect to terebelliform relationships, analyses of the mitochondrial and combined dataset recovered the same topology. Therefore, in the inlet only this part of the mitochondrial ML tree is shown and no outgroups. Only bootstrap values above 50 are shown. In the inlet, bootstrap values of the mitochondrial analysis are given at the first position and of the combined analysis at the second.

Symplesiomorphy trap

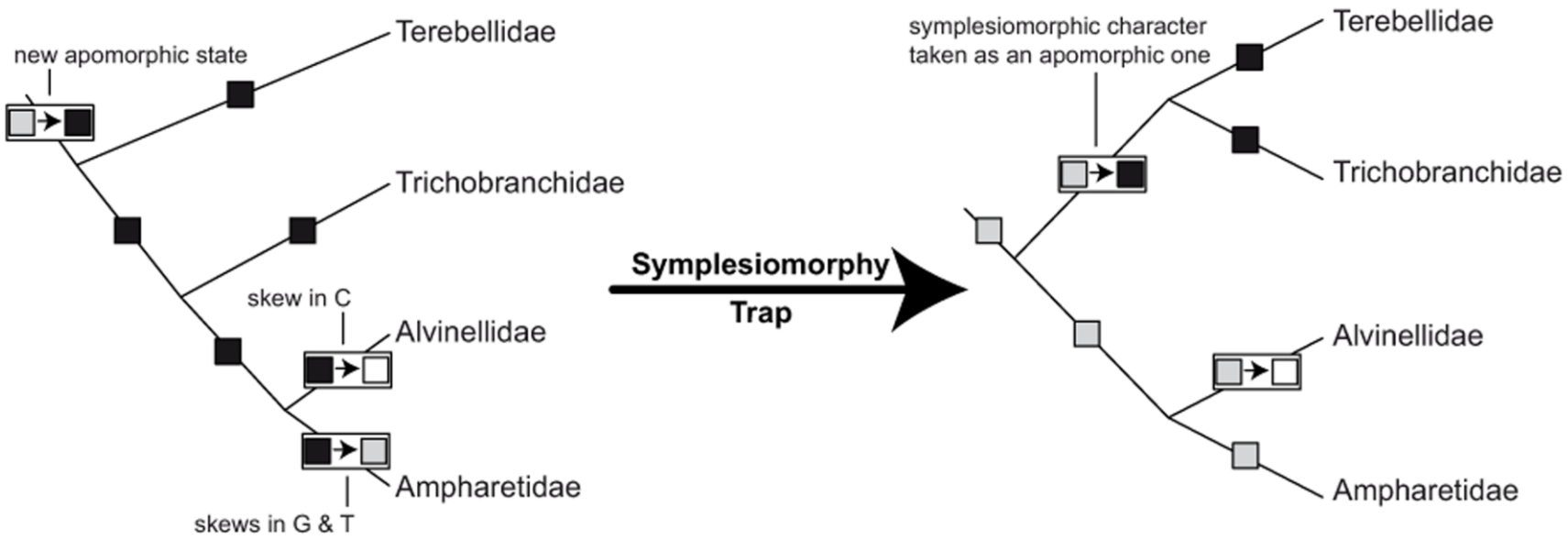


Figure 8 Schematic representation of the effect of biases with respect to the symplesiomorphy trap in our terebelliform example. White, grey and black boxes indicate different character states as well as the possible change of one state to another along a branch.

RCFV

Compositional bias and heterogeneity –
RCFV (relative composition frequency variability)

compositional bias
(state-specific RCFV)

$$RCFV = \sum_{i=1}^{n} \sum_{j=A}^{T \text{ or } U \,/\, W} \frac{\left| \mu_{ij} - \overline{\mu_j} \right|}{n}$$

taxon-specific RCFV (across all states)

Phillips & Penny (2003); Zhong et al. (2011); Kück & Struck (2014)

Skew values

skew values of nucleotides –
A/T & G/C – skewed strands
A/G & C/T – skewed classes

$$A/T = \frac{\mu_A - \mu_T}{\mu_A + \mu_T}; G/C = \frac{\mu_G - \mu_C}{\mu_G + \mu_C}; A/G = \frac{\mu_A - \mu_G}{\mu_A + \mu_G}; C/T = \frac{\mu_C - \mu_T}{\mu_C + \mu_T}$$

Perna & Kocher (1995); Zhong et al. (2011)

Direct compositional frequencies

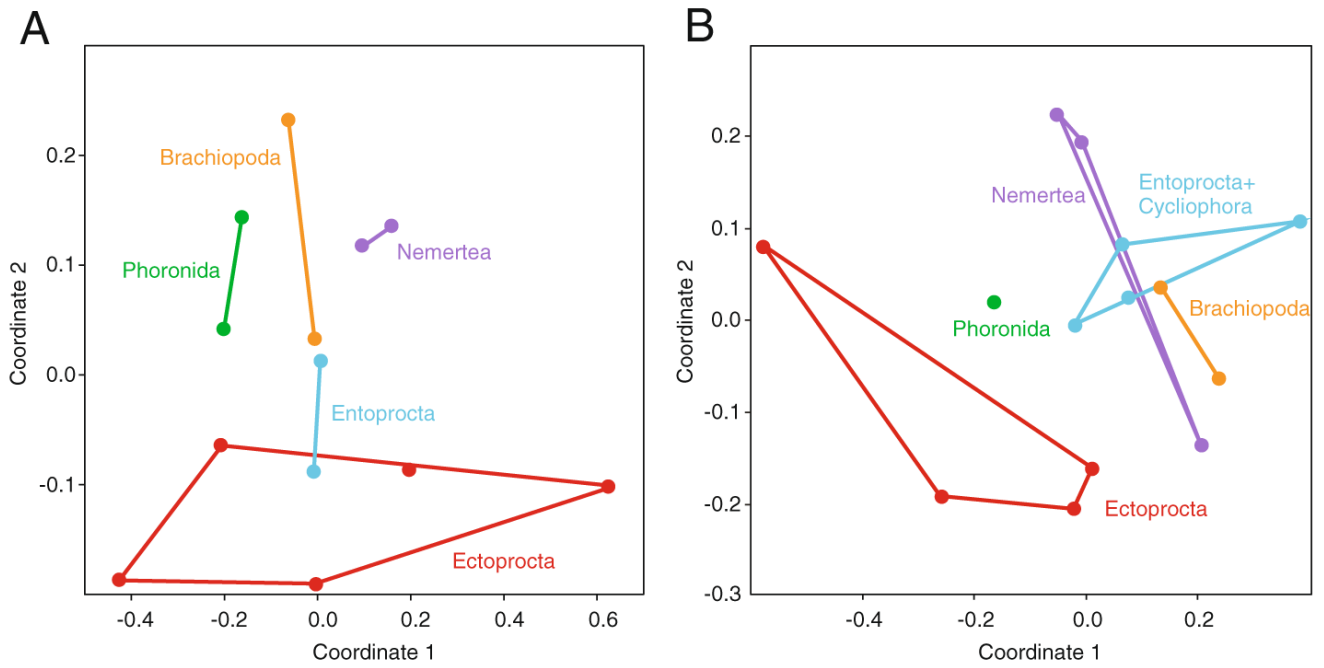
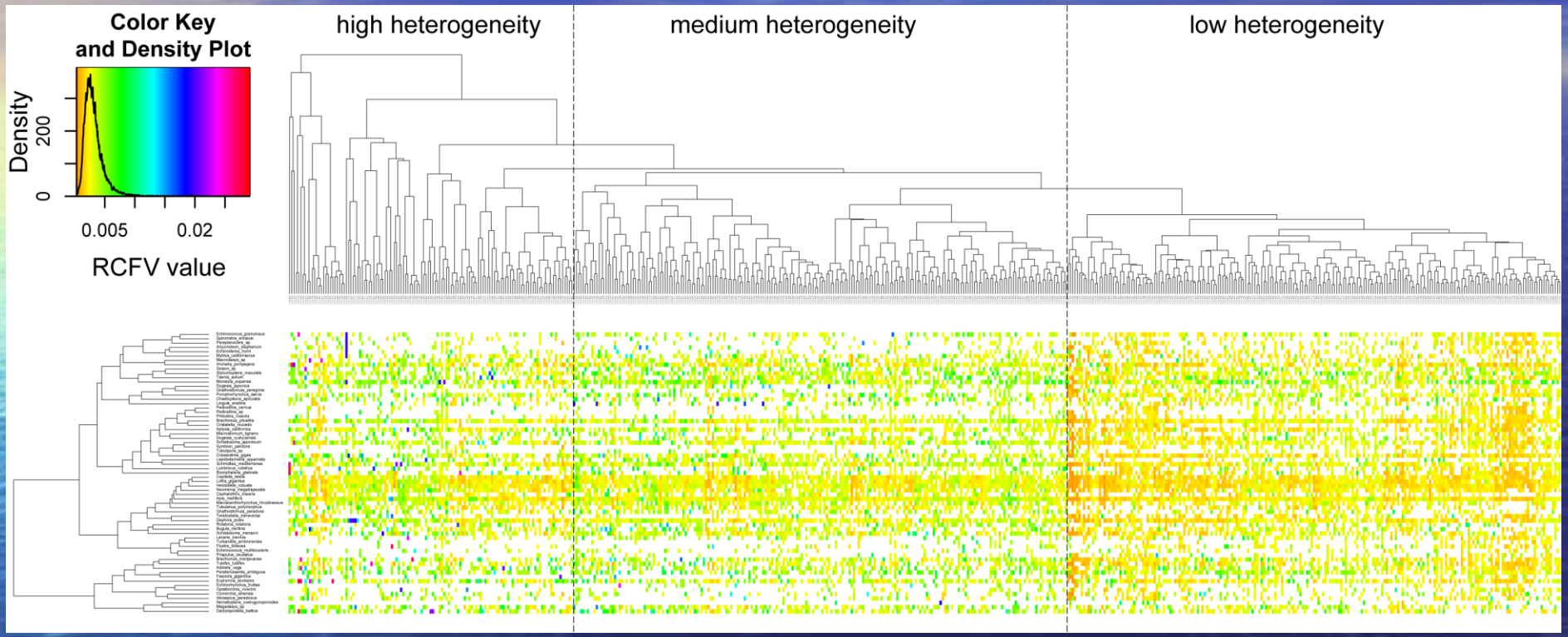


Figure 6 Non-metric multidimensional scalings of compositional distances between amino acid sequences. Scaling of distances between focal taxa using **(A)** the ribosomal protein dataset of Nesnidal et al. [35] and **(B)** the dataset used in this study.

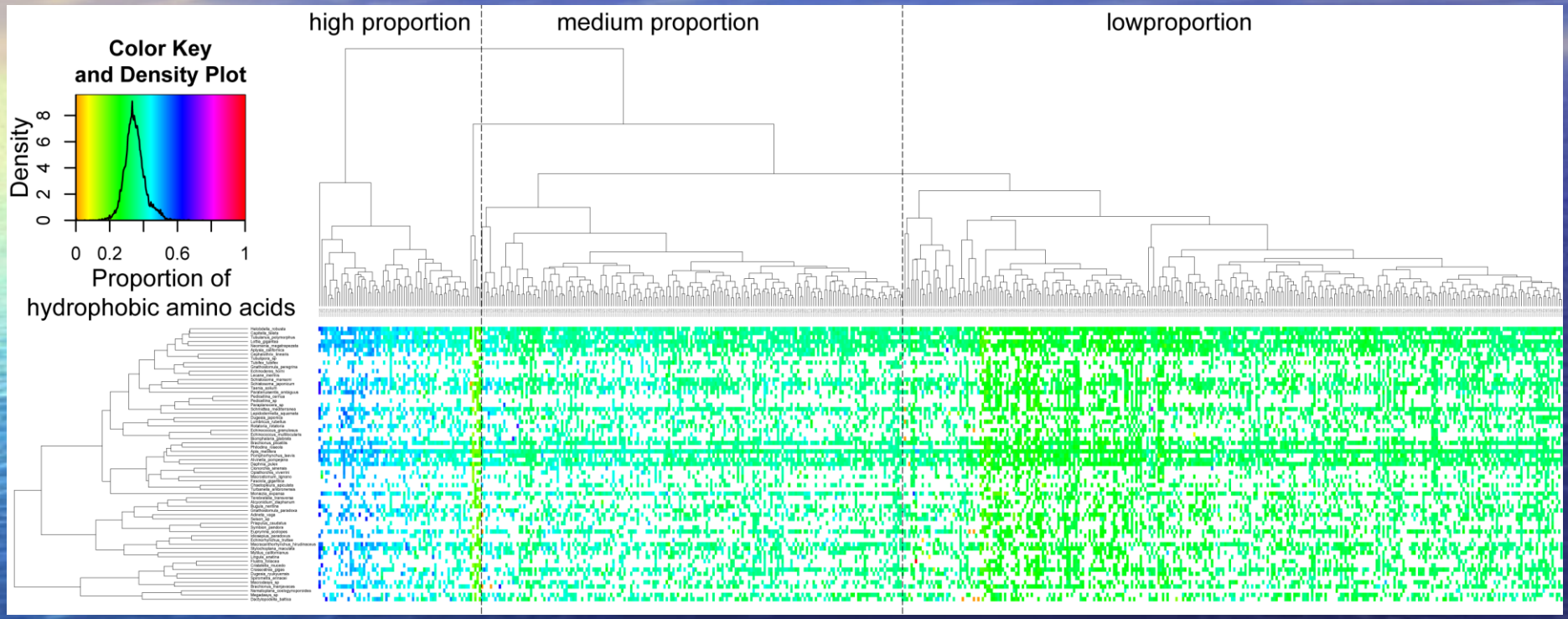
Nesnidal et al. (2013)

Heatmap of RCFV



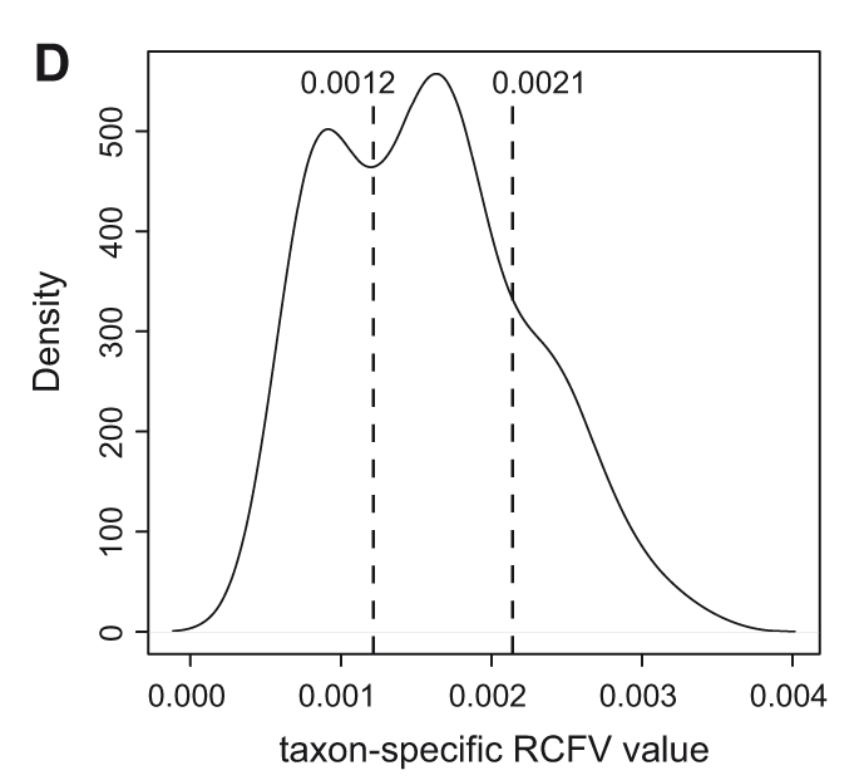
Struck et al. (2014)

Heatmap of amino acid frequencies



Struck et al. (2014)

Density plots



Golombek et al. (2015)