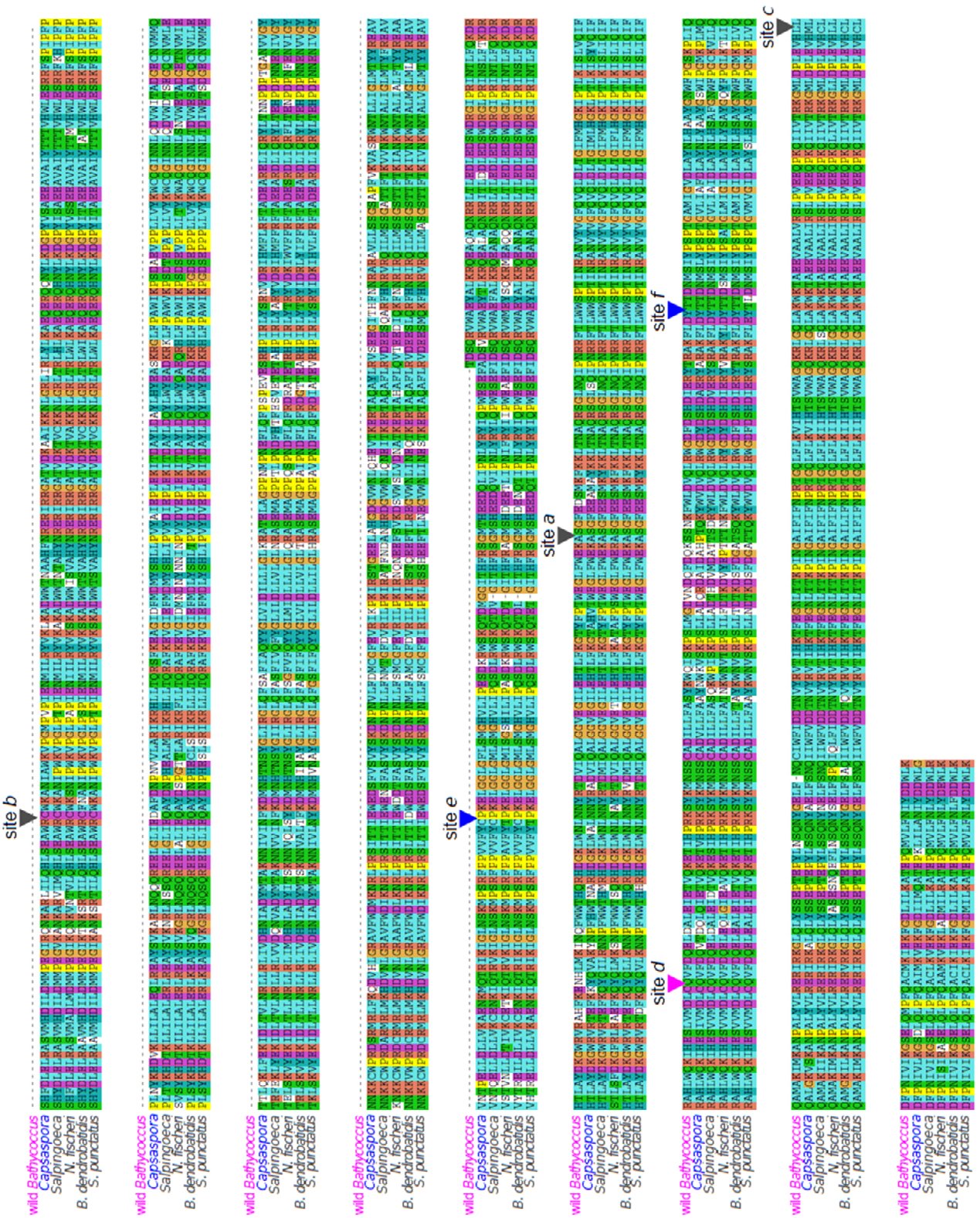
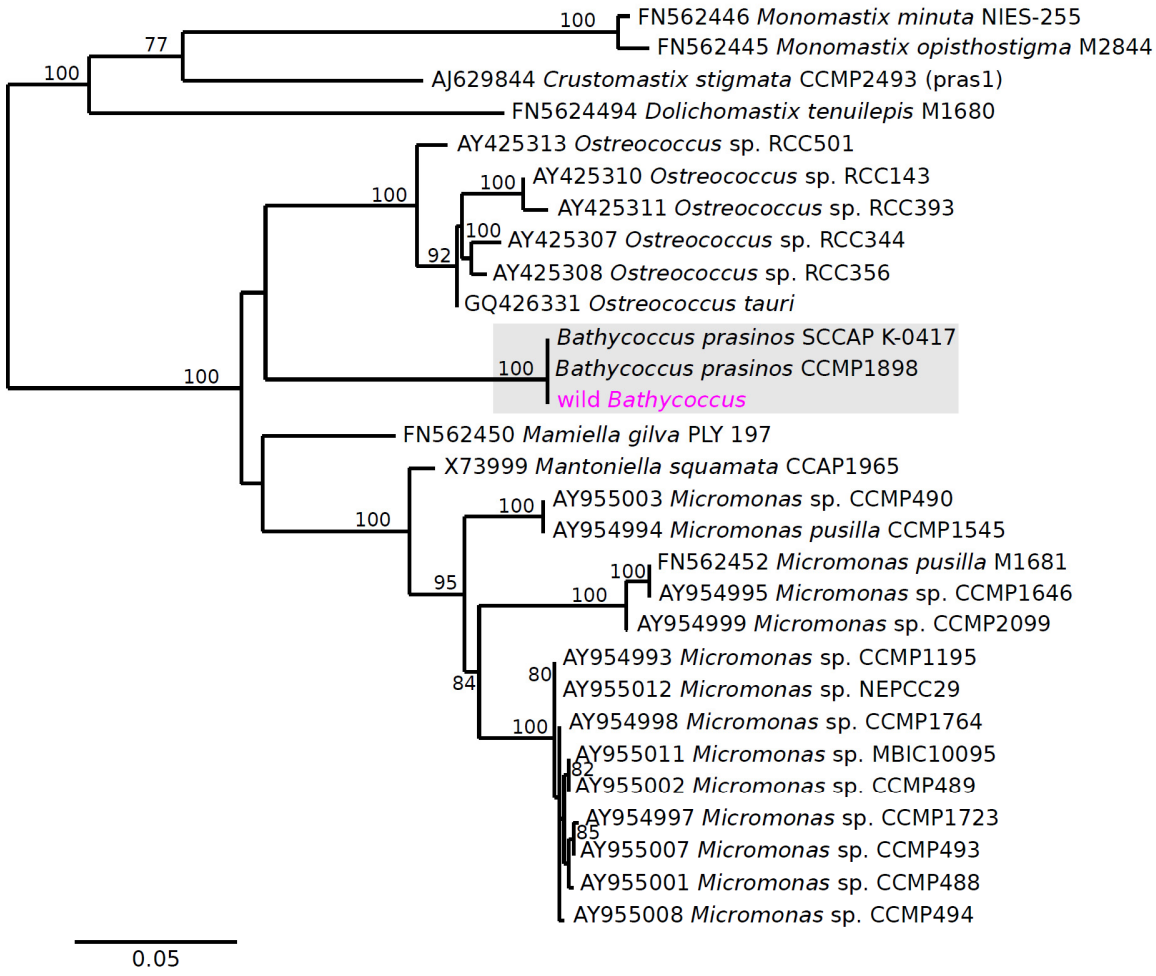
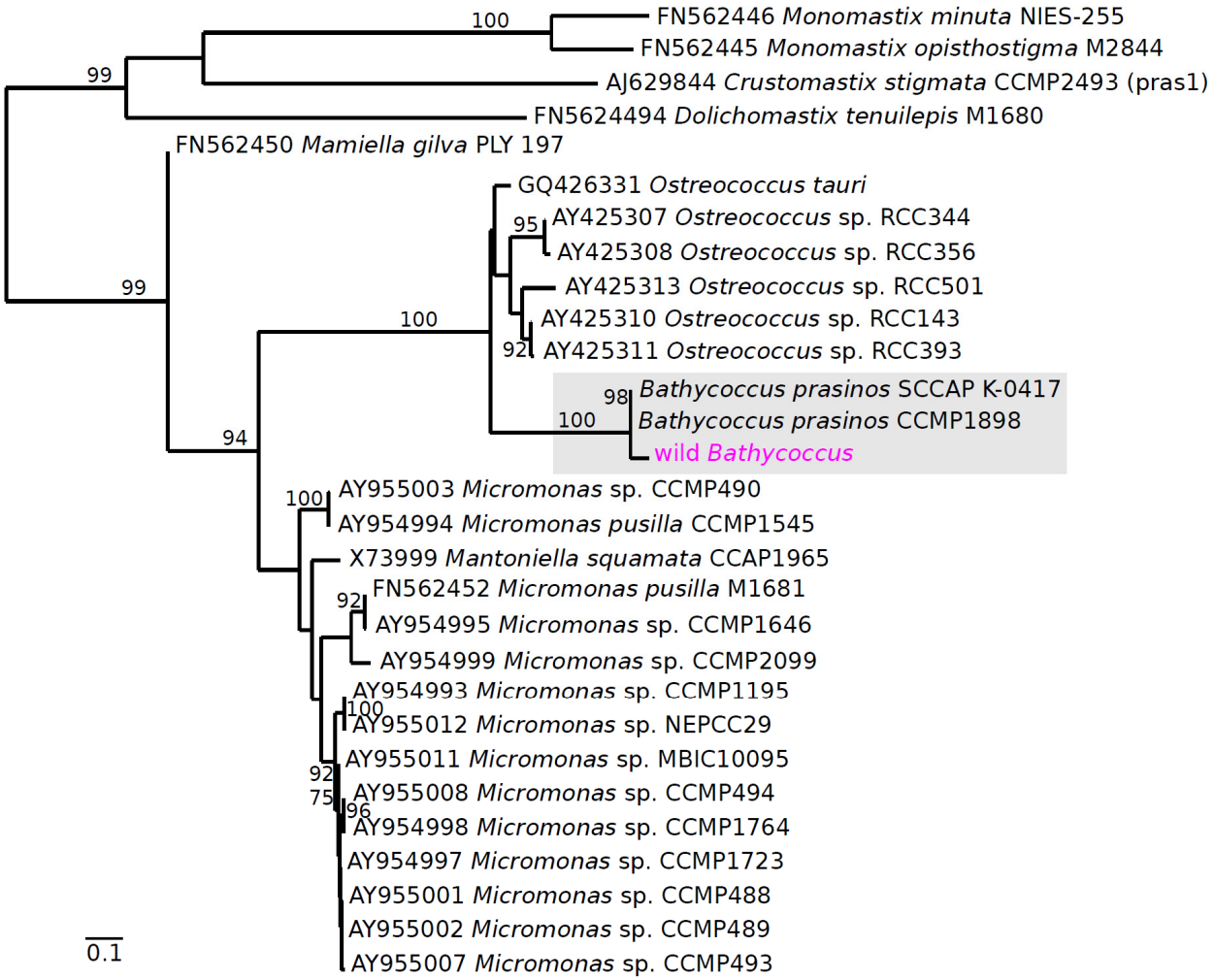
Monier et al. Supplementary Figure 1



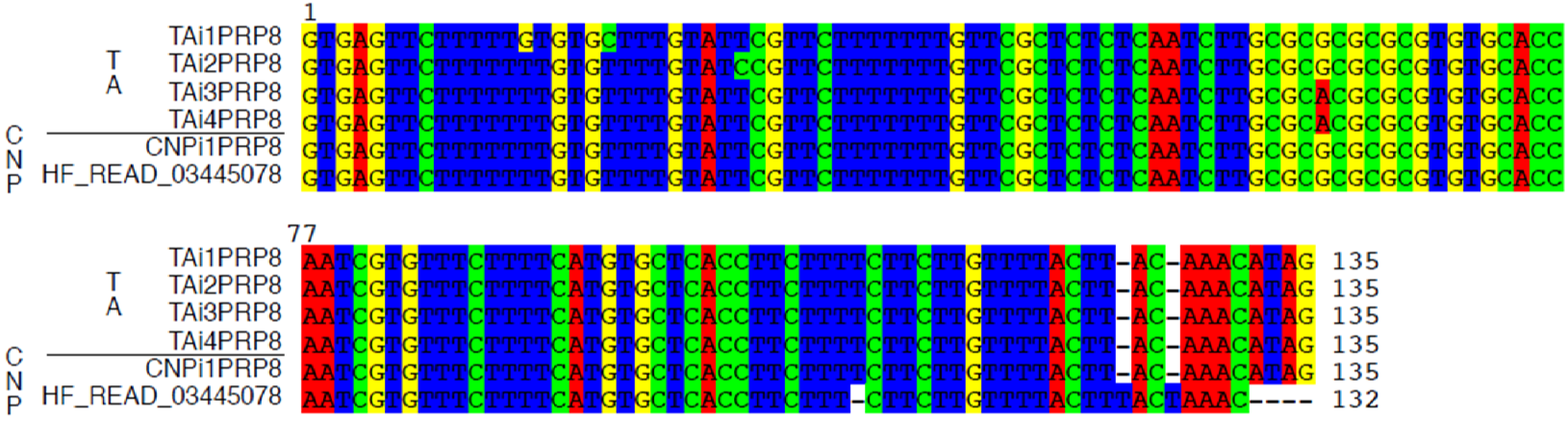
Supplementary Figure 1 - Alignment of flanking sequences of PRP8 intein hosts. Alignment of the *prp8* gene segment that hosts inteins and intein insertion positions in fungal representatives (sites *a*, *b*, and *c*; grey arrows), and those deposited in InBase as part of this study: the wild *Bathycoccus* population (site *d*; fuchsia arrow), *Salpingoeca rosetta* (site *a*; InBase identifier: Sro-PRP8) and *Capsaspora owczarzaki* (site *b*, and blue arrows, *e* and *f*, which were 480, 542 and 536 residues in length; InBase identifiers: Cow-PRP8-1, -2 and -3, respectively). Forty-four fungal PRP8 inteins are known (only a representative selection is shown here). Note that for HE domains within full inteins (not shown) substitutions at two aspartates in the HE enzyme's active site, one in block D and one in E (Figure 2), reportedly render HEs inactive towards DNA double-strand cleavage (1). These aspartate residues are conserved in intein-encoded HEs from *Bathycoccus* (site *d*) and *C. owczarzaki* sites *b* and *e*, supporting the idea that they are active or have been recently. *C. owczarzaki* site *f* HE has the block E aspartate only and *S. rosetta* has only a mini-intein.



Supplementary Figure 2 - Wild *Bathycoccus* phylogenetic relationships with other Mamiellales based on 18S rRNA genes. Maximum-likelihood reconstruction using model TIM2+I+G (best fitting as determined in jModelTest with AIC criterion) was based on a multiple sequence alignment of 1,679 nucleotide positions, including the wild *Bathycoccus* and CCMP1898 18S rDNA sequenced herein. The entire rRNA operon from the wild *Bathycoccus* (fuchsia) was present on a single metagenomic scaffold in the targeted metagenome. The wild *Bathycoccus* 18S rRNA branched with sequences from cultured *Bathycoccus* strains with 100% support and no divergence (nucleotide identity between the wild population and CCMP1898 was 99.9%). Node support ≥75% (100 bootstrap replicates) and substitutions per site (scale bar) are shown.

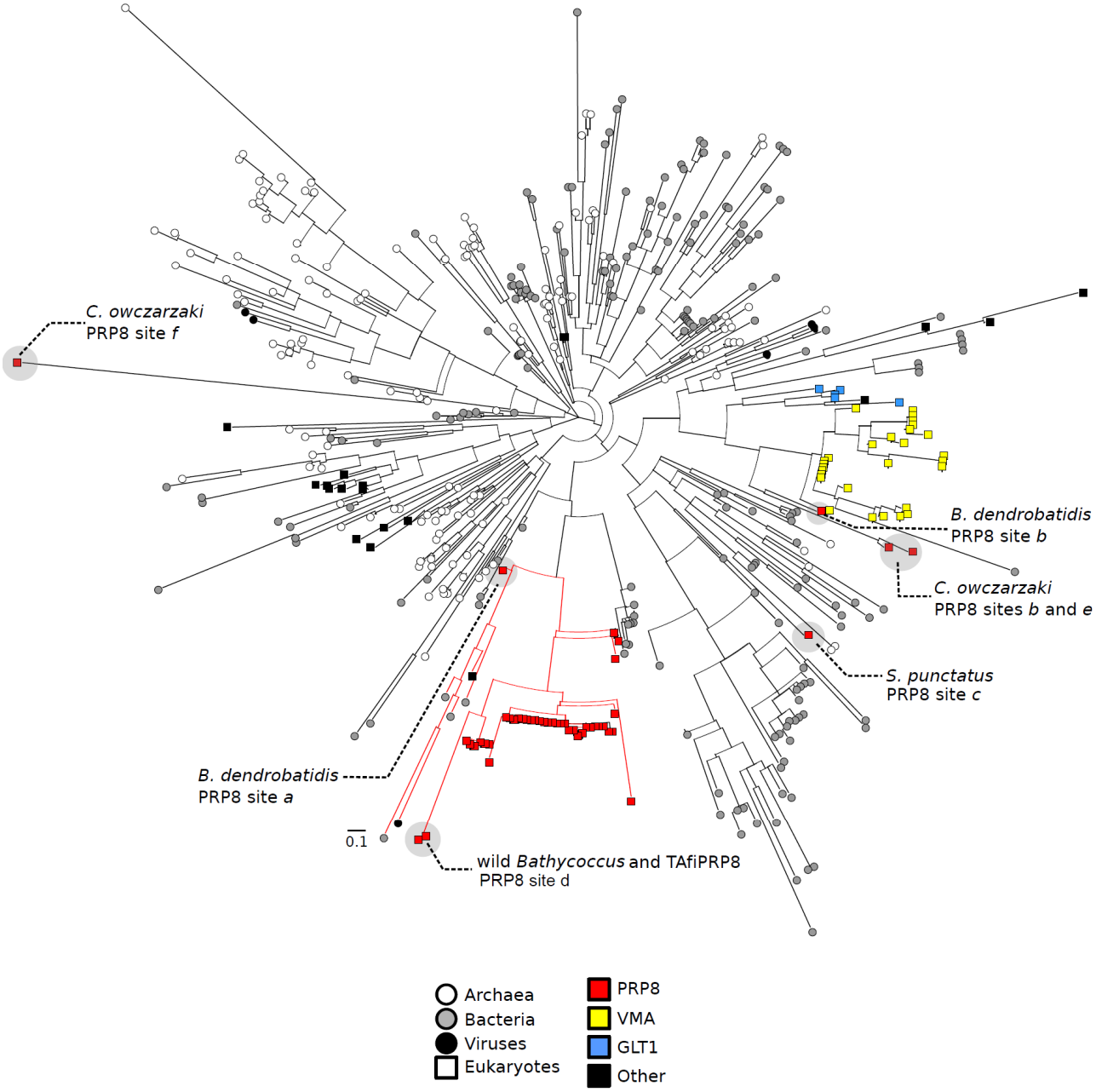


Supplementary Figure 3 - Wild *Bathycoccus* phylogenetic relationships with other Mamiellales based on ITS2 and 5.8S rRNA genes. Maximum-likelihood reconstruction using model GTR+I+G (best fitting as determined in jModelTest with AIC criterion) based on a multiple sequence alignment of 430 nucleotide positions, with sequences generated herein from cultured *Bathycoccus* strain CCMP1898 and wild *Bathycoccus* (fuchsia), wild *Bathycoccus* sequence presented slight divergence from its cultured counterparts, but less divergence than observed between different *Ostreococcus* clades or between different *Micromonas* clades. Results from the independent phylogenetic analyses of the 18S rDNA and this phylogeny are consistent with known relationships between *Bathycoccus* and other Mamiellales genera. Node support ≥75% (100 bootstrap replicates) and substitutions per site (scale bar) are shown.

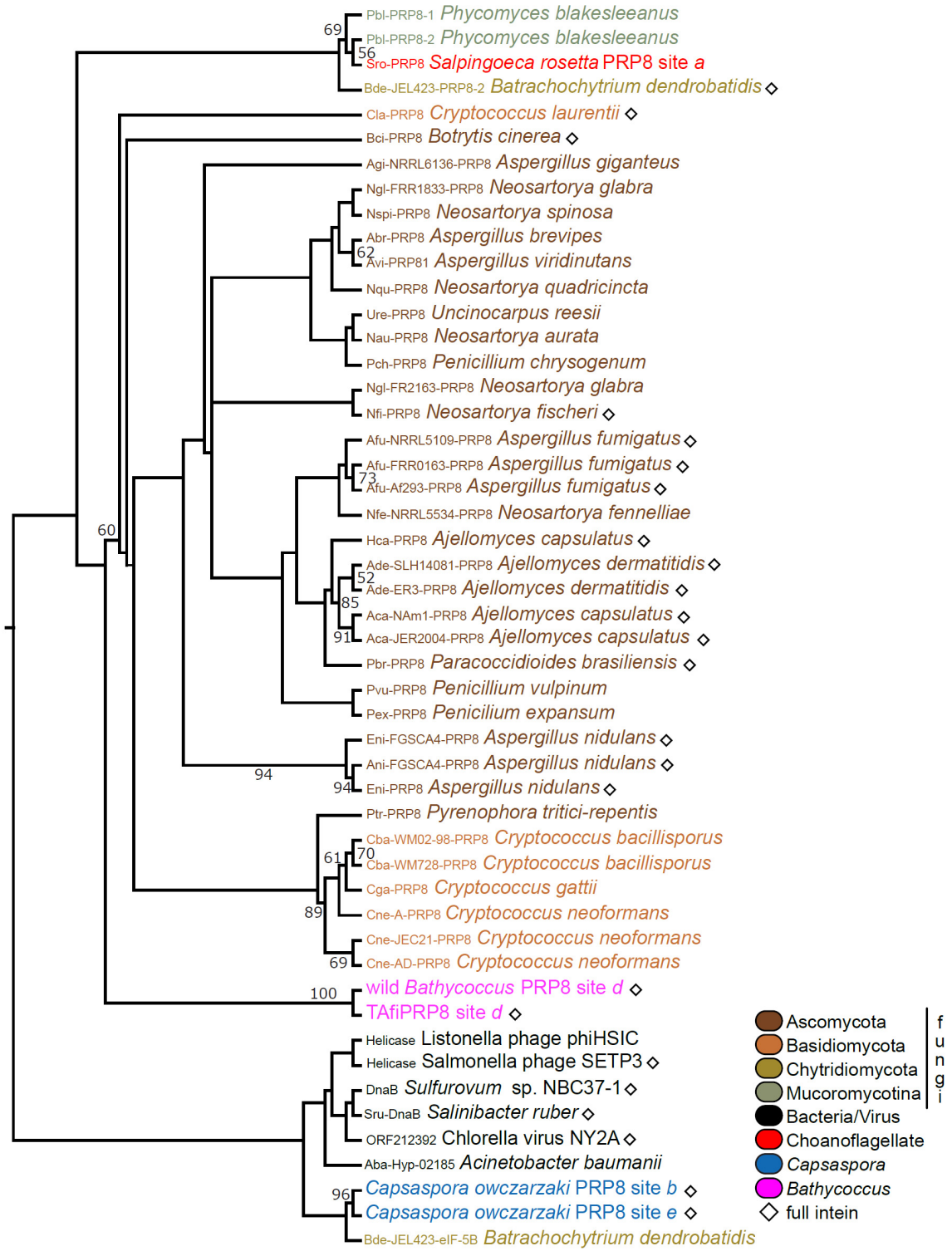


Supplementary Figure 4 - Alignment of putative *prp8* intron sequences. All environmental sequences with intervening sequences suggestive of introns are shown aligned. HF_READ_03445078 represents a 454-metagenomic sequence while all others are the result of PCR and cloning (with Sanger sequencing). Note that HF_READ_03445078 and CNPi1PRP8 appear to be identical assuming that incongruous ‘t’s” in the former represent homo-polymer sequencing issues (a gap or deletion in the Pacific 454-read) which are a well known problem of the 454-sequencing platform.

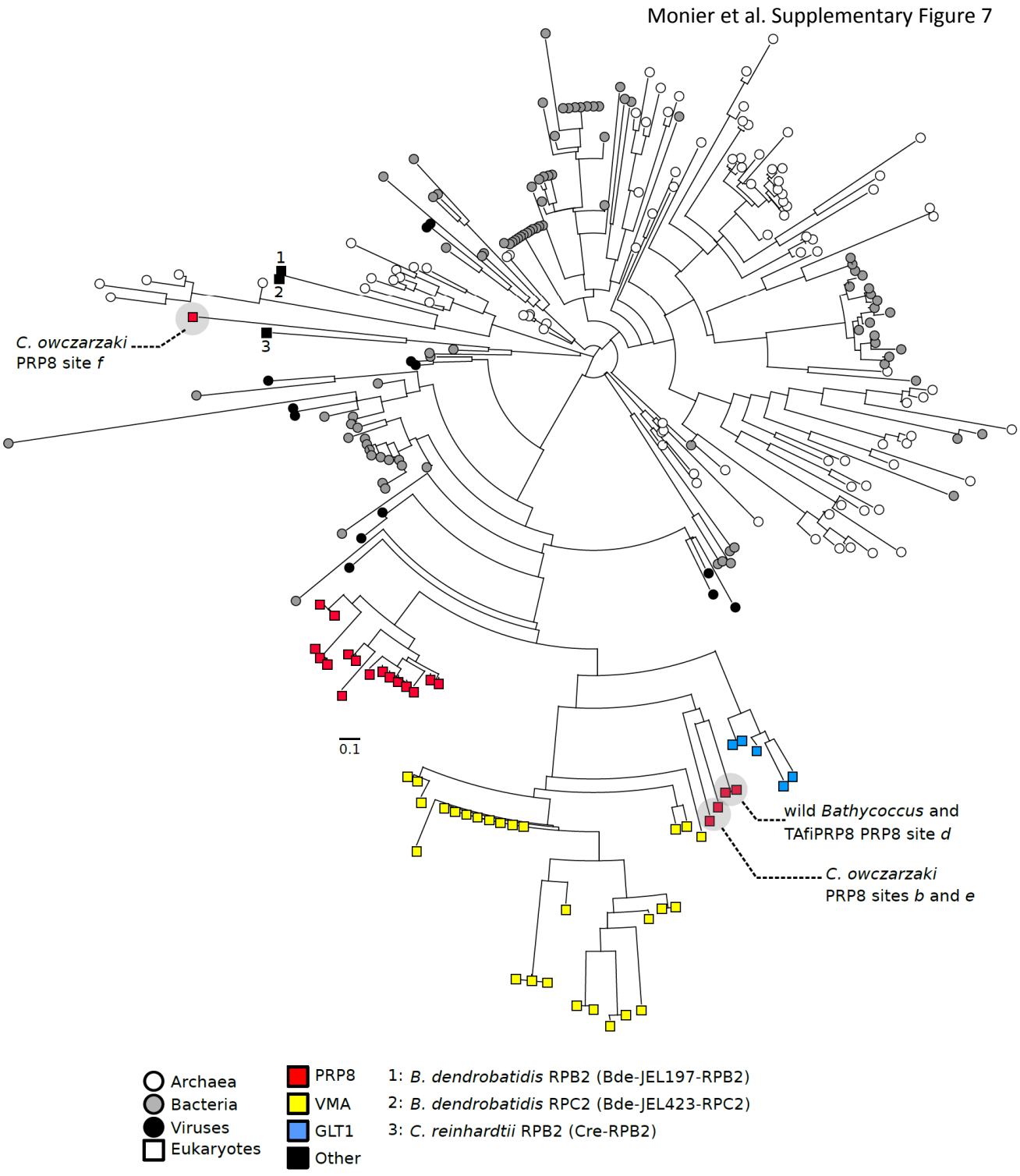
Monier et al. Supplementary Figure 5



Supplementary Figure 5 - Phylogenetic relationships of intein splicing domains. The complete phylogenetic tree displayed as a subtree in Figure 5. *B. dendrobatidis* (BdeJEL423-PRP8-1) and *S. punctatus* (Spu-PRP8), which have PRP8 inteins inserted at sites *b* and *c*, respectively, did not belong to the primary PRP8 clade and are indicated by dashed lines. Note that although the only eukaryotic nucleus-encoded inteins are in fungi and one each in *D. discoidium* and *C. reinhardtii*, there are seven others in eukaryotes but these are in chloroplast-encoded genes (across five algal genera, all within the supergroup Plantae) not nucleus-encoded genes. Note also that intein-coding regions of TAfiPRP8 and the FACS-sorted wild *Bathycoccus* population had 98% nucleotide identity across the entire sequences and both had conservation in all splicing motif and HE blocks (Figure 2). SH-like statistical support as percent of 1,000 replicates was computed but for simplicity shown on the subtree only (Figure 5).

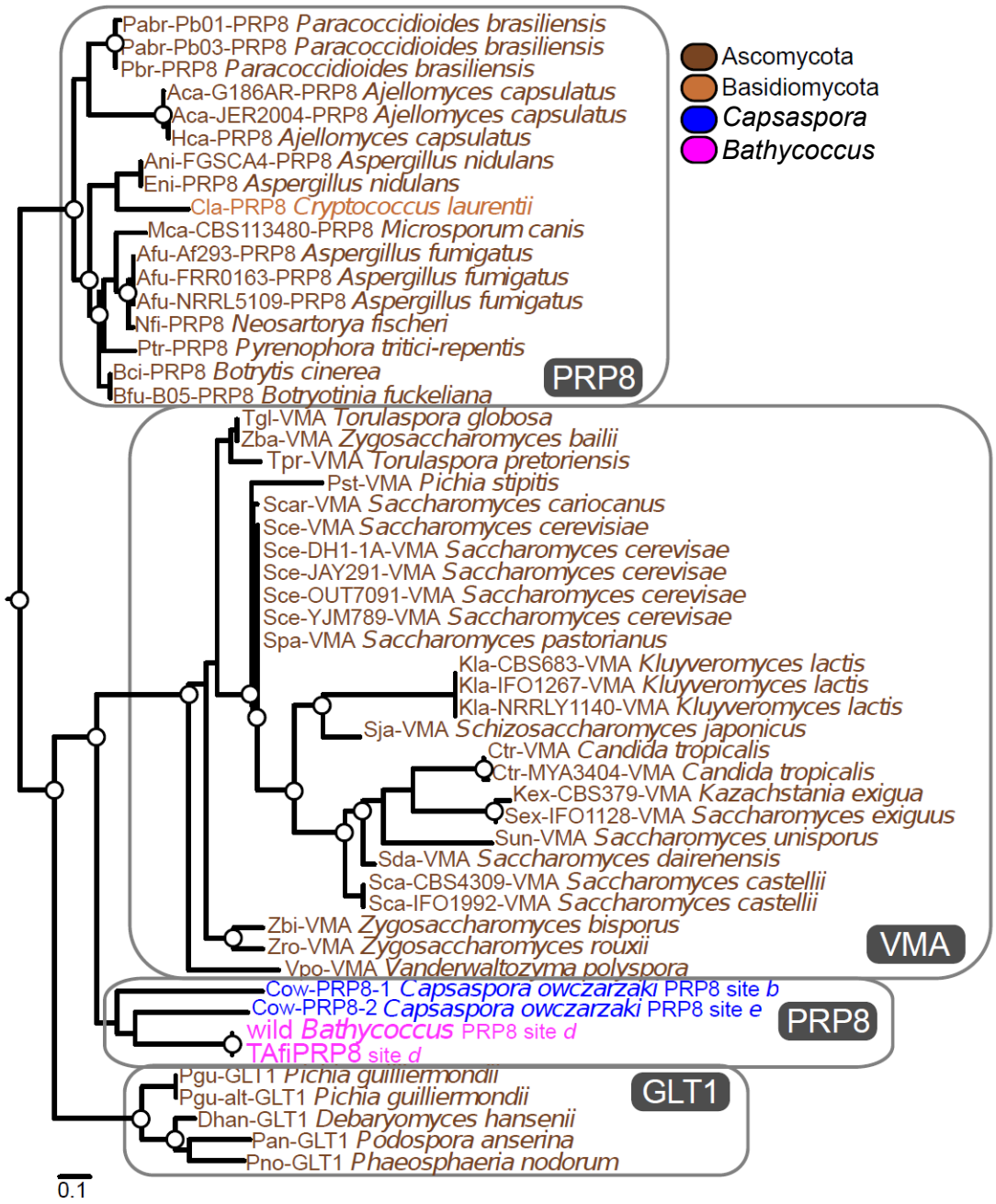


Supplementary Figure 6 - Cladogram displaying the distance-based phylogenetic relationships of PRP8-intein splicing domains, using the same data as for Figure 5 and Supplementary Figure 5. NEB InBase identifiers precede species names. The phylogenetic tree was reconstructed using BIONJ; 100 bootstrap replicates were performed and the consensus tree was generated with PHYLIP. Only nodes with support ≥60% are shown.



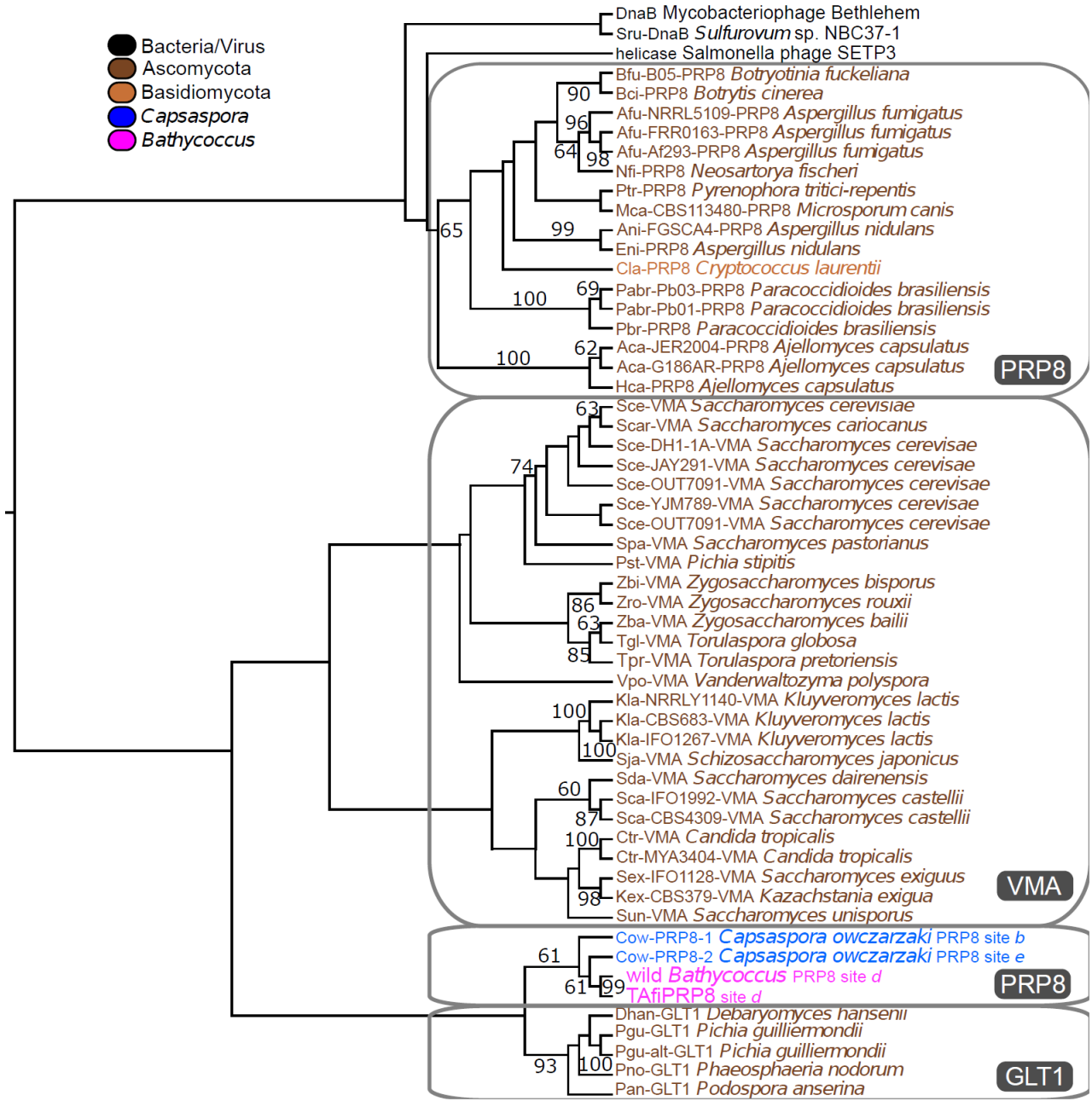
Supplementary Figure 7 - Phylogenetic relationships of intein-located HE domains. The complete phylogenetic tree displayed as a subtree in Supplementary Figure 8. HE genes fall into six structural families (2) and all HEs from nucleus-encoded full-inteins belong to the LAGLIDADG family. Here, we tested evolutionary relationships for all 256 known HE within this family that are in full-inteins, using 49 amino-acid positions from defined HE blocks. The phylogenetic reconstruction shows the eukaryotic full-intein-HEs branching together, even though they are from non-allelic insertion sites, except those of *C. owczarzaki* PRP8 site *f* (Cow-PRP8-3) and three

RNA polymerase full-inteins (numbered 1, 2 and 3). Most grouped according to their host gene, e.g., all HEs from fungal PRP8-inteins branched together (Supplementary Figure 8). However, HEs from *Bathycoccus* and *C. owczarzaki* sites *b* and *e* PRP8-inteins (Cow-PRP8-1 and -2, respectively) grouped together in an unsupported clade that was apart from fungal homologs. HE divergence levels appear greater than for intein splicing domains, as seen previously for fungal PRP8 full-inteins (3). Apart from the outlier PRP8 HEs (in *Bathycoccus* and *C. owczarzaki*), the highest divergence in eukaryotic intein-contained HEs was seen among those from VMA, in agreement with (3). The *Bathycoccus* HE domains appear to be more divergent from other PRP8-encoded HEs than their intein-splicing domains, similar to observations for Ascomycota and Basidomycota (3). The two *C. owczarzaki* intein-encoded HEs (sites b and e) grouped within the eukaryotic intein-encoded HE subtrees (Supplementary Figures 8 and 9). Their corresponding intein splicing domains were outside the PRP8-intein subtree in the approximate-maximum likelihood reconstruction (Figure 5) and in an unsupported position in the subtree constructed using distance-based methods (Supplementary Figure 6). This discrepancy between reconstruction approaches in part reflects the divergence between splicing motifs of these two *C. owczarzaki* inteins and those from other PRP8-inteins.
NEB InBase identifiers precede species names. SH-like statistical support was computed as the percent of 1,000 replicates but is shown only on the subtree (Supplementary Figure 8) for figure simplicity.



Supplementary Figure 8 - This subtree shows HEs from eukaryotic nuclear genes and is derived from an analysis of 256 HEs from full-inteins only (Supplementary Figure 7) and was computed using approximate maximum-likelihood, as implemented in FastTree. Forty-nine amino-acid positions were used. NEB InBase identifiers precede species names. PRP8 inteins are at site *a* unless indicated otherwise in name. Node support ≥75% (white dots) and substitutions site^{-1} (scale bar) are shown.

Monier et al. Supplementary Figure 9



Supplementary Figure 9 - This cladogram shows HEs from eukaryotic nuclear genes and is derived from an analysis of 256 HEs from full-inteins only (Supplementary Figure 7). It was computed using distance methods in BIONJ and 49 amino-acid positions . NEB InBase identifiers precede species names. PRP8 inteins are at site *a* unless indicated otherwise in name. 100 bootstrap replicates were performed and the consensus tree was generated with PHYLIP. Node support is only given when ≥60%.

Additional citations included in supplementary figure legends:

1. Theodoro RC, Volkmann G, Liu XQ, & Bagagli E (2011) PRP8 intein in Ajellomycetaceae family pathogens: sequence analysis, splicing evaluation and homing endonuclease activity. *Fungal Genet Biol* 48(2):80-91.
2. Taylor GK & Stoddard BL (2012) Structural, functional and evolutionary relationships between homing endonucleases and proteins from their host organisms. *Nucleic Acids Res*:1-12.
3. Butler MI, Gray J, Goodwin TJ, & Poulter RT (2006) The distribution and evolutionary history of the PRP8 intein. *BMC Evol Biol* 6:42.