Systematic Biology (2023), **0**, 0, *pp.* 1–63
doi:10.1093/sysbio/output

A step in the deep evolution of Alvinellidae (Annelida: Polychaeta): a phylogenomic comparative approach based on transcriptomes

Pierre-Guillaume Brun[1], Stéphane Hourdez[2], Marion Ballenghien[1], Yadong Zhou[3], Jean Mary[1*] and Didier Jollivet[1*,**]

[1] *Station Biologique de Roscoff, Sorbonne Université-CNRS, UMR 7144, Place G. Teissier, 29280 Roscoff, France*

[2] *Observatoire Océanologique de Banyuls, Sorbonne Université-CNRS, UMR 8222, 1 Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France*

[3] *Key Laboratory of Marine Ecosystem Dynamics, 310000 Hangzhou, China*

**JM and DJ are joint senior authors*

***jollivet@sb-roscoff.fr*

Abstract

Alvinellid worms are a family of endemic and closely related species from deep-sea hydrothermal vents in the Pacific and Indian Oceans. These annelid worms, sister group to the Ampharetidae, occupy a wide range of ecological niches, some of which include the most thermotolerant marine animals described to date such as the Pompeii worm *Alvinella pompejana*, and other species living at much lower temperatures such as *Paralvinella grasslei* or *Paralvinella pandorae*. The phylogeny of this family has not yet been studied extensively. It is, however, a complex case with conflicting molecular phylogenies, the main question being the monophyletic or polyphyletic character of the genus *Paralvinella*. We carried out a comprehensive study of the phylogeny of this family using the best molecular data currently available from RNAseq datasets. The study is based on the assembly of several hundred transcripts for 11 of the 14 species described or in description. The results obtained by the most popular phylogenetic inference models (gene concatenation and maximum likelihood, or coalescent-based methods from gene trees) are compared, and an

© The Author 2023. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

attempt is made to use sequence insertion and deletion information to evaluate and strengthen our choice over the different phylogenies using a newly-developed maximum likelihood method.

Although our study does not allow to definitively assert the phylogeny of the Alvinellidae (three species are still missing), we propose to support the initial hypothesis of the monophyly of the *Paralvinella* proposed by Desbruyères and Laubier on the basis of the morphology of the species, in which the species *Paralvinella pandorae* and *Paralvinella unidentata* are basal in the genus *Paralvinella* and grouped together within the subgenus *nautalvinella*. Following a clock calibration, the radiation of the Alvinellidae, dated between 55 and 78 Myr, took place very rapidly, resulting in high rates of incomplete lineage sorting between the first ancestors and probable gene transfers between the *Alvinella*, *Nautalvinella* and the rest of the *Paralvinella* lineages.

Key words: Alvinellidae, hydrothermal vents, phylogenomics, RNAseq, transcriptomics, indel.

Introduction

Discovered in the late 1970s, hydrothermal vents represent extreme environments with particularly intriguing ecosystems. These habitats, associated with submarine volcanism resulting from the geological activity of oceanic ridges and back-arc basins, offer a large quantity of reduced compounds and energy on which mainly chemosynthetic and highly specialized communities are based (Monaco and Prouzet, 2015). Hydrothermal vents are unstable environments, both in time and space. The point emissions of hot fluid at hydrothermal sites, clustering different chimneys located a few to hundreds meters apart, are synchronized by subterranean fissure networks (Chevaldonné et al., 1997). These emissions, by mixing with cold, deep seawater, produce important environmental gradients,

both physical (temperature) and chemical (pH, oxygen, hydrogen, sulfide), in which dense communities adapted to this diversity of ecological niches develop (Lelièvre et al., 2018). In addition, community networks defined several biogeographic provinces with faunal similarities and transition zones, for example between the Western Pacific and the Indian Ocean, or between the East Pacific Rise and the Indian Ocean Ridge through the linear system of interconnected oceanic ridges which allows for the dissemination of the vent species (Tunnicliffe, 1988; Matabos et al., 2011; Moalic et al., 2012; Nakamura et al., 2012).

After the initial discovery of vent communities on the Galapagos rift and the description of emblematic symbiotic species such as *Riftia pachyptila* or *Bathymodiolus thermophilus* (Corliss et al., 1979), the Pompeii worm *Alvinella pompajena* Desbruyères & Laubier, 1980, was the first chimney species described during another cruise at 21°N on the East Pacific Rise (EPR) (Desbruyères and Laubier, 1980). Collected from the newly-discovered "black smoker" type of chimneys at the 21°N on the EPR, the two ontogenic forms living in association were later found to be two distinct species, *A. pompejana* and *Alvinella caudata* (Desbruyères and Laubier, 1982; Jollivet and Hourdez, 2021). The two *Alvinella* species were initially described aberrant forms in the family Ampharetidae on the basis of retractile filiform buccal tentacles, notopods with simple dorsal setae and ventral uncini tori, although some obvious specificities such as the lack of separation between the head and the rest of the body and modifications of the notopods immediately led the authors to suggest the creation of the subfamily Alvinellinae (Desbruyères and Laubier, 1980). Later, the subfamily Alvinellinae was split into two genera, *Alvinella* and *Paralvinella*, following the discovery of the species *Paralvinella grasslei* Desbruyères & Laubier, 1982, at the EPR and the Guaymas basin. The distinction between the two genera is based in particular on the shape of the mouth apparatus and the shape of the gills, which are lamellar in *Alvinella* and filamentous in *Paralvinella*. New specificities are observed, which further distinguish Alvinellinae from other Ampharetidae: unicini have only two teeth (with the exception of *Paralvinella unidentata*), and a process

of cephalization of the gills is suggested, resulting in the insertion of the first pair of gills on the first asetigerous segment and the 3 other pairs on the fused second and third segments, in contrast to more typical Ampharetidae (Desbruyères and Laubier, 1982).

Subsequently, new species of Alvinellinae were discovered, all endemic to hydrothermal vents in the Pacific Ocean: *Paralvinella pandorae irlandei* Desbruyères & Laubier, 1986 at the EPR and Guaymas basin (Desbruyères and Laubier, 1986), *Paralvinella pandorae pandorae* Desbruyères & Laubier, 1986 (Desbruyères and Laubier, 1986), *Paralvinella palmiformis* Desbruyères & Laubier, 1986 (Desbruyères and Laubier, 1986), *Paralvinella dela* Detinova, 1988 (Detinova, 1988), all these species living at the Juan de Fuca Ridge, *Paralvinella hessleri* Desbruyères & Laubier, 1989 in the Marianas back-arc basins and Okinawa Trough (Desbruyères and Laubier, 1989), *Paralvinella bactericola* Desbruyères & Laubier, 1991 at the Guaymas basin (Desbruyères and Laubier, 1991), *Paralvinella sulfincola* Desbruyères & Laubier, 1993 found on top of vent chimneys on the Juan de Fuca Ridge (Desbruyères and Laubier, 1993), *Paralvinella fijiensis* Desbruyères & Laubier, 1993 also found on similar 'hot' habitats in the south-west Pacific (Tunnicliffe et al., 1993), and finally *Paralvinella unidentata* Desbruyères & Laubier 1993 living in colder habitats of the south-west Pacific (Desbruyères and Laubier, 1993). The progressive species enrichment associated with the geographic sampling of these vent worms and resulting studies combining both morphological and molecular traits to assess their position in the Terebelliformia now demonstrate that the Alvinellidae form a monophyletic group, not nested within the Ampharetidae, which now represents its sister group (Rousset et al., 2003; Glasby et al., 2004; Stiller et al., 2020).

More recently, a new species, *Paralvinella mira* Han, Zhang, Wang & Zhou, 2021, genetically close to *P. hessleri* has been described outside of the Pacific Ocean, from the hydrothermal vents of the Carlsberg Ridge in the Indian Ocean (Han et al., 2021). The alvinellid worms' distribution has thus been expanded beyond Pacific waters, supporting the sampling of a small population of individuals attributed to the genus *Paralvinella*, but

not formally described, observed at the Solitaire field on the Central Indian Ridge (Nakamura et al., 2012). At present, no species of Alvinellidae have been observed neither on the Central Indian Ridge nor at the Southern Indian ridge, although sampling of vent communities has received less attention in this region than in the Pacific and Atlantic oceans (Nakamura et al., 2012; Han et al., 2021). Finally, another building parchment-tube species, whose description is still in progress, was first discovered in the Manus Bassin in 2011, then from the Nifonea vent site along the volcanic arc of Vanuatu. This species is referred to as *Paralvinella* sp. nov. in our article.

Annelids, especially Terebelliformia, show a great morphological diversity and a great variety of ecological niches, with habitats ranging from coastal muddy areas to coral reefs or from cold Antarctic waters to hydrothermal vents (Stiller et al., 2020). In the sister group Ampharetidae, which also co-occurs at hydrothermal vents, shifts toward chemosynthesis-based ecosystems, both from seep to vent and vent to seep environments, has occurred several times independently followed by diversification (Eilertsen et al., 2017). From one shared ancestor, Alvinellidae probably have a long evolutionary history of speciation since their colonization of the deep hydrothermal environment. A molecular dating proposed by Jollivet and Hourdez establishes an origin for the family between 70 and 200 Ma, depending on the genetic markers used and the number of calibration dates available (Jollivet and Hourdez, 2021). This predates the last major anoxic events which affected the deep ocean at the end of the Cretaceous period (Jacobs and Lindberg, 1998) but coincides with fossil records dating from the same geologic period and resembling tubes currently produced by vestimentiferan tubeworms and Alvinnelid worms (Vrijenhoek, 2013).

While Fontanillas et al. hypothesized in 2017 that the ancestor of the Alvinellidae could be a thermophilic species by observing the amino acid composition of ancestral protein sequences (Fontanillas et al., 2017), the family has now diversified to occupy a wide variety of vent habitats, resulting from strong environmental gradients and observable

within the biogenic structures of tube-worm bushes from basaltic diffusing cracks to the tube aperture of sibonoglid worms in the Eastern Pacific or at the edge of the hairy snails beds in the Western Pacific (Lelièvre et al., 2018). Alvinellid worms contain in particular two of the most thermotolerant animals described to date, *A. pompejana* on thewalls of black smokers on the EPR and *P. sulfincola*, its ecological homolog inhabiting vent chimneys walls on the Juan de Fuca Ridge (Desbruyères and Laubier, 1986). Experiments in pressurized aquaria have shown that both species live at temperatures exceeding 40°C, and can sustain temperatures up to 55°C for relatively short periods of time (Girguis and Lee, 2006; Ravaux et al., 2013). In the aquatic environments, only *Potamocypris* ostracods from Oregon hot vents have been described as being able to live at temperatures above 50°C (Chevaldonné et al., 2000). However, other alvinellid species probably live at lower temperatures. *P. palmiformis* appears to thrive at temperatures below 30°C, while the thermal preference of *P. grasslei* is around 15°C (Desbruyères and Laubier, 1982; Girguis and Lee, 2006; Cottin et al., 2008). *P. p. pandorae* and *P. p. irlandei* have even been observed in the coldest parts of hydrothermal sites, where temperatures do not exceed 2-5°C (Desbruyères and Laubier, 1986). However, these species also behave as pioneer species in warmer, newly activated sites (20-25°C) forming dense populations with siboglinid worms such as *Tevnia jerichonana* (Jollivet and Hourdez, 2021). Experiments in pressurized aquaria also show that these animals may exhibit a pronounced eurythermal character, being able to grow over a relatively wide temperature range (Dilly et al., 2012; Jollivet and Hourdez, 2021).

As a consequence of the extreme conditions of the hydrothermal environment, and very steep gradients, Alvinellidae exhibit a range of physiological and functional adaptations, including enlarged gill surfaces allowing to cope with the hypoxic conditions created by the mixing of vent fluid with deep-sea water (Hourdez and Lallier, 2007). They also have respiratory pigments with higher affinities for oxygen, whose affinity is not affected by the presence of sulfide and weakly affected by temperature (Somero et al.,

1989; Hourdez and Lallier, 2007), and an oxygen species removal proteins arsenal that may represent a way to adapt to high temperatures (Mary et al., 2010; Dilly et al., 2012). Alvinellid species also exhibit adaptations interpreted as a direct consequence of the wide temperature range at which they thrive. Species of the genus *Alvinella* secrete a thick tube from their ventral shield that protects them from high temperatures (Le Bris and Gaill, 2007), as does the newly sampled *P.* sp. nov.. The hot species *P. sulfincola* and *P. hessleri* or *P. mira* also produce a thick and mineralized tube, while colder species such as *P. grasslei*, *P. palmiformis*, *P. p. irlandei* produce a thinner mucus cocoon. Enzymes of warm species *A. pompejana*, *A. caudata*, *P. sulfincola* and *P. hessleri*, were less affected by higher temperature exposure than their homologs from colder species (*P. palmiformis*, *P. grasslei*, *P. p. pandorae*, *P. p. irlandei*) (Jollivet et al., 1995). Finally, the over-expression of several proteins of *A. pompejana* has been shown to be more thermostable than their homologs in homeotherms (Chevaldonné et al., 2000).

The diversity of these specific adaptations is particularly intriguing. The question naturally arises of the evolutionary path taken by the alvinellid worms, as well as the different stages that allowed its dispersal and colonization of hydrothermal vents throughout the Pacific to the Indian Ocean, including their putative initial link with the hydrothermal paleocoastal environment. A necessary step in understanding this evolutionary history is the establishment of a robust phylogeny of the family, including all currently known and described species.

Multiple phylogenies of the Alvinellidae have been proposed. Based on the careful observation of morphological characters, Desbruyères and Laubier divided the family into two monophyletic genera, *Alvinella* and *Paralvinella* (Desbruyères and Laubier, 1993). Species of the genus *Alvinella* have gills with flattened-leaf secondary filaments, the anterior fourth and fifth setigers are modified with the insertion of two pairs of hooks, and a pair of thick tentacles is inserted on the buccal apparatus in males. Species of the genus *Paralvinella* have stalked bipennate gills with 2 rows of cylindrical filaments and no

filaments at the tip of the stem, the seventh setiger is modified with a single pair of hooks, and the modified buccal tentacles of males involved in the male-female pairing are different (either three-lobed, coiled and tapered or absent) (Jollivet and Hourdez, 2021). Other non-morphological differences between the two genera are noted: *Alvinella* species secrete thick parchment-like tubes including inorganic material and have epibiotic bacteria, whereas *Paralvinella* species secrete mucus tubes (with the exception of *P.* sp. nov.) or cocoons and are devoid of epibiotic bacteria (Desbruyères and Laubier, 1991).

Recent molecular phylogenies, on the other hand, have cast doubt on this view of two monophyletic genera in the phylogeny of the Alvinellidae. A phylogeny established by Jollivet and Hourdez on the basis of the mitochondrial gene Cox1 proposed that the species *P. p. irlandei* is basal to the family, and that the genus *Paralvinella* is therefore polyphyletic (Jollivet and Hourdez, 2021). This relationship between species was also recently found with high support (posterior probability 0.91) by Stiller et al. on the basis of several hundred orthologous genes derived from transcriptomic datasets (Stiller et al., 2020). However, other molecular data support the monophyly of the genus *Paralvinella*. Thus, the same study by Jollivet and Hourdez proposes a second phylogeny, based on a set of 278 orthologous transcripts, in which the genus *Paralvinella* remains monophyletic and *P. p. irlandei* is only basal to the other species of *Paralvinella* (Jollivet and Hourdez, 2021). This last phylogenetic work was in agreement with an earlier study by Jollivet et al. which proposed, based on enzyme isoforms similarities, that the species *P. p. pandorae*, *P. p. irlandei* and *P. unidentata* form a group within *Paralvinella* and that the genus *Paralvinella* is monophyletic (Jollivet et al., 1995). Interestingly, studies supporting the monophyly of the *Paralvinella* species include data for *P. unidentata*, in contrast to studies suggesting that the species *P. p. pandorae* and *P. p. irlandei* were the first to diverge within the Alvinellidae.

Desbruyères and Laubier had already proposed the grouping of the species *P. p. pandorae*, *P. p. irlandei* and *P. unidentata* within the subgenus *Nautalvinella*, basal

within the genus *Paralvinella*, considering the comb-like insertion of the secondary filaments of the gills as well as the loss of modified tentacles on the buccal apparatus of the males. However, some peculiarities are noted in the species *P. unidentata*, notably the leaf-like shape of the secondary filaments of the gills, similar to those of the *Alvinella* species (Desbruyères and Laubier, 1993; Jollivet and Hourdez, 2021). The phylogenetic relationship between the three *Nautalvinella* species and the *Alvinella* species is therefore a key to understand the phylogeny of the Alvinellidae.

Here, we propose a new molecular phylogeny of the family Alvinellidae based on orthologous genes derived from RNAseq datasets, notably improved by a new sequencing of the species *P. unidentata* and other species from both the western Pacific and the Indian Ocean. The phylogeny therefore includes all the described species of Alvinellidae species, with the exception of *P. p. pandorae*, and the rare species *P. dela* and *P. bactericola* for which no exhaustive molecular data exist to date. This phylogeny is rooted with species of the sister group Ampharetidae, as well as more distantly-related species of Terebellidae and the coastal pectinariid worm *Pectinaria gouldii*, in accordance with the currently accepted phylogeny of the Terebelliformia worms (Rousset et al., 2007; Stiller et al., 2020). This new phylogeny is based in particular on the sequences of several hundred partial cDNAs and their amino acid translation, as well as an attempt to use sequence insertion/deletion events that are inferred from the sequence alignments. Finally, the results obtained by different methods are compared, between a concatenation + Maximum Likelihood approach on the one hand, and a gene trees + coalescent approach on the other hand.

Materials and Methods

Animal Collection, Sequencing and Assembly

Transcripts of *A. pompejana* were predicted from the genome assembled at the chromosomal level by R. Copley (NCBI accession number PRJEB46503) with the Augustus

webserver (Hoff and Stanke, 2013). The software was first trained on cDNA sequences provided by the MPI that contained previous *A. pompejana* EST assemblies obtained from a Sanger sequencing (Genoscope project, see Gagnière et al. for details (Gagnière et al., 2010)). The prediction was performed on scaffolds of more than 60k bases (95.8% of the assembled genome). The resulting coding sequences were filtered with Kraken v.2.0.9 and then Transdecoder v.5.5.0 to ensure data homogeneity among all final transcriptomes.

Transcriptome reads for the species *P. gouldii*, *Anobothrus* sp., *Amphicteis gunneri*, *Amphisamytha calrdarei* and *Hypania invalida* were downloaded from the NBCI SRA database (accession numbers SRR2057036, SRR11434464, SRR11434467, SRR11434468, SRR5590961) (Kocot et al., 2016; Stiller et al., 2020). Total RNA extraction from flash-frozen tissue were performed with Trizol. Libraries of *P. gouldii* were sequences with Illumina HiSeq, libraries of *Anobothrus* sp., *A. gunneri* and *A. carldarei* were obtained with Illumina Hiseq 4000, and Illumina Hiseq 2500 was used for *H. invalida*.

An EST library was previously obtained for the species *P. sulfincola* in the framework of a Join Genome Initiative project (NCBI accession number PRJNA80027), led by P.R. Girguis and S. Hourdez (Girguis and Lee, 2006). The animals were collected on the Juan de Fuca Ridge during the jdFR oceanographic cruise in August and September 2008, with the submersible Alvin on board of the R/V Atlantis. Reads were obtained by 454 Roche technology, and 24,702 transcripts were assembled using Newbler (Margulies et al., 2005).

Other Alvinellid species were collected during several oceanic cruises from 2004 to 2019 using the N/O L'Atalante and either the ROV *Victor6000* or the manned submersible Nautile with the exception of *P. palmiformis*, which was also collected with *P. sulfincola* during the same jdfR cruise on Juan de Fuca. Information about the sampling location of the alvinellid worms collected along the East Pacific Rise are provided in Fontanillas et al. 2017 (Fontanillas et al., 2017). Other alvinellid species and the vent terebellid were sampled on different vent sites of the western Pacific back-arc basins during

the Chubacarc 2019 cruise with the tele-manipulated arm of the ROV *Victor6000* and brought back to the surface in an insulated basket. *P. hessleri* was sampled from Fenway in the Manus basin, *P. fijiensis* from Big Papi (Manus Basin) and Tu'i Malila (Lau Basin), *P. unidentata* from Fenway (Manus Basin), *P.* sp. nov. from the volcano South Su (Manus Basin) and the vent terebellid Terebellidae gen. sp. (not yet described) at Snow Cap (Manus Basin). Finally, *P. mira* was sampled from the Wocan vent field in the northwest Indian Ocean on the Carlsberg Ridge by HOV Jialong during the DY38 cruise in March 2017 (Han et al., 2021). *Melinna palmata* and *Neoamphitrite edwardsii*, which are shallow-water species, were respectively sampled in the bay of Morlaix and Roscoff, France. Total RNA extraction from flash-frozen tissue were performed with Trizol. RNAseq libraries were produced at Genome Québec following a polyA purification of mRNAs, and sequenced using the Novaseq technology.

With the exception of *A. pompejana* and *P. sulfincola* species, all transcriptomes were assembled using a common procedure. The reads were first cleaned of adapters and trimmed with Fastp v.0.20 to retain reads with a phred-score above 25 for each base (Chen et al., 2018). Kraken v.2.0.9 was then used to remove reads corresponding to prokaryotic contamination (Wood and Salzberg, 2014). The reads retained at this stage were assembled *de novo* using the method implemented by Trinity v.2.9.1 (Grabherr et al., 2011). Finally, the assembled transcripts with an identity greater than 99% on 50 consecutive bases were assembled with cap3 v.10.2011 (Huang and Madan, 1999). Results of the sequencing, filtration steps, and assembly metrics are detailed in Supplementary data Table 1, as well as the associated command lines. Transcriptomes are available from the NCBI BioProject repository (*A. caudata* XXX, *P. unidentata* XXX, *P. p. irlandei* SRX1843887, *P. palmiformis* XXX, *P. grasslei* XXX, *P. mira* XXX, *P. hessleri* XXX, *P.* sp. nov. XXX, *P. fijiensis* XXX, Terebellidae gen. sp. XXX, *M. palmata* XXX, *N. edwardsi* XXX).

The Open reading Frames were then determined with Transdecoder v.5.5.0, trained on the 500 longest sequences present in each transcriptome. The predicted regions in the

reading frame were retained, stripped of the 3' and 5' UTR regions. The metrics associated with these different steps are also presented in Supplementary data Table 1.

In total, 25 transcriptomes were obtained, from 19 different Terebelliformia species including 11 species of Alvinellidae.

Search for orthologous genes and Bioinformatic Processing

Orthologous sequences were determined from the assembled transcriptomes using Orthograph v.0.7.2 (Petersen et al., 2017). Following the recommendations of Orthograph, a reference database of 1997 orthogroups (OG) was first constructed from the ODB9 database (Zdobnov et al., 2021). These OGs correspond to single copy genes in all lophotrochozoan species available when downloading ODB9 (*Lottia gigantea*, *Crassostrea gigas*, *Biomphalaria glabrata*, *Capitella teleta*, *Helobdella robusta*). Then, each Terebelliformia transcriptome was blasted against these OGs with Orthograph to identify orthologous genes. These steps attempted to identify all orthologous genes present in Terebelliformia that are expected to be single copies. The results of these ortholog identification steps are detailed in Supplementary data Table 1. We then obtained the following datasets:

1. a set of 966 orthologous genes, in nucleotides, and present in at least 20 transcriptomes (with the constraint that *P. p. irlandei* must be present for each gene);

2. a set of 478 orthologous genes, derived from dataset (1) and translated in amino acids. These genes are found in at least 20 transcriptomes (with the constraint that *P. p. irlandei*, which had the less populated transcriptome must be present for each gene), and were filtered to reduce amino-acid composition biases between species;

3. a set of 895 orthologous genes, encoded in nucleic acids and concatenated into a supergene. This set of genes derives from the set of 966 genes (1), but with additional

constraints. Sequence alignments must contain at least the same gene from 20 transcriptomes, within each alignment sequences shorter than 25% of the total length of the alignment are deleted, and aligned gene fragments of less than 100 nucleotides between two gaps are considered unreliable and deleted. The resulting supergene contains 594,101 aligned nucleotide sites (*P. p. irlandei* being the least represented species with 288,184 sites). Here, we give priority to the quality of alignment and completeness of individual sites over the total number of genes;

4. a set of 385 orthologous genes, translated in amino acids and concatenated into a supergene. This set of proteins derives from the first set of 478 amino-acid sequences (2), with the same additional constraints as for dataset (3), and aligned gene fragments of less than 30 amino acids between two gaps are considered unreliable and deleted. The resulting supergene contains 96,432 aligned residues (*P. gouldii* being the least represented species with 45,945 sites), giving priority to the quality of alignment and completeness of individual sites over the total number of genes;

5. 2 sets of 2,050 and 1416 sequence indels, shared by at least 23 transcriptomes, obtained from the first 478 protein dataset. The set of 1416 indels corresponds to a more conserved subset for which the quality of the alignment is the most reliable.

The nucleotide sequences were aligned with MACSE v.2.05, which is an aligner respecting the codon structure of the coding sequences (Ranwez et al., 2011). The amino acid sequences were aligned with Probcons v.1.12 (Do et al., 2005). The indels were deduced from the alignments obtained by Probcons on the protein sequences.

Phylogenetic inference and molecular dating

Phylogenetic inference using a supergene and the ML approach A global phylogeny was first evaluated from the supergenes of datasets 3 and 4 with the maximum likelihood approach implemented in RAxML-ng v.1.1.0 (Kozlov et al., 2019) with 100 bootstrap

replicates, either under the GTR+Γ model with 4 discrete Γ classes for nucleotide mutation rates, or under the JTT+Γ model with 4 discrete Γ classes for amino acid mutation rates. The choice of the JTT matrix is justified by the fact that this model is the most suitable for 65% of individual gene trees according to a model cross-validation procedure implemented in Phylobayes (see below). The best ML trees were then compared with consensus trees obtained from the single gene coalescence approach described below.

Phylogenetic inference using the gene coalescence approach The choice of the evolutionary model used to establish the amino-acid translated gene trees was made according to the recommendations of the cross-validation procedure implemented in Phylobayes v.4.1 (Lartillot and Philippe, 2004; Lartillot et al., 2007). For each gene:

- the tree topology was evaluated from the run of 2 MCMC chains, until convergence of both chains (effective size > 50, maximum discrepancy < 0.3, which measures the number of true independent samples for each parameter and the similarity of this parameter distributions between the two chains). This exploration of the tree universe was done for a series of evolutionary models involving either empirical mixture models which take into account the differences in equilibrium frequencies across sites, namely C20 (Si Quang et al., 2008) or WLSR5 (Wang et al., 2008)), associated with empirical amino-acid substitution matrices (WAG, JTT or LG), or empirical models based on a mixture of 3 mutation matrices (UL3 (Le et al., 2008)), with 4 discrete Γ classes of mutational rates;
- a MCMC chain was run for each model using an initial fixed topology obtained under that model. Genealogies were saved every 5 cycles, until 500 trees were obtained. This search was performed with 5 replicates to perform a 5-fold cross validation;
- the cross-validation likelihood scores were calculated for each of these replicates by comparing the parameterization obtained for the two chains (burn-in 100) trained on the learning set (80% of the original gene alignment) and its predicting power on the

test set (20% of the original gene alignment);

- the chosen model corresponds to the one with the best likelihood score;

In the end, the model selection procedure was a 5-fold cross validation on 5 replicates, among the models combining C20 or WLSR5 with the matrices WAG, JTT, and LG or the model UL3 (7 different models). The gene topology corresponds to the tree obtained in the first step after convergence of the 2 MCMC chains under the best selected model. Command lines are detailed in supplementary data. The cross-validation procedure being computationally intensive, the selection of models for each gene was done more parsimoniously than the systematic evaluation of each model, according to a hierarchical procedure motivated by the observation that the C20+JTT model is the best model for 3/4 of the genes when all models are exhaustively evaluated on a subsample of 47 randomly chosen genes (see supplementary data Table 2). Thus, the C20+JTT, C20+WAG and UL3 models are first compared. If C20+WAG is the best of the three models, it is then compared with C20+LG. If UL3 emerges as the best of the 3 models, it is compared with WLSR5+JTT (the comparison is potentially continued with WLSR5+LG). Otherwise, C20+JTT is retained as the best model.

The amino-acid encoded genes were then tested to see wether the best evolutionary model was able to describe observed sequences' features (namely substitution rates, homoplasies, divergence, the proportion of invariant sites, and especially the assumption of amino-acid frequency homogeneity which is the invariance of the amino-acid frequency distribution over time). For this purpose, the Phylobayes function *ppred* was used on the initial chain corresponding to the best model, with a burn-in of 100 and the sampling of one genealogy every 5 cycles. Only genes for which the evolutionary model was suitable, i.e. well-describing these features with a $p - value > 0.1$, were retained. These genes form the dataset (2) and (4), which explains why there are fewer genes in these datasets compared to their untranslated counterparts from dataset (1) and (3).

Less care was needed when filtering the phylogenetic signal in nucleotide

alignments. Only the third nucleotide position of each codon was removed to limit homoplasy and phylogenetic signal saturation.

Fitting the most probable alternative tree topologies associated with the deepest nodes to the sequence datasets Fifteen specific trees differing in their topologies at the two deepest nodes surrounding the root of the family Alvinellidae, named T1 to T15, were exhaustively evaluated against datasets 3 and 4 with BaseML (GTR+Γ with 4 discrete classes of Γ) for nucleotide mutation rates, or CodeML (JTT+Γ with 4 discrete classes of Γ) for amino-acid mutation rates (Yang, 2007). The 15 topologies were also evaluated using Phylobayes under the GTR+Γ model with 4 discrete classes of Γ proposed by this software for amino acid sequences. In this case, 4 MCMC chains were run until convergence (maximum discrepancy < 0.3, effective size > 50). The log likelihoods of the trees of each topology were sub-sampled with a burn-in of 200 cycles, one genealogy sampled every 5 cycles. The resulting likelihood distribution was fitted according to a normal distribution.

These topologies were also evaluated using the coalescent method for datasets 1 and 2. Gene trees were obtained with RAxML-ng (GTR+Γ for nucleotide sequences and JTT+Γ for amino acid data, 100 bootstrap replicates, from 15 random starting trees). The species trees were then evaluated with MP-EST v.2.0 (Liu et al., 2010). The gene trees obtained with Phylobayes from the amino acid sequences are not necessarily binary, and cannot be processed with MP-EST. For trees produced by Phylobayes, we used ASTRAL v.5.7.8 (Zhang et al., 2018) which can use non-binary trees. Branches with low bootstrap support (¡0.1) could also be contracted beforehand to create a polytomy, according to the recommendations of the coalescent-based species tree software ASTRAL.

Estimating the age of the radiation event that led to the diversity in the family Alvinellidae The most probable alvinellid phylogeny was dated with the software Phylobayes. The uncorrelated Gamma model (Drummond et al., 2006), Cox-Ingersoll-Ross (CIR) model (Lepage et al., 2007) and the log normal model (Thorne et al., 1998) were

tested on the dataset 4. Considering that the oldest known Terebelliformia fossils date from the Carboniferous/Devonian time (Sepkoski Jr., 2002), the age of the tree's root (divergence between *P. gouldii* and other terebellid, ampharetid and alvinellid species) is given a Gamma distribution *prior* with a mean of 330 million years and a standard deviation of 200 million years. This places 95% of the distribution between 300 and 360 million years ago. The two calibration points then used within the Alvinellidae correspond to the divergence between the species *P. palmiformis* and *P. grasslei*, estimated 23 to 34 millions of years ago (Ma), as a consequence of the subduction of the Farallon plate under the North American plate (Tunnicliffe, 1988)), as well as the divergence between the two geographic forms of the species *P. fijiensis* from the Manus and Lau basins, estimated to first expand between 2 and 4 Ma (Desbruyères et al., 2006; Boulart et al., 2022).

Phylogeny based on indel data

The indel data (insertion-deletion) were obtained from the translated gene alignments into amino acids according to Probcons. Indel are defined as a position or a set of positions where one or more amino acids is/are absent in at least one of the aligned sequences (a '-' character in the alignment). In addition, indels must be bound on both sides by homologous amino-acid sequences. This excludes leading and trailing gaps of each aligned gene, located in the 5' and 3' UTR (Simmons and Ochoterena, 2000). To this extent, only indels defined with at least 23 transcriptomes were retained.

To extract the phylogenetic information from indels, a Maximum Likelihood model was used to assess the different tree topologies given the probability of observing indels patterns between species. We make two important assumptions (Simmons and Ochoterena, 2000):

- For a given topology, if a residue or a set of residues are present in two different sequences, these residues must be present in the last common ancestor protein sequence (principle of character homology);

- A residue lost during the evolution of proteins is lost forever (a subsequent insertion does not recover the homology). Therefore, if a residue is present in one ancestor and its subsequent descendants, it must be present in all intermediate ancestors.

The model, therefore, is not reversible and distinguishes between insertions and deletions through the presence of an outgroup in the evaluated topology. The probabilities of occurrence of an insertion or deletion are described by a power law (Gu and Li, 1995; Fletcher and Yang, 2009).

The maximum likelihood model is optimised according to 7 parameters:

- r, $l_1(0)$ and $l_2(0)$, which are the ratio between the instantaneous probabilities of insertions against deletions, and the instantaneous size distributions of insertions and deletions;
- T, a scalar applied to the branch lengths;
- I, which is the proportion of invariant sites;
- σ, analogous to the discrete Γ law used to described variable site rates;
- π, the rate of change of a class along σ.

In addition, the model is optimised on two datasets corresponding to different levels of confidence in the alignment of gaps, considering that alignment errors can distort the information in phylogenies established from indels (Donath and Stadler, 2018):

- a "soft filter" dataset in which only indels of 10 amino acids or less explained by a maximum of 4 insertion or deletion events are retained. In addition, the event/length ratio is limited to 2. This set contains 2050 indels;
- a "hard filter" dataset in which only indels of 5 amino acids or less explained by a maximum of 1 insertion or deletion event are retained. This set contains 1416 indels.

The details of the model equations and parameters are presented in the supplementary data.

Gene trees and number of informative sites

Usually gene tree topologies may diverged from each other and from the species tree depending on the specific history of these genes. The aim of this process is to evaluate the potential discrepancy between gene and species topologies. To do this, we compare the likelihood scores of the 15 most relevant topologies at the level of each gene. The orthologous genes previously used in the gene coalescence approach were sub-divided into categories according to their informativeness in order to test whether the low or very high number of informative sites at a given gene could bias the branching of some specific lineages at the deepest nodes of the consensus tree.

The informativeness of the genes is estimated from the number of informative sites that specifically distinguish species associated with the deepest branches in the alvinellid phylogeny: i.e. *A. pompejana* +*A. caudata*, *P. p. irlandei* and *P. unidentata*. A given site, nucleotide or amino acid, is considered informative if it exhibits at least two different states (not a gap) between these 4 species. In the case where the site is present in 3 different states between *Alvinella*, *P. p. irlandei* and *P. unidentata*, it is mandatory that one of these three states is also present in at least one species outside this group of targeted species. A site is therefore informative as soon as the tree topologies have different likelihoods at this site.

For each gene, the likelihoods of either the coding or protein sequence alignments for 15 alternative tree topologies were obtained with RAxML, either under the JTT+Γ or GTR+Γ model with 4 discrete classes of Γ and subsequently ranked. Because likelihoods are highly variable depending on the length of alignments, they must therefore be normalized. To do this, we first assimilate the likelihood of a topology to the probability of occurrence of this topology given all constrained parameters (branch lengths and Γ). We also consider that the set of possible tree topologies is restricted to the n topologies tested. Therefore, if P_{T_i} is the probability of occurrence of a given topology i , we want $\sum_{i=1}^{n} P_{T_i} = 1$ and therefore, if the Log likelihood of T_i is $L_{T_i} = log(P_{T_i})$, we should have

$\sum_{i=1}^{n} e^{L_{T_i}} = 1$

For a topology k, we can deduce that its probability of occurrence is thus $P_k = e^{L_{T_k}} / \sum_{i=1}^{n} e^{L_{T_i}}$ given that the true topology has to be one of the n tested topologies.

To avoid underflow problems due to very small likelihoods, we instead preferred to calculate this probability using a likelihood difference in which

$$P_k = e^{L_{T_k} - L_{max}} / \sum_{i=1}^{n} e^{L_{T_i} - L_{max}}$$

where L_{max} is the maximum likelihood obtained for this gene under the most likely topology. If $L_{T_i} - L_{max} < -30$, we consider that $e^{L_{T_i} - L_{max}} \approx 0$. This transformation of the likelihoods into probabilities of occurrence across n topologies allows to provide the same weight to each gene in the tree comparisons.

We have later limited this approach by considering that only the three most relevant topologies. In this case, the calculation is performed by summing only on the reduced topological space for likelihoods obtained with RAxML on both protein and coding sequences. If the choice between topologies is further reduced to only two possibilities for a given gene, then the choice between the two alternative topologies will follow a Bernoulli distribution of parameter 0.5 if randomly drawn. If the choice is repeated identically and independently for g genes, we can therefore test the bias in the choice of one topology over the other by confronting it with a statistics that follows a binomial distribution of parameters $(g, 0.5)$.

Statistic tests on the significance of the topology choice

The gene-to-gene evaluation of the tree topology based on a likelihood-constrained approach with a reference selected among 15 prior topologies also allows the reconstruction of the species tree by a coalescent approach. Indeed, methods for estimating species trees by coalescence are sensitive to errors in the estimation of individual gene trees (GTEE) (Molloy and Warnow, 2018). We therefore evaluated species trees with ASTRAL from gene trees optimized by maximum likelihood among our 15 user (reference) topologies, taking

into account the uncertainty in the estimation of these gene trees.The underlying assumption of this study is that imposing a series of topologies derived from well-resolved groups of species defined from the initial species tree strongly reduces the overall GTEE at problematic nodes, and more specifically at the root of Alvinellidae. This also prevents likelihoods from converging to outlier trees with aberrant grouping of species at the gene level. We therefore assume that outside the root of Alvinellidae, incomplete lineage sorting (ILS) or gene flow events are negligible in the evolutionary history of species divergence.

Using either the coding (dataset 1) or the protein (dataset 2) sequence alignments, the evaluation is performed 100 times from a set of randomly resampled genes (100 bootstraps). The tree topology uncertainty at the gene level was taken into account with the previously-described procedure: for each gene, 100 trees are drawn from our set of 15 alternative trees with respect to their likelihoods. For example, a tree topology whose likelihood is 8 times higher than another one will be drawn in average 8 times more than this second topology. In the analysis, each gene is therefore weighted at a value of 100, and the distribution of topologies corresponds to the relative likelihood score of the topologies.

The species trees are evaluated by ASTRAL, which gives a normalized quartet score reflecting the agreement between the gene trees and the species tree. To simultaneously compare the scores obtained with dataset 1 and 2, we calculate the Euclidean distance of each topology from the point (1,1), which would correspond to an ideal topology maximizing both the quartet scores for the amino-acid data and for the nucleotide data. The performance of the best ranked topology (TX) averaged on all bootstraps is then tested against each other topology. For this purpose, for each bootstrap replicate, the difference between the distance of the tested topology with TX is calculated. The difference distribution is then fitted to a Gaussian distribution to calculate the probability that the difference is positive (probability that TX is further away from (1,1) than the tested topology given a randomly drawn gene set).

Results

Phylogenetic inference A first phylogeny of the family Alvinellidae has been established based on the whole transcriptome dataset from alignments of either nucleotide or amino-acid sequences, according to the two methods of phylogenetic inference: either concatenation of the coding sequences (CDS) and maximum likelihood, or coalescence from single-gene trees.

Both methods consistently group our terebellomorph species into 3 main clades (Fig. 1), which are the Alvinellidae family, the Ampharetidae family with *Anobothrus* sp., *H. invalida*, *A. carldarei* and *A. gunneri*, and a third group composed of the Terebellidae species (Terebellidae gen. sp. and *N. edwardsi*) as well the Melinnidae species (*M. palmata*). This suggests that the Melinnidae family is closer to the Terebellidae than to the Ampharetidae. This placement is found with all methods with a fairly strong support (bootstrap values of 85 and 90 for respectively nucleotide or amino acid sequences when using the maximum likelihood method). In contrast, the placement of the species *Anobothrus* sp. in the family Ampharetidae is more uncertain. While methods using protein sequences place the species at the root of Ampharetidae with high support (82 bootstrap for concatenated-genes-ML method, and 100 for the single-gene coalescent-based method with MP-EST), methods using nucleotide sequences strongly disagrees with this topology (bootstrap 13 and 0 respectively) and traces the divergence of *Anobothrus* sp. back to before the separation of the families Alvinellidae and Ampharetidae (see supplementary data Fig. 1). For the following, however, we chose to present the protein-derived phylogeny that groups *Anobothrus* sp. with the Ampharetidae, as supported by morphology and other studies on the phylogeny of the Terebellomorphs and Ampharetidae (Eilertsen et al., 2017; Stiller et al., 2020).

Within the Alvinellidae, all methods give a maximum bootstrap value to the grouping of the 2 *Alvinella* species. In the other genus, with the exception of *P. unidentata* and *P. p. irlandei*, the topology of the other species of *Paralvinella* is also resolved

identically regardless of the method used with a bootstrap value of 100. On the other hand, the placement of the species *P. unidentata*, and particularly *P. p. irlandei*, remains uncertain. Depending on the number of genes considered, the evolutionary method used or the phylogenetic inference method employed, the placement of these two species in relation to the group of the other *Paralvinella* species as well as to the genus *Alvinella* is fluctuating at the root of Alvinellidae with two main alternative trees shown in Figure 1a. and 1b. For this reason, we have chosen to address more specifically the resolution of these deep nodes with a topology fixed for the other species, as it is well resolved for the rest of the tree.

Evaluating bifurcations at the root of Alvinellidae In order to solve the Alvinellid topology, we exhaustively compared the phylogenetic results obtained for all possible root topologies, given the outgroup subtree (Terebellidae+Ampharetidae+*P. gouldii*) which remains quite stable (except the position of *Anobothrus* sp.) regardless of the used method. Figure 2 lists all these topologies, numbered from 1 to 15, in which the subtree containing the *Paralvinella* species (without *P. unidentata* and *P. p. irlandei*) are considered resolved and constrained.

The results of the different inference methods on the different datasets (CDS or translated proteins) are presented in Table 1. From these analyses, three topologies stand out:

- Topology 6, in which the species *Alvinella* diverge at the family root, and all the *Paralvinella* species form a monophyletic group within which *P. p. irlandei* and *P. unidentata* are sister species. This topology is strongly supported by maximum likelihood methods on concatenated nucleotide data (column 1). The ML method on amino acid sequences also ranks this topology among the best (third best after T7 and T8 according to CodeML, column 2) and T6 also represents the best topology according to Phylobayes (column 3), although the 95% confidence interval on likelihoods sampled by Phylobayes is not able to separate T6 from the alternative

topologies T7, T8, T9;

- Topology 9, which places *P. p. irlandei* at the root of Alvinellidae, and groups together the species of *Alvinella* with *P. unidentata*. This topology is more supported by the single-gene coalescent methods (columns 4 and 5). T9 is also the third best topology according to ASTRAL, after T1 and T14 (column 6). Finally, T9 is the best topology according to "soft filtered" indel data (column 7) and the fourth best for "hard filtered" indel data (column 8, on par with T1);
- Topology T1, which groups on one side *Alvinella*+*P. unidentata*, and on the other side *Paralvinella*+*P. p. irlandei*. T1 obtains the best result according to ASTRAL (column 6, although the optimization of branch length by ASTRAL proposes a length of 0.03 coalescent unit for the branch leading to the *P. p. irlandei* +*Paralvinella* group, which makes the topology very close to T9) and represents the fourth best result according to the 'hard filtered' indel data.

The topologies T10, T11 and T12, which place *P. unidentata* at the root of the family, are all strongly rejected by all methods. Interestingly, the evaluation of the trees according to ASTRAL and the GAP ML 'hard filter' method gave similar results, where topologies T13, T14 and T15, which cluster the species of *Alvinella* with *P. p. irlandei* and *P. unidentata* and thus promote the polyphyletic status of the alvinellid worms, perform well, in conjunction with topologies 1 and 9. However, as topologies 13, 14 and 15 are clearly rejected by the other methods, we can consider that these topologies are not relevant, and that T1 and T9 should represent the best topologies selected by the GAP ML "hard filter" method. We also tested the GAP ML method with a soft filter on the data, where all indels of less than 40 amino acids and for which up to 10 insertion and deletion events are allowed (data not shown). This dataset of a greater number of indels (2500) ranks the topologies in the following order: T13>T6=T1=14>T15>T9. A larger dataset therefore ranks T1 and T6 better than T9. However, this dataset is arguably much noisier. In the lack of a method to properly discriminate the quality of gap formation in

the alignment, we rather should favor a more conservative approach that increases the quality of the indel alignment, at the expense of the phylogenetic signal.

In conclusion, if we restrict ourselves to the comparison of T1, T6 and T9 topologies, T6 is the topology with the best likelihood values with methods using a set of concatenated genes for both nucleotide and amino acid sequences, while T1 and T9 perform better with single-gene coalescent inference method. In the single-gene approach, T9 performs better with MP-EST, while T1 works better under ASTRAL. The indel approach promotes either the topology T6 or T9 depending on the number of indels used ('soft' vs 'hard' filtering).

Indel-based phylogeny Table 2a. shows the results of the first round of parameter optimization for r, $l_1(0)$ and $l_2(0)$, on the 15 tested topologies. As the topologies are very similar, the parameters achieved are very close from tree to tree. However, these parameters are strongly influenced by the filter applied on the indel data before the analysis. Increasing the number of indels increases the insertion/deletion ratio, which varies from 0.35 to 1.44, and tends to decrease the average length of indels (although the average length of insertions is systematically higher than that of deletions). These observations can be easily explained by the hypothesis of site homology that is used in the model: amino acid residues that are present at the same position in two contemporary species are considered homologous, and must be present in all the ancestors of these species until their last common ancestor. Therefore, the inclusion of numerous but less reliable sites tends to increase the number of false homologies in alignments (false gaps), implying that the model must accommodate a larger number of deletions to explain the loss of amino acids in the present-day species. In addition, by considering longer sites, the model must mechanically increase the average length of insertions and deletions.

For the second round of optimization, we optimized the likelihood by constraining r, $l_1(0)$ and $l_2(0)$ to the values obtained under the "soft filter" in Figure 2a. These intermediate values are close to the empirical assessment made by Gu, Li and Cartwright

in humans and rodents (Gu and Li, 1995; Cartwright, 2009). The fixation of parameters to constant values at this stage implies that the likelihood differences obtained in a second step between the three topologies do not depend on the best evaluation of these parameters, which are difficult to estimate as they are not described by analytical functions. It should be noted, however, that repeating the likelihood optimisation with the other parameter sets did change the absolute likelihood, but not the relative likelihoods. Thus, the observed differences in likelihoods between the topologies remain true with the other values for r, $l_1(0)$ and $l_2(0)$ summarized in 2a.

The results of this second optimization are presented in Table 2b., for both the 'soft filter' ($L \leqslant 40$, $n \leqslant 10$, $n/L \leqslant 2$) and the 'hard filter' datasets ($L \leqslant 5$, $n \leqslant 1$, $n/L \leqslant 1$). For the three most probable topologies, the optimisation gave satisfactory results where the estimated parameters have very similar values under each filtering condition, which is expected considering that the differences between the topologies are small. For the 'hard filter' condition, the branch lengths obtained are very short, on the order of 10,000 times smaller than the branch lengths in "number of expected amino acid substitution" units. Only the parameter π does not converge to a single value, which is explained by the zero dispersion of indel evolutionary rates ($\sigma = 0$). Under the soft filtering conditions, the branch lengths of the tree drastically increase, with the parameter T multiplied by 5 compared to the hard filter. Moreover, 92% of the sites are considered invariant, which implies that the indel variable parts of the coding sequences experience about one indel every 150 substitutions. Finally, the σ parameter converges to a value around 1.2. The evolutionary rates of the sites are therefore modulated by a factor of 10^x, where x follows a normal distribution of parameters $(0, 1.2)$, which is in practice a dispersion of the rates between $\times 0.01$ and $\times 100$. This implies that he fastest-evolving regions of transcripts, probably the least constrained N-terminal and C-terminal tails, have similar branch lengths to those obtained from the amino-acid substitutions.

For both indel filtering conditions, the T1 and T9 topologies obtain better

likelihood scores compared to T6. These two topologies cluster the species *A. pompejana*, *A. caudata* and *P. unidentata*, and promote the polyphyletic status of the genus *Paralvinella* when compared to the T6 topology which instead clusters *P. unidentata* with *P. p. irlandei* at the base of the genus *Paralvinella*.

Finding the most probable topology out of T1, T6 and T9 We then focused on the evaluation of the three topologies that appear to be the most relevant, namely T1, T6 and T9 to solve the phylogeny of the alvinellid worms.

On one hand, topologies T1 and T9 cluster the two *Alvinella* species with *P. unidentata*. In one case, *P. p. irlandei* clusters with the other *Paralvinella* (T1), while in the other case (T9), *P. p. irlandei* is the first species to diverge within the family, leading to the polyphyletic status of the genus *Paralvinella*. On the other hand, T6 separates the family into the two previously-described genera *Alvinella* and *Paralvinella*, and clusters the species *P. unidentata* and *P. p. irlandei* as expected from the morphology (i.e. subgenus *Nautalvinella*).

In order to evaluate these three topologies, we estimated their likelihood on each gene individually, either on their coding nucleotide sequences (GTR+Γ model) or their translated protein sequences (JTT+Γ model). As shown in Figure 3a. and 3c., most genes are in agreement with the topologies T6 and T9, with a highest number of genes in favor of T6. The choice between these topologies does not seem to depend on the number of informative sites for the nucleotide sequences (Fig. 3b.). For proteins, it is however possible that the number of genes in favor of T6 slightly increases with the number of informative sites when compared to T9, but this does not have a strong influence either (Fig. 3d.). Thus, the conflicting phylogenetic signal between the most probable topologies is not due to a lack or loss of the phylogenetic signal at some slow-evolving genes for the resolution of the ancestral node in the Alvinellidae.

Considering that topologies T1 and T9 both cluster the species *Alvinella* and *P. unidentata* together, we restricted the final analysis to the comparison of topologies T6

and T9. This is because we have a joint phylogenetic signal favoring T1 and T9, and the removal of the T1 topology from the analysis tends to overall reassign a greater number of genes to T9. The T6 and T9 topologies are shown in more detail in Figure 1, where the branch lengths are optimized with CodonML under the JTT+Γ model. In both scenarios, the radiation of Alvinellidae is a fast event, considering that the branches separating the *Alvinella*, *P. p. irlandei*, *P. unidentata* and the other *Paralvinella* are of the order of 0.01, leading to very few diagnostic mutations associated with their separation.

Even in the latter case, T6 remains the most suitable topology for most genes. For coding sequences of transcripts, T6 was the best topology for 509 genes, compared to 453 for T9 ($p - value = 0.038$). In the case of their translated proteins, T6 was also the best topology for 255 genes, compared to 211 for T9 ($p - value = 0.023$).

Therefore, although the coalescent methods involving MP-EST and ASTRAL, and the indel method preferred T1 and T9 topologies, it appears that most genes better fit T6 if we constrain the evolutionary model of the genes to one of these 3 topologies first. T6 is therefore the preferred topology, both in the case of gene concatenation with free topology optimization, and in the case of individual gene trees if we restrict phylogenetic exploration to the most likely topologies.

In a final approach, we attempted to reconstruct the species tree using the coalescent method from gene trees constrained to one of the 15 topologies. The idea was to reduce the gene tree estimation error (GTEE) by assuming that incomplete lineage sorting (ILS) and inter-species gene flow were negligible at all nodes of the tree with the exception of the two incriminated nodes at the root of the family Alvinellidae because of the initial fast radiation of the family. As shown in Figure 4a., Topology T7 had on average the highest quartet scores bot for nucleotide-encoded genes and their translated amino-acid counterparts, although this result was sensitive to the resampling of genes. Figure 4b. displays the combined quartet scores obtained by each topology as the distance to an hypothetical ideal topology that would maximize the quartet scores for both types of

sequences. In this representation, T7 obtains the best combined score, although not statistically better than T4, T5, T6 and T9 ($p - value = 0.14, 0.05, 0.16, 0.06$ respectively). Again, two types of topologies stand out, where *P. p. irlandei* is paraphyletic with the other *Paralvinella* species (T7), or monophyletic (T4, T5, T6). Compared to previous analyses, the T9 and T1 topologies are less supported, although T7 was only significantly better than T1 ($p - value = 9.4 \times 10^{-3}$) but not T9. As before, we tried to restrict single-gene topologies to T4, T5, T6 and T7. Pairwise comparisons consistently concluded that T6 was better supported over the other topologies (proteins: T6, 278 genes *vs.* T4, 156 genes, $p - value = 0$; T6, 277 genes *vs.* T5, 193 genes, $p - value = 6 \times 10^{-5}$; T6, 259 genes *vs.* T7, 192 genes, $p - value = 9 \times 10^{-4}$; nucleotides : T6, 569 genes *vs.* T4, 391 genes, $p - value = 0$; T6, 526 genes *vs.* T5, 441 genes, $p - value = 3 \times 10^{-3}$; T6, 531 genes *vs.* T7, 432 genes, $p - value = 8 \times 10^{-4}$). Genes for which two topologies have the same likelihood are excluded from the comparisons.

Molecular dating of the alvinellid radiation In a final step, we made a first attempt to estimate the age of the alvinellid worm's radiation, under the hypothesis that the T6 topology was true. We used two calibration points corresponding to the subduction of the Farallon plate under the North American plate (Tunnicliffe, 1988), between 24 and 34 Ma, which separates the species *P. grasslei* and *P. palmiformis*, as well as the opening of the Manus and Lau basins between 2 and 4 Ma, which separates two geographic and divergent individuals of *P. fijiensis* (Desbruyères et al., 2006; Boulart et al., 2022). The root of the tree, which corresponds to the radiation of the Terebelliformia, was obtained from a *prior* distribution with a mean 330 Ma and a 200 Myr standard deviation, according to the age of the first fossils attributed to the order Terebelliformia (Sepkoski Jr., 2002).

The dating was performed according to three models: log Normal (Thorne et al., 1998), CIR process (Lepage et al., 2007) and uncorrelated Gamma multipliers (Drummond et al., 2006) under Phylobayes. The three models give different results for the age of the radiation, ranging from 40 to 100 Ma. However, the log Normal model appears to

significantly underestimate ages, since the divergence *Amphisamytha* and the *Hypania* species is estimated at 49 Myr, which appears low compared to the estimated age of the *Amphisamytha* radiation around 50 Ma (Stiller et al., 2013). Furthermore, the root of the tree is 67 millions of years (Myr) old, which is far from our expectations based on the age of the Terebellomorph fossils, dating from the Devonian-Carboniferous period (Sepkoski Jr., 2002). On the other hand, the uncorrelated Gamma multipliers model, which assumes that the evolutionary rates between tree branches are uncorrelated with time, places the age of the tree's root at 299 Myr (123-563 Ma), which is closer to the expected age. The radiation of the Alvinellidae is estimated 101 Ma (49-193 Ma), and that of the *Paralvinella* genus around 92 Ma (45-177 Ma). Yet the 95% confidence intervals are very large with this model, and the estimated ages are far from the ages obtained under the *prior* without calibrations which dates the root 74 Ma (between 43 and 193 Ma). At last, the CIR model proposed the most consistent estimates that are presented in Figure 5. The age of the root is 133 Myr (113-159 Ma). The age of the Alvinellidae radiation stands at 65 Ma (55-78 Ma) and the radiation of the *Paralvinella* lineage is estimated to be 64 Ma (54-76 Ma). These ages are also in good agreement with the *prior* run without calibrations (age of the root ranged between 34 and 199 Myr). In these two last models, the radiation of the Alvinellidae was a rapid event, probably occurring in the Late Cretaceous within a few million years.

Discussion

Monophyly of the genus Paralvinella

A key question in this phylogeny is whether the genus *Paralvinella* is mono- or polyphyletic. Indeed, the initial description of the family by Desbruyères and Laubier subdivided the family into two genera, mainly based on the shape of the gills (4 pairs of lamellar gills in *Alvinella*, versus simple cylindrical gills with no filaments at the tip in

Paralvinella), the position and shape of the first ventral uncini (on segment 9 in *Paralvinella* or 9 and 10 in *Alvinella*) and the position of the first setiger (segment 3 in *Paralvinella* and 6 in *Alvinella*), as well as the presence of a digitiform lobe on the notopodia (Desbruyères and Laubier, 1986; Jouin and Gaill, 1990; Jollivet and Hourdez, 2021). The authors therefore proposed a taxonomic arrangement in which *P. unidentata* and *P. pandorae* are two sister species, basal to the *Paralvinella* lineage (Desbruyères and Laubier, 1993). Interestingly, this proposal is consistent with the T6 topology, which is well supported by our set of concatenated transcripts, as well as individual gene trees constrained among the most relevant topologies. The dichotomy between the genera *Alvinella* and *Paralvinella* was also suggested Jollivet et al. who proposed the monophily of *Paralvinella* on the basis of 14 allozymes or a set of 278 concatenated orthologous sequences (Jollivet et al., 1995; Jollivet and Hourdez, 2021). Interestingly, these two articles, like this study, include sequences for the species *P. unidentata*, often missing in others despite its definite importance in establishing the phylogeny considering its intermediate position between *P. p. irlandei*, *Alvinella* and other species of *Paralvinella*. We therefore retain T6, as shown in Figure 5, as the preferred phylogeny for the Alvinellidae, well supported by molecular data and initially hypothesized by Desbruyères and Laubier based on morphological characters.

However, comparing with other topologies, notably T7 and T9, does not allow us to exclude with certainty the possibility of a paraphily of *Paralvinella*. Topologies T7, T8 and T9, in which *P. p. irlandei* is the first species to diverge within the Alvinellidae, would all agree with several phylogenies proposed on the basis of COI sequences and transcriptomic sequences as well as some morphological characters (Vrijenhoek, 2013; Stiller et al., 2020; Jollivet and Hourdez, 2021). In our analysis, the T9 topology is well supported by the coalescence of gene trees (using a methodology similar to the one employed by Stiller et al. (Stiller et al., 2020)), as is the T7 topology to a lesser extent. In the present study, we have taken particular care to ensure that each orthologous gene cluster used has a sequence for

P. p. irlandei. However, the sequencing quality is often lower for this species and the sequences are much more fragmented and divergent than for the other species. To better resolve this ancestral node, it would be necessary to improve the sequence data for this species and to get additional information from its sibling species *P. p. pandorae* from Juan de Fuca. Finally, it is worth noting that T9 clusters the two *Alvinella* species with *P. unidentata*. This topology thus raises the question of the monophily of *Alvinella* in the traditional view that separates this genus from all other species of Alvinellidae. This grouping indeed fits better with the gills' morphology: the filaments of *P. unidentata* have the particularity to be lamellar, which is unique compared to other *Paralvinella* species, and similar to the two *Alvinella* species. However, the overall structure of the gill, with a tip devoid of filaments and funnel-like inserted secondary filaments, brings the species closer to other *Paralvinella* and especially to *P. p. irlandei* (Desbruyères and Laubier, 1993). In this topology, the lack of filaments at the tip of the gills and the cylindrical shape of the secondary filaments would be symplesiomorphic characters, while the acquisition of the lamellar shape of the gills would be a synapomorphic character shared by the *Alvinella*+*P. unidentata* group, which contradicts the original hypothesis by Desbruyères and Laubier (Desbruyères and Laubier, 1993). The flat or funnel-like organization of the gills would result either from an evolutionary convergence between *Alvinella* and *Paralvinella* (with the exception of *P. p. irlandei* and *P. unidentata*), or from a convergence towards the funnel-like form by *P. unidentata* and *P. p. irlandei*. Considering T6 as the true topology of the species, the similarity between the gills of *Alvinella* and *P. unidentata* could also be attributed to introgressive gene flow between the initial radiating lineages. The study of the gills in Alvinellidae is a good example of the important morphological variations that can be observed within the Terebelliformia at the level of palps, setae and gills, as well as the shape and position of the uncini (Day, 1964; Stiller et al., 2020), which make this character challenging to establish their relative importance in relation to each other from a phylogenetic perspective.

The last interesting topology obtained from the molecular data corresponds to T1, which represents an intermediate position between T6 and T9. In this case, we find the *Alvinella*+*P. unidentata* group of T9, but also the monophily of the other *Paralvinella* species with *P. p. irlandei* as in T6. However, we note that by limiting the possible topologies for the gene trees to T1, T6 and T9, the genes for which T1 is preferred are in their great majority more in agreement with T9 as second choice than with T6. For these genes, the clustering of *P. unidentata* with *Alvinella* outweighs the monophily of *Paralvinella* (although T6 remains the topology preferred by the majority of the genes).

The monophily of *Paralvinella*, according to the T6 topology, seems to be the preferred solution. However, many genes are inconsistent with this view and tend to bring *P. unidentata* and *Alvinella* species closer together. This grouping thus raises the question of the definition of the genus *Alvinella*, which clusters together two particularly closely-related species (divergence of 0.83% according to the JTT+Γ model on amino acid sequences) and for which the synapomorphies between *A. pompejana* and *A. caudata* are not informative in the resolution of the family root. Investigating the relation between *P. unidentata* and *P. p. irlandei*, and therefore improved sequencing of the species *P. p. irlandei*, is essential. Nevertheless, this radiation is undoubtedly a difficult case to resolve where multiple short branches ($< 1\%$) are buried deep in the tree topology. The radiation leading to the separation of the *Alvinella*, *P. unidentata*, *P. p. irlandei* and other *Paralvinella* lineages is likely a fast event, probably comprised between 1 and 10 million years. If genealogical species (group of organisms which are more closely related to each other than to any organism outside the group) are considered to be monophyletic at 50% of the loci, genealogical species status is reached after 4-7 N generations after isolation of the initial populations, but complete lineage sorting still requires 9 to 12 N generations (Hudson and Coyne, 2002). In previous studies, Chevaldonné and Jollivet estimated a census population size of about 45,000 individuals for *A. pompejana* per chimney, while Jang et al. estimated the ancestral population for *A. pompejana* of about 1.2 million

individuals at the scale of the East Pacific Rise about 4.2 million years ago (Chevaldonné and Jollivet, 1993; Jang et al., 2016). Considering effective population size between 500,000 and 1 million for the population of the first Alvinellidae ancestor, the status of genealogical species can only be reached after 5 to 12 million generations, which is longer than the time span of the radiation assuming annual generations. This estimate is in good agreement with the observation that for at least *P. unidentata*, *P. hessleri* and *P. fijiensis*, present-day populations that are separated for 3 to 4 million years still display very short coalescent times (see supplementary data Fig. 2 and 3). Furthermore, this is consistent with the large uncertainty in the topology at the root of the family depending on the genes studied, where no topology is drastically discredited by another (see supplementary data Fig. 4).

Consequently, in the case of the Alvinellidae, in addition to the quality of the sequencing for *P. p. irlandei*, the rapid radiation between the first ancestral lineages raises questions about the relevance of a binary tree of species at the root of the family. However, the divergence between *P. unidentata* and *P. p. irlandei* (topology T6 which supports the *Paralvinella* monophily) or between *P. unidentata* and *Alvinella* (T9 and T1, which does not resolves the question of the *Paralvinella* monophily), was probably concomitant to the first speciation event, probably less than 10 Myr after the radiation of the family. The relationship between *P. unidentata* and *P. p. irlandei* is therefore central to the resolution of the phylogeny.

Division of the genus Paralvinella

The genus *Paralvinella* has been sub-divided by Desbruyères and Laubier into three subgenera on the basis of morphological characters (Desbruyères and Laubier, 1993):

- *P. Paralvinella*, which groups the species *P. palmiformis*, *P. grasslei*, *P. fijiensis* and *P. sulfincola*;
- *P. Miralvinella*, which groups together the species *P. hessleri*, *P. dela* and *P. bactericola*;

- *P. Nautalvinella*, which includes the species *P. unidentata*, *P. p. irlandei* and *P. p. pandorae*.

Han et al. also proposed that the species *P. mira* should be related to the subgenus *Miralvinella*, and more particularly to *P. hessleri*, considering the shape of the gills and of the buccal apparatus despite the different implantation of the tentacles (Han et al., 2021).

Several criteria are used to differentiate these subgenera, notably the presence of specialized sexual tentacles (used in the male-female pairing during reproduction) on the mouth apparatus of males in *P. Paralvinella* (three-lobed) and *P. Miralvinella* (pointed and coiled) which are absent in *P. Nautalvinella*, the shape of the gills (filaments on the opposite sides in *P. Paralvinella* and *P. Miralvinella*, and comb-like in *P. Nautalvinella*), as well as the presence of digitiform notopodial lobes on the anterior part of *P. Paralvinella* and *P. Miralvinella* which are absent in *P. Nautalvinella*. Other morphological criteria are less clear in differentiating these subgenera, such as the number of segments of the animals (55 to 180 setigers in *P. Paralvinella versus* 50 to 75 in *P. Miralvinella* and *P. Nautalvinella*), or the position of the first uncini (setigers 12 to 26 in *P. Paralvinella*, 15 to 26 in *P. Miralvinella* and 5 to 32 in *P. Nautalvinella*) (Desbruyères and Laubier, 1993; Jollivet and Hourdez, 2021; Han et al., 2021).

Considering the T6 topology, the definition of the subgenus *Nautalvinella* appears robust even if the three species of this group, namely *P. unidentata*, *P. p. irlandei* and *P. p. pandorae* are highly divergent. The acquisition of the comb shape of the gills, the loss of the pair of gills on the buccal apparatus of males and the loss of the notopodial lobes are three relevant synapomorphies that have occurred once in the *Nautalvinella* lineage. If, on the other hand, *P. unidentata* is considered to be more closely related to the two *Alvinella* species (topologies 1 and 9) with leaf-shaped branchial filaments, then the subgenus *Nautalvinella* contains only the species *P. p. irlandei* and *P. p. pandorae*. *Alvinella* species have notopodial lobes, a pair of large tentacles on the mouth apparatus and the gill filaments are inserted on opposite sides. Assuming topologies 1, 7 and 9, these

three traits must have been gained at least twice in the branches leading to *P. p. irlandei* and *P. unidentata* (or two losses in the branches leading to *Alvinella* and all other *Paralvinella*). The other most parsimonious topologies regarding the evolution of these criteria are T3, T5 and T13, but none of these topologies are well supported by the molecular dataset (Table 1). It is also noteworthy that an analysis of 278 loci (Jollivet and Hourdez, 2021) had concluded to a topology where *P. p. irlandei* +*P. p. pandorae* and *P. unidentata* were basal to the genus *Paralvinella*, although they were not sister species. This topology, close to T4 in our study, is only well supported in the analysis combining coalescent-based approach with constrained gene trees (Fig. 4).

The other two subgenera *Miralvinella* and *Paralvinella* differ mainly in the shape of the buccal apparatus, which has a pair of three-lobed sexual tentacles in the subgenus *Paralvinella* and a pair of pointed tentacles in *Miralvinella*. The phylogeny of these species is well resolved with maximum bootstrap confidence regardless of the used phylogenetic method. The species *P. mira*, which has the same pointed tentacle shape (Han et al., 2021), effectively clusters with *P. hessleri*, which is the only *Miralvinella* species included in this phylogeny. According to Han et al., *P. mira* is morphologically closer to *P. hessleri* than to *P. dela* or *P. bactericola* (which are two sister species respectively inhabiting the Juan de Fuca Ridge and the EPR) (Desbruyères and Laubier, 1993; Jollivet and Hourdez, 2021). This also makes sense considering the geographical distribution of these species (*P. hessleri* occupying the Mariana and Manus basins while *P. mira* is found in the Indian Ocean). Molecular data are not available for the rarer species *P. dela* and *P. bactericola*, but on the basis of morphological characters specific to *Miralvinella*, a monophyletic origin of these species with a topology ((*P. dela*,*P. bactericola*),(*P. hessleri*,*P. mira*)) can be envisioned with confidence. The subgenus *Paralvinella*, in the proposed phylogenies, is however no longer monophyletic since the *Miralvinella* species are included in this group. The three-lobed tentacles of the mouth apparatus would then represent an symplesiomorphy of the genus *Paralvinella*, derived later as pointed tentacles in

miralvinella.

In conclusion, the initial subdivision into subgenera proposed by Desbruyères and Laubier within the genus *Paralvinella* still holds for the subgenus *Miralvinella*, but not for the subgenus *Paralvinella* which is polyphyletic with two groups of species: one living in colder conditions than the other. Depending on the chosen topology (T6, T7 or T9), the subgenus *Nautalvinella* may or may not be considered valid, but it is noteworthy that this initial proposal, established from a detailed morphological description of these species, is one of the preferred solutions on an otherwise hard-to-establish molecular phylogeny. Indeed, T6 is the only topology including both the monophyly of the genus *Paralvinella* and the *Nautalvinella* lineage. The ongoing description of new geographic species from this latter subgenus will probably help to solve this specific clade.

Spreading of the family in and outside of the Pacific Ocean

Regardless which of the major topologies is chosen, most of the basal species in the family (*A. pompejana*, *A. caudata*, *P. p. irlandei*, *P. grasslei* and *P. palmiformis*) are endemic to the East Pacific Rise (EPR) or the Juan de Fuca Ridge (JdF) (Desbruyères and Laubier, 1986). This distribution suggests an origin of the family in the Eastern Pacific Ocean, followed by the colonization of the Western Pacific. This is consistent with the distribution of the sister family Ampharetidae, for which species diversity was the greatest in the eastern Pacific as well, which may suggest a common origin for both families in the now-extinct Pacific-Farallon ridge. This ridge has been active since the Early Jurassic (200 Ma) and has been proposed to have played a central role in the initial dispersal of the Pacific hydrothermal fauna (Mammerickx et al., 1980; Bachraty et al., 2009).

Depending on the model used for the molecular dating, radiation of the Alvinellidae family occurred between 65 (CIR model, 95% confidence interval 55-78 Myr) and 100 million years ago (uncorrelated gamma model, 49-193 Myr). Worm imprintings on fossil tubes dating from 95 million years ago in the Bayda massive sulfide deposit of the Samail

ophiolite, Oman, exhibit a structure very similar to those of species such as *A. pompejana* or *Riftia pachyptila* (concentric structure, interlamination of pyrite with void probably originally filled with protein assemblies) (Haymon et al., 1984; Gaill and Hunt, 1986). These fossils suggest a hydrothermal sea-floor similar to the modern sources of the EPR and JdF, suggesting that Alvinellid worms may have had a long speciation history dating back to the Cretaceous (Haymon et al., 1984; Vrijenhoek, 2013), consistent with both of these dates. Using 278 orthologous genes or the *Cox1* gene, Jollivet and Hourdez proposed two dates for the radiation of the Alvinellidae, one at 200 Ma and the other at 70 Ma (Jollivet and Hourdez, 2021). The first dating clearly appears too old but was associated with only one calibration date. Indeed, molecular phylogeny analyses show that most of the invertebrate taxa underwent a radiation in the Cenozoic (about 66 Ma), and the fossil record exhibits important transitions of the chemosynthetic faunas during the late Mesozoic (145-66 Ma). These transitions are best explained by the recolonization of the deep environment following two mass extinctions caused by major anoxic/dysoxic events. The first during Cenomanian/Turonian stage (94 Ma, for which the benthic foraminiferous record gives no real evidence of a mass extinction), and the second during the Paleocene/Eocene stage associated with deep-water warming (57 Ma, for which 50% of the benthic foraminifera became extinct) (Vrijenhoek, 2013). These events argue more in favor of a radiation dating around 60 Myr, as obtained by the CIR model. This is in good agreement with the dating of Jollivet and Hourdez established from *Cox1* with two calibration dates, although the phylogeny of the Alvinellidae used in this case is closer to T7/T9, and the time of radiation between *P. p. irlandei*, *Alvinella* and the group of other *Paralvinella* extends over 20 Myr, which is a much longer time than in our results (5 Myr). Finally, the CIR model proposes a colonization of the Western Pacific (divergence between *P. grasslei* +*P. palmiformis* and the other *Paralvinella* species) at 50 Myr (95% confidence interval: 44-59), which is in good agreement with the activity of a mid-ocean ridge North of New Guinean 55 Ma, as well as the fossilization of the Kula Ridge 43 Ma. These ridges

have both been proposed to have allowed the spread of species (and more specifically those which were more likely to disperse near the seafloor) between the western and eastern Pacific hydrothermal basins, which are now more isolated from each other (Hessler and Lonsdale, 1991; Bachraty et al., 2009). The dates proposed by the CIR model are therefore the most consistent with the geological events that are thought to have played a major role in structuring the Pacific hydrothermal basins and thus the vent communities.

In this hypothesis, the divergence between the Alvinellidae and Ampharetidae occurs at about 100 Ma, at a time when high spreading rates and off-ridge volcanism formed sub-aerial volcanic environments, which are now submerged in the Pacific (Röhl and Ogg, 1996). These emerged masses may have served as refuges for vent faunas during episodes of deep-sea anoxia from the Aptian to the Cenomanian (125 to 94 Ma) (Jacobs and Lindberg, 1998). In contrast to the Ampharetidae, for which colonization of the deep hydrothermal environment seem to have happened several times independently, all the Alvinellidae are endemic to the hydrothermal environment and it is possible that the ancestor of the family was already endemic to Pacific vents. Fontanillas et al. proposed that the ancestor of the family was adapted to warmer environments (Fontanillas et al., 2017), which is in good agreement with the higher temperatures of 10 to 15 °C of the Pacific deep waters during the late Mesozoic between 100 and 60 Myr, with a peak reached between 55 and 60 Myr (Savin, 1977; Vrijenhoek, 2013). This warmer temperature and the sink of massive surface productions caused a period of anoxia and a mass extinction of the deep environment, allowing recolonization by new taxa. Subsequently, the family may have spread to the Western Pacific by the active fault system between 55 and 43 Myr that connected the two Pacific hydrothermal basins, now mostly connected by the Antarctic Ridge (Hessler and Lonsdale, 1991; Bachraty et al., 2009; Moalic et al., 2012). Interestingly, the colonization of the Pacific Ocean must have taken place in several successive stages, given the alternance in the phylogeny between species associated with the eastern Pacific (*A. pompejana*, *A. caudata*, *P. p. irlandei*, *P. grasslei*, *P. palmiformis*,

P. sulfincola) or with the western Pacific (*P. hessleri*, *P.* sp. nov., *P. fijiensis*, *P. unidentata*) (Desbruyères and Laubier, 1993). The divergence times between these different groups are all dated within a period between 41 and 44 Myr (95% confidence interval: 35-55), which is notably compatible with the activity of the Kula Ridge 43 Ma (Hessler and Lonsdale, 1991). The diversification of species in the Manus and Lau basins (*P. fijiensis* and *P.* sp. nov.) is dated between 19 and 28 Myr, which precedes the separation of these basins about 10 Ma (Boulart et al., 2022).

The recent description of *P. mira* in the Indian Ocean, at Wocan and Daxi vents, shows that the Alvinellidae have spread outside the Pacific Ocean. The divergence between *P. mira* and *P. hessleri* is dated at 31 Myr (24-38). The colonization of the Indian Ocean was therefore from a Western Pacific population, at a time when the Caroline Basin floor was formed (34-29 Ma) and the Banda Arc was active (Parker and Gealey, 1985). The connection between the western Pacific and the Indian Ocean through Northern Australia was closed between 60 and 40 Ma, with the advance of the Philipines and the opening of the China Sea (Parker and Gealey, 1985; Moalic et al., 2012). Although the complex tectonics of this region make it difficult to reconstruct geological ridge connectivity between basins, it is more likely that the colonization of the Indian Ocean passed through the Antarctic Ridge, whose spreading started 50 Ma and connects the Indian Ocean with the Pacific Ocean (Parker and Gealey, 1985). This ridge is indeed proposed to connect the main hydrothermal basins, from the EPR to the Atlantic Ocean through the Pacific Ocean (Moalic et al., 2012). Although Alvinellid worms are rare in the Indian Ocean, an undescribed species of *Paralvinella* was found in the Solitaire field on the Central Indian Ridge (Nakamura et al., 2012). The sampling of this species may be of particular interest to investigate the possibility that the CIR acted as a stepping stone to the Wocan and Daxi sites in the dispersal of Alvinellidae. At present, no Alvinellidae have been observed in the Atlantic Ocean, as potential alvinnelid-related sequences sampled at the mid-Atlantic ridge (López-García et al., 2003) are found after verification to belong to the

ampharetid genus *Amphisamytha*, which is known to thrive at this location (Eilertsen et al., 2017). Nevertheless, the presence of Alvinellidae on the CIR makes this possibility credible, as this ridge acts as a bridge between the Pacific and Atlantic Oceans in the dispersal of hydrothermal species (Tunnicliffe, 1988; Moalic et al., 2012; Watanabe and Beedessee, 2015).

Molecular dating would also gain in accuracy through a more comprehensive sampling of the different species of the family. Indeed, only two calibration points have been used here. The first point corresponds to the opening of the Manus (3-4 Ma) and Lau (1-2 Ma) basins, which are assumed to form two disjunct ridge systems in the New Guinea arc populations. The second refers to the subduction of the Farallon Plate beneath the North American Plate, which caused a vicariance event between hydrothermal faunal populations distributed between the modern JdF and the EPR between 34 and 23 Myr (Tunnicliffe, 1988). This vicariance event has already been used to date other phylogenies of vent species distributed between the EPR and JdF, if we consider that these species spread mostly along the axial valley of oceanic ridges (Stiller et al., 2013; Jollivet and Hourdez, 2021). This calibration is particularly interesting in the case of the Alvinellidae, since it separates several pairs of sibling species distributed throughout the phylogeny: *P. palmiformis* and *P. grasslei*, *P. p. irlandei* and *P. p. pandorae*, as well as *P. dela* and *P. bactericola* (Tunnicliffe, 1988; Desbruyères and Laubier, 1993; Jollivet and Hourdez, 2021). However, there are no molecular data for the rarer species *P. dela* and *P. bactericola*. Similarly, only the *Cox1* gene of *P. p. pandorae* is available, which allowed Jollivet and Hourdez to estimate a 70 Myr old radiation of the Alvinellidae with two calibration points, using the species pairs *P. palmiformis* with *P. grasslei* and *P. p. irlandei* with *P. p. pandorae* (Jollivet and Hourdez, 2021). In addition, using the species *P. palmiformis* and *P. grasslei*, Chevaldonné et al. established a molecular clock of *CO1* in the Alvinellidae corresponding to a divergence of 0.13%/Myr (Chevaldonné et al., 2002). Better sampling of the rare species of the family, particularly the *P. Miralvinella*

species and *P. p. pandorae*, would therefore help to improve the estimate of the species' dispersion date over both the Pacific and the Indian oceans.

Finally, if the ecological diversification and geographic spread of Alvinellidae are undeniable over geological time scales, the life cycle of Alvinellidae and the dispersal ability of species are still poorly documented. Our implicit hypothesis is that the dispersal of Alvinellidae is primarily driven by the tectonic dynamics of hydrothermal sites, and that it therefore occurs near the seafloor along the axial graben of oceanic ridges (Chevaldonné et al., 1997). In *A. pompejana*, modelling of the birth and death of hydrothermal vent sites hints that the displacementof venting regions (vent fields) allows for the homogenization of populations at the ridge scale, without the need of a great dispersal capacity provided that the ridge is not physically offset by geological discontinuities such as transform faults (Jollivet et al., 2022). Moreover, for *P. grasslei*, it has been suggested that short-term and oriented dispersal may occur by zoochory within a given active field, with larvae observed being carried by hydrothermal bythograeid crabs (Zal et al., 1995). The possibility of long-term dispersal of the larvae by marine currents was also proposed because larvae are lecithotrophic and can remain for a long period of time in the water column (McHugh, 1989). In *A. pompejana*, larvae are indeed able to stop their development when temperatures drop but resume development at 10°C, and are unable to develop at temperatures above 20°C, showing that the larvae develop outside the adult colonies but should meet an active vent site to develop again and settle (Pradillon et al., 2005). Furthermore, little is known about the early stages of colonization of newly formed hydrothermal vents, whose permeable and unstable matrix is not immediately colonized by invertebrates (Eilertsen et al., 2017). In *A. pompejana*, the study of the phosphoglucomutase (PGM1) gene shows that the 90 allele is more thermostable than the 78 and 100 alleles, and more frequent in newly-formed colonies (Piccino et al., 2004; Bioy et al., 2022). Such allelic specialization for warm habitats could be a clue for the colonization of new territories by alvinellid worms and may explain some of the geographic

discrepancies found among some species. Thus, in addition to Pacific tectonic events that promote reductions and expansions of hydrothermal populations along mid-ocean ridges and back-arc basins, the study of life-history traits of the Alvinellidae, in particular the dispersal and settlement patterns of larval and post-larval stages, is essential to understand the expansion capabilities of this family.

Methodological control of the proposed phylogenies

Constrained topologies To identify the most likely tree topology for the Alvinellidae, we chose to exhaustively evaluate the most relevant topologies via different phylogenetic inference methods. These analyses provided robust inferences for some specific groupings. Among them, the *Paralvinella* species (without both *P. unidentata* and *P. p. irlandei*) were thus constrained, considering that this group was correctly resolved with reliable and consistent species positioning for all methods (concatenation, coalescent-based approach, on coding sequences or translated proteins). The outgroup species (Ampharetidae, *M. palmata*, *N. edwardsi*, Terebellidae gen. sp. and *P. gouldii*) were also constrained as an outgroup subtree. The grouping of *M. palmata* with Terebellidae gen. sp. and *N. edwardsi* was robust with a high bootstrap value on the concatenated sequences (>80) and maximal score with the coalescent-based approach, which is consistent with the phylogeny proposed by Stiller et al. that clusters the Mellinidae and the Terebellidae together with strong support (Stiller et al., 2020). The species positioning in the Ampharetidae is more questionable. The branch leading to *Anobothrus* sp. is indeed basal to the family Ampharetidae from analyses using amino-acid encoded sequences (bootstrap>80), but the nucleotide data suggest an older origin of the branch, whose separation predates the divergence between the Ampharetidae and Alvinellidae families. Here we chose to consider only the topology in which Ampharetidae species form a monophyletic group (Eilertsen et al., 2017; Stiller et al., 2020).

The question whether the Alvinellidae form a monophyletic sister group to the

Ampharetidae or whether the family should be included within the Ampharetidae is indeed not fully resolved. Both families have retractable bucal filaments attached to the dorsal curtain, which is an important trait in establishing the monophyly of Ampharetidae. While several morphological and molecular studies have concluded about the clear separation of the two families (Rousset et al., 2003; Stiller et al., 2020), Eilertsen et al. suggested that a broader phylogeny of the Ampharetidae based on molecular data (mtCOI, mitochondrial 16S rDNA, nuclear 18S and 28S rDNA) include Alvinellidae (Eilertsen et al., 2017), with a trifurcation between the Alvinellidae, *Amphysamytha* and *Anobothrus*+*Amphicteis*lineages. In contrast to this last result, our data clearly brought the *Amphisamytha* species (*A. carldarei*) closer to the *Amphicteis* species (*A. gunneri*). Moreover, considering that the Mellininae are no longer considered a subfamily of Ampharetidae but a true family (Mellinidae), the trifurcation proposed by Eilertsen et al. is solved in our analysis because of the divergence between the Ampharetidae (excluding the Mellinidae) and the Alvinellidae. It is this topology that we have used for further analysis, in accordance with phylogenies including more Ampharetidae species (Stiller et al., 2020).

Data quality Phylogenetic reconstruction is sensitive to differences in species sequence composition. Indeed, similar compositional biases can lead to evolutionary convergence events that can be mistaken for synapomorphy. Conversely, the absence of compositional bias in some species compared to others can have the same effect (the symplesiomorphy trap), which is a special case of long branch attraction. This phenomenon has been observed in mitochondrial data of Ampharetidae and Alvinellidae, which exhibit strong amino-acid and base composition biases (Zhong et al., 2011). In addition, Alvinellidae colonize environments with a wide temperature range, which is known to have an effect on the amino-acid composition of biomolecules between mesophilic and thermophilic eukaryotes (Wang and Lercher, 2010; van Noort et al., 2013; Bock et al., 2014).

Evaluation of composition biases in the orthologous gene clusters shows strong

disparities between species (see supplementary data Fig. 5), in particular for *P. p. irlandei* which shows very different indices from other species for CvP bias, PAYLE *vs.* DGMS and purine load, criteria suggested to discriminate between warm and cold species (Hickey and Singer, 2004; Fontanillas et al., 2017). Therefore, we carefully filtered out genes that exhibited too much compositional heterogeneity according to Phylobayes for further analysis. In addition, we tried to reconstruct the phylogeny according to the Phylobayes+ASTRAL method by first recoding the amino acids according to their physicochemical properties (Dayhoff 6 coding). The recoding of the sequences does limit the heterogeneity of composition, but leads to a result identical to the phylogeny without recoding. We have therefore preferred to limit ourselves to the filtering of sequences encoded in amino acids.

The composition of *P. p. irlandei* sequences is particularly different from the average composition of all other species, especially for serine contents, CvP bias, and purine load. The sequences of *P. p. irlandei* are more fragmented than for the other species, which could have an effect on the average composition. One might expect that the strong compositional heterogeneity of *P. p. irlandei* would tend to make this species diverge more rapidly among the Alvinellidae in the phylogenetic reconstruction, which surely contributes to the good scores of the T7, T8 and T9 topologies. However, as the average composition of *P. p. irlandei* diverges not only from the other species of Alvinellidae, but also from the species chosen as outgroup (Ampharetidae + Terebellidae + *P. gouldii*), one must be cautious with these results which suggest that the sequences of *P. p. irlandei* may have evolved very rapidly, and that the placement of the basal species to the Alvinellidae may be a typical case of long-branch attraction.

A final difficulty with this phylogeny regards the rapid radiation of the family that probably occurred at the end of the Cretaceous in less than 10 million years, suggested by the short branches at the root of the family compared to the leaves that accumulated many more mutations. Early Cenozoic radiation is suggested for most dominant invertebrate

taxa (Vrijenhoek, 2013), and this rapid divergence can lead to a high rate of incomplete lineage sorting (ILS) and results in larger gene tree estimation errors (GTEE) (Molloy and Warnow, 2018). Under these conditions, the ILS can lead to a case where the most frequent gene trees differ from the actual species tree. We must therefore be especially vigilant about the main biases that can lead to a difference between the gene trees and the species tree, namely ILS, inter-species gene flow, gene duplications and GTEE (Cai et al., 2021).

To limit the effects of gene duplications, we focused on orthologous gene clusters considered single-copy in lophotrochozoans in ODB9 (see Material & Methods), and this effect should therefore not be a concern in our analysis. It is suggested that in the case of very high ILS, coalescent-based methods give superior results to concatenation methods for estimating the true species tree. Indeed, concatenation methods impose the same evolutionary history on all genes, and do not take into account the possibility of gene flow or ILS. However, the superiority of coalescent methods only appears in cases of very high level of ILS. For low or medium ILS, both methods produce equivalent results, with an increased reliability for the concatenation method which benefits from longer alignments that allow the reduction of GTEE (Molloy and Warnow, 2018). Generally speaking, in the case of ILS or large amounts of inter-species gene flows, the coalescent method should be preferable provided that GTEE is low with, in particular, sufficiently long gene alignments and a large enough number of genes (several hundred) (Nute et al., 2018). In the case where the ILS or gene flow are sufficiently low, or the GTEE is high, concatenation methods give more reliable results (Roch and Warnow, 2015). A simulation by Molloy and Warnow on 26 species and 1000 genes, conditions relatively similar to our study, further shows that filtering genes on the basis of GTEE is only beneficial if the ILS is low. Otherwise, increasing the number of genes is preferable, but the large GTEE severely limits the ability of both methods to find true species trees (Molloy and Warnow, 2018).

In our study, GTEE seems to be the first concern. Indeed, many gene topologies exhibit aberrant results, with notably the species *P. p. irlandei* which position is very

fluctuating and can cluster with outgroup species, probably due to its heterogeneity in sequence composition and long branch attraction. We therefore considered in a second step that ILS and gene flow were negligible at all nodes of the tree, which are globally very well supported by all methods, except for the basal nodes of the Alvinellidae lineage. We recomputed gene trees in this reduced topology space corresponding to trees T1 to T15, with the objective of limiting the GTEE on these nodes by imposing a strong *prior* on the possible topologies. In this framework, most of the genes agreed with the T6 topology (52.9% of the genes for nucleotide-encoded sequences compared to T9; 54.7% for protein sequences against T9). Surprisingly, we find that all 15 tested topologies are supported by a significant number of genes, regardless of the number of informative sites at the gene level (see supplementary data Fig. 6). If the preference between gene topologies was only due to the GTEE, we would expect to see some topologies prevailing over others as the phylogenetic signal becomes stronger. This result suggests that other processes such as ILS and gene flow must be involved in the discordance between gene trees. Furthermore, unlike ILS which randomly segregates alleles between lineages at a node, gene flow occurs specifically between two lineages. Therefore, in the case of high gene flow from one lineage to another, one would expect to see some secondary topologies globally more represented than others, reflecting gene flow between the clustered branches. In our case, topologies T6, T9, and perhaps T5, T7 and T13 are the most represented at the level of the gene trees, which could suggest gene flow between the oldest ancestors of the family (considering that the topologies that group *P. p. irlandei* with the *Alvinella*, such as T2, T12 and T15, are less supported). However, as all topologies are fairly well represented at the level of the gene trees, it is very probable that high ILS occurs at the root of Alvinellidae.

We therefore evaluated each topology with ASTRAL from these constrained gene trees, assuming that ILS and gene flow were negligible in the rest of the tree. This approach aims to reduce the GTEE, which is an essential assumption of the coalescent-based methods. In this case, topologies T7, T4, T6 and T5 are the four preferred solutions. Here

again, the grouping of *P. p. irlandei* with the other *Paralvinella* is debatable (T7 and T4, which have very similar results). T7 had also obtained a good evaluation in the concatenation plus the maximum likelihood method for the protein sequences. In general, the comparison of T6 with all other relevant topologies (T1, T4, T5, T7, T9) shows that most of the genes are in agreement with the T6 topology (55.1% of the genes for nucleotide-encoded sequences compared to T7; 57.4% for protein sequences against T7). We therefore conclude that at this stage, T6 remains the most likely topology. However, ILS which must have been important at the time of the radiation of the Alvinellidae, as well as gene transfers that probably took place notably between the *Alvinella* species and *P. unidentata*, make the resolution of this phylogeny very delicate. The sequencing effort, which includes several hundred genes, is already very substantial and it is unlikely that increasing the number of sequences will improve this result significantly. On the other hand, the sequences for *P. p. irlandei* are much more fragmented than for other species, which may significantly increase the GTEE regarding the placement of this species.

indel inferrence The indel likelihood model used for species-tree evaluation aims to use an alternative type of mutational information that is not sensitive to the base/codon composition biases of transcripts or translated coding sequences,namely insertions or deletions of sequences. We also estimated intantaneous indel parameters r, l_1 and l_2 that characterize the ratio between the instantaneous probabilities of occurrence of insertions against deletions, as well as the size distributions of insertions and deletions.

For insertions, l_1 ranged between 2 and 2.36, and for deletions, l_2 ranged between 2.39 and 2.66. This is rather in agreement with the values obtained empirically by Gu and Li and Cartwright on human and rodent sequences, which estimate l_1 between 1.68 and 1.96 and l_2 between 1.77 and 2.32, longer in non-coding regions (1.6-1.7) than in pseudo genes where l is greater (1.7-2.3) (Gu and Li, 1995; Cartwright, 2009). In our study, indel appear to be quite short, which is expected as we investigate the evolution of coding sequences and consider the instantaneous distribution lengths, while these studies observe

the outcome of several successive indels events that can accumulate into longer indels.

Indeed, Gu and Li also estimated r to be between 0.29 and 0.52, which is more in line with the parameters we estimated when considering a soft filter on indels in our alignment ($r = 0.35$). In our data, however, r is probably underestimated by alignment errors that will create false homologies between sequences around an indel site. This homology imposes on the model that the sequence was present in the ancestral state, with the consequence that the optimization must compensate by increasing the rate of independent deletions in the daughter branches. The "hard filter", where only short gaps with a single insertion or deletion event are considered over the entire phylogeny, should give a better estimate of the parameter as the alignment quality for theses gaps is much higher. In this case, the probability of an insertion is higher than for a deletion ($r = 1.44$).

This newly-developed indel-based model seems to be quite promising, and especially provided results that correlate well with the coalescent-based methods (best topologies in the "hard filter" condition are T1, T9, T13, T14, T15 as for Phylobayes+ASTRAL). However, the size of the dataset (2050 indels) and the quality of the alignment are two important limitations of the method at the moment. Indeed, we obtained substantial differences depending on the filter applied to the indels. Some indels have much higher evolutionary rates than others, which is more likely to be related to misalignments than to a biological reality. Ideally, the two steps of alignment and phylogenetic likelihood calculation should be conducted in parallel according to the same evolutionary model. Nonetheless, this seems to be an interesting model to explore in future works, as the evaluation of phylogenies from well-defined gaps gave fairly consistent results despite this two constraints.

Supplementary Material

Data available from the Dryad Digital Repository: `http://dx.doi.org/XXXX.[NNNN]`.

Conflict of interest

The authors declare no conflict of interest.

References

Bachraty, C., Legendre, P., and Desbruyères, D. (2009). Biogeographic Relationships Among Deep-Sea Hydrothermal Vent Faunas at Global Scale. *Deep Sea Research Part I: Oceanographic Research Papers*, 56(8):1371–1378.

Bioy, A., Le Port, A.-S., Sabourin, E., Verheye, M., Piccino, P., Faure, B., Hourdez, S., Mary, J., and Jollivet, D. (2022). Balanced Polymorphism at the Pgm-1 Locus of the Pompeii Worm Alvinella pompejana and Its Variant Adaptability Is Only Governed by Two QE Mutations at Linked Sites. *Genes*, 13(2):206.

Bock, T., Chen, W.-H., Ori, A., Malik, N., Silva-Martin, N., Huerta-Cepas, J., Powell, S. T., Kastritis, P. L., Smyshlyaev, G., Vonkova, I., Kirkpatrick, J., Doerks, T., Nesme, L., Baßler, J., Kos, M., Hurt, E., Carlomagno, T., Gavin, A.-C., Barabas, O., Müller, C. W., van Noort, V., Beck, M., and Bork, P. (2014). An Integrated Approach for Genome Annotation of the Eukaryotic Thermophile *Chaetomium thermophilum*. *Nucleic Acids Research*, 42(22):13525–13533.

Boulart, C., Rouxel, O., Scalabrin, C., Le Meur, P., Pelleter, E., Poitrimol, C., Thiébaut, E., Matabos, M., Castel, J., Tran Lu Y, A., Michel, L. N., Cathalot, C., Chéron, S., Boissier, A., Germain, Y., Guyader, V., Arnaud-Haond, S., Bonhomme, F., Broquet, T., Cueff-Gauchard, V., Le Layec, V., L'Haridon, S., Mary, J., Le Port, A.-S., Tasiemski, A., Kuama, D. C., Hourdez, S., and Jollivet, D. (2022). Active Hydrothermal Vents in the Woodlark Basin May Act as Dispersing Centres for Hydrothermal Fauna. *Communications Earth & Environment*, 3(1):64.

Cai, L., Xi, Z., Lemmon, E. M., Lemmon, A. R., Mast, A., Buddenhagen, C. E., Liu, L.,

and Davis, C. C. (2021). The Perfect Storm: Gene Tree Estimation Error, Incomplete Lineage Sorting, and Ancient Gene Flow Explain the Most Recalcitrant Ancient Angiosperm Clade, Malpighiales. *Systematic Biology*, 70(3):491–507.

Cartwright, R. A. (2009). Problems and Solutions for Estimating Indel Rates and Length Distributions. *Molecular Biology and Evolution*, 26(2):473–480.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics*, 34(17):i884–i890.

Chevaldonné, P., Fisher, C. R., Childress, J. J., Desbruyères, D., Jollivet, D., Zal, F., and Toulmond, A. (2000). Thermotolerance and the 'Pompeii worms'. *Marine Ecology Progress Series*, 208:293–295.

Chevaldonné, P. and Jollivet, D. (1993). Videoscopic Study of Deep-Sea Hydrothermal Vent Alvinellid Polychaete Populations: Biomass Estimation and Behaviour. *Marine Ecology Progress Series*, 95:251–262.

Chevaldonné, P., Jollivet, D., Desbruyères, D., Lutz, R. A., and Vrijenhoek, R. C. (2002). Sister-Species of Eastern Pacific Hydrothermal Vent Worms (Ampharetidae, Alvinellidae, Vestimentifera) Provide New Mitochondrial COI Clock Calibration. *Cahiers de Biologie Marine*, page 4.

Chevaldonné, P., Jollivet, D., Vangriesheim, A., and Desbruyères, D. (1997). Hydrothermal-Vent Alvinellid Polychaete Dispersal in the Eastern Pacific. 1. Influence of Vent Site Distribution, Bottom Currents, and Biological Patterns. *Limnology and Oceanography*, 42(1):67–80.

Corliss, J. B., Dymond, J., Gordon, L. I., Edmond, J. M., von Herzen, R. P., Ballard, R. D., Green, K., Williams, D., Bainbridge, A., Crane, K., and van Andel, T. H. (1979). Submarine Thermal Springs on the Galápagos Rift. *Science*, 203(4385):1073–1083.

Cottin, D., Ravaux, J., Léger, N., Halary, S., Toullec, J.-Y., Sarradin, P.-M., Gaill, F., and Shillito, B. (2008). Thermal Biology of the Deep-Sea Vent Annelid *Paralvinella grasslei*: *In Vivo* Studies. *Journal of Experimental Biology*, 211(14):2196–2204.

Day, J. H. (1964). A Review of the Family Ampharetidae (Polychaeta). *Annals of the South African Museum. Annale van die Suid-Afrikaanse Museum*, 48:97–120.

Desbruyères, D., Hashimoto, J., and Fabri, M.-C. (2006). Composition and Biogeography of Hydrothermal Vent Communities in Western Pacific Back-Arc Basins. In Christie, D. M., Fisher, C. R., Lee, S.-M., and Givens, S., editors, *Geophysical Monograph Series*, volume 166, pages 215–234. American Geophysical Union, Washington, D. C.

Desbruyères, D. and Laubier, L. (1980). *Alvinella pompejana* gen. sp. nov., Ampharetidae Aberrant des Sources Hydrothermales de la Ride Est-Pacifique. *Oceanologica Acta*, 3:267–274.

Desbruyères, D. and Laubier, L. (1982). *Paralvinella grasslei*, new Genus, new Species of Alvinellinae (Polychaeta: Ampharetidae) from the Galapagos Rift Geothermal Vents. *Proceedings of the Biological Society of Washington*, 95:484–494.

Desbruyères, D. and Laubier, L. (1986). Les Alvinellidae, une Famille Nouvelle d'Annélides Polychètes Inféodées aux Sources Hydrothermales Sous-Marines: Systématique, Biologie et Ecologie. *Canadian Journal of Zoology*, 64(10):2227–2245.

Desbruyères, D. and Laubier, L. (1989). *Paralvinella hessleri*, New Species Of Alvinellidae (Polychaeta) from the Mariana Back-Arc Basin Hydrothermal Vents. *Proceedings of the Biological Society of Washington*, 102:761–767.

Desbruyères, D. and Laubier, L. (1991). Systematics, Phylogeny, Ecology and Distribution of the Alvinellidae (Polychaeta) from Deep-Sea Hydrothermal Vents. *Ophelia Supplement*, 5:31–45.

Desbruyères, D. and Laubier, L. (1993). New Species of Alvinellidae (Polychaeta) from the North Fiji Back-Arc Basin Hydrothermal Vents (Southwestern Pacific). *Proceedings of the Biological Society of Washington*, 106:225–236.

Detinova, N. (1988). New Species of Polychaetous Annelids from Hydrothermal Vents of the Juan-de-Fuca Ridge (Pacific-Ocean). *Zoologichesky Zhurnal*, 67(6):858–864.

Dilly, G. F., Young, C. R., Lane, W. S., Pangilinan, J., and Girguis, P. R. (2012). Exploring the Limit of Metazoan Thermal Tolerance via Comparative Proteomics: Thermally Induced Changes in Protein Abundance by Two Hydrothermal Vent Polychaetes. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3347–3356.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic Consistency-Based Multiple Sequence Alignment. *Genome Research*, 15(2):330–340.

Donath, A. and Stadler, P. F. (2018). Split-inducing indels in phylogenomic analysis. *Algorithms for Molecular Biology*, 13(1):12.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, 4(5):e88.

Eilertsen, M. H., Kongsrud, J. A., Alvestad, T., Stiller, J., Rouse, G. W., and Rapp, H. T. (2017). Do Ampharetids Take Sedimented Steps Between Vents and Seeps? Phylogeny and Habitat-Use of Ampharetidae (Annelida, Terebelliformia) in chemosynthesis-Based Ecosystems. *BMC Evolutionary Biology*, 17(1):222.

Fletcher, W. and Yang, Z. (2009). INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution*, 26(8):1879–1888.

Fontanillas, E., Galzitskaya, O. V., Lecompte, O., Lobanov, M. Y., Tanguy, A., Mary, J., Girguis, P. R., Hourdez, S., and Jollivet, D. (2017). Proteome Evolution of Deep-Sea Hydrothermal Vent Alvinellid Polychaetes Supports the Ancestry of Thermophily and

Subsequent Adaptation to Cold in Some Lineages. *Genome Biology and Evolution*, page evw298.

Gagnière, N., Jollivet, D., Boutet, I., Brélivet, Y., Busso, D., Da Silva, C., Gaill, F., Higuet, D., Hourdez, S., Knoops, B., Lallier, F., Leize-Wagner, E., Mary, J., Moras, D., Perrodou, E., Rees, J.-F., Segurens, B., Shillito, B., Tanguy, A., Thierry, J.-C., Weissenbach, J., Wincker, P., Zal, F., Poch, O., and Lecompte, O. (2010). Insights into Metazoan Evolution from *Alvinella pompejana* cDNAs. *BMC Genomics*, 11(1):634.

Gaill, F. and Hunt, S. (1986). Tubes of Deep Sea Hydrothermal Vent Worms *Riftia pachyptila* (Vestimentifera) and *Alvinella pompejana* (Annelida). *Marine Ecology Progress Series*, 34:267–274.

Girguis, P. R. and Lee, R. W. (2006). Thermal Preference and Tolerance of Alvinellids. *Science*, 312(5771):231–231.

Glasby, C. J., Hutchings, P. A., and Hall, K. (2004). Assessment of Monophyly and Taxon Affinities within the Polychaete Clade Terebelliformia (Terebellida). *Journal of the Marine Biological Association of the United Kingdom*, 84(5):961–971.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome. *Nature Biotechnology*, 29(7):644–652.

Gu, X. and Li, W.-H. (1995). The Size Distribution of Insertions and Deletions in Human and Rodent Pseudogenes Suggests the Logarithmic Gap Penalty for Sequence Alignment. *Journal of Molecular Evolution*, 40(4):464–473.

Han, Y., Zhang, D., Wang, C., and Zhou, Y. (2021). Out of the Pacific: A New Alvinellid

Worm (Annelida: Terebellida) From the Northern Indian Ocean Hydrothermal Vents. *Frontiers in Marine Science*, 8:669918.

Haymon, R. M., Koski, R. A., and Sinclair, C. (1984). Fossils of Hydrothermal Vent Worms from Cretaceous Sulfide Ores of the Samail Ophiolite, Oman. *Science*, 223(4643):1407–1409.

Hessler, R. R. and Lonsdale, P. F. (1991). Biogeography of Mariana Trough Hydrothermal vVnt Communities. *Deep Sea Research Part A. Oceanographic Research Papers*, 38(2):185–199.

Hickey, D. A. and Singer, G. A. (2004). Genomic and Proteomic Adaptations to Growth at High Temperature. *Genome Biology*, page 7.

Hoff, K. J. and Stanke, M. (2013). WebAUGUSTUS - a Web Service for Training AUGUSTUS and Predicting Genes in Eukaryotes. *Nucleic Acids Research*, 41(W1):W123–W128.

Hourdez, S. and Lallier, F. H. (2007). Adaptations to Hypoxia in Hydrothermal-Vent and Cold-Seep Invertebrates. *Reviews in Environmental Science and Bio/Technology*, 6(1-3):143–159.

Huang, X. and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9):868–877.

Hudson, R. R. and Coyne, J. A. (2002). Mathematical Consequences Of The Genealogical Species Concept. *Evolution*, 56(8):1557–1565.

Jacobs, D. K. and Lindberg, D. R. (1998). Oxygen and Evolutionary Patterns in the Sea: Onshore/Offshore Trends and Recent Recruitment of Deep-Sea Faunas. *Proceedings of the National Academy of Sciences*, 95(16):9396–9401.

Jang, S.-J., Park, E., Lee, W.-K., Johnson, S. B., Vrijenhoek, R. C., and Won, Y.-J. (2016).

Population Subdivision of Hydrothermal Vent Polychaete *Alvinella pompejana* across Equatorial and Easter Microplate Boundaries. *BMC Evolutionary Biology*, 16(1):235.

Jollivet, D., Chevaldonné, P., and Planque, B. (2022). Hydrothermal-Vent Alvinellid Polychaete Dispersal in the Eastern Pacific. 2. A Metapopulation Model Based on Habitat Shifts. page 16.

Jollivet, D., Desbruyères, D., Ladrat, C., and Laubier, L. (1995). Evidence for Differences in the Allozyme Thermostability of Seep-Sea Hydrothermal Vent Polychaetes (Alvinellidae): a Possible Selection by Habitat. *Marine Ecology Progress Series*, 123:125–136.

Jollivet, D. and Hourdez, S. (2021). 7.7.4 Alvinellidae Desbruyères & Laubier, 1986. *Pleistoannelida, Sedentaria III and Errantia I*, 3:18.

Jouin, C. and Gaill, F. (1990). Gills of Hydrothermal Vent Annelids: Structure, Ultrastructure and Functional Implications in Two Alvinellid Species. *Progress in Oceanography*, 24(1-4):59–69.

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J. T., Moroz, L. L., Lieb, B., and Halanych, K. M. (2016). Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. *Systematic Biology*, page syw079.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference. *Bioinformatics*, 35(21):4453–4455.

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model. *BMC Evolutionary Biology*, 7(Suppl 1):S4.

Lartillot, N. and Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Le, S. Q., Lartillot, N., and Gascuel, O. (2008). Phylogenetic Mixture Models for Proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):3965–3976.

Le Bris, N. and Gaill, F. (2007). How Does the Annelid *Alvinella pompejana* Deal with an Extreme Hydrothermal Environment? *Reviews in Environmental Science and Bio/Technology*, 6(1-3):197–221.

Lelièvre, Y., Sarrazin, J., Marticorena, J., Schaal, G., Day, T., Legendre, P., Hourdez, S., and Matabos, M. (2018). Biodiversity and Trophic Ecology of Hydrothermal Vent FFauna Associated with Tubeworm Assemblages on the Juan de Fuca Ridge. *Biogeosciences*, 15(9):2629–2647.

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. (2007). A General Comparison of Relaxed Molecular Clock Models. *Molecular Biology and Evolution*, 24(12):2669–2680.

Liu, L., Yu, L., and Edwards, S. V. (2010). A Maximum Pseudo-Likelihood Approach for Estimating Species Trees under the Coalescent Model. *BMC Evolutionary Biology*, 10(1):302.

López-García, P., Philippe, H., Gail, F., and Moreira, D. (2003). Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. *Proceedings of the National Academy of Sciences*, 100(2):697–702.

Mammerickx, J., Herron, E. M., and Dorman, L. (1980). Evidence for Two Fossil Spreading Ridges in the Southeast Pacific. *Geological Society of America Bulletin*, 91(5):263.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes,

X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome Sequencing in Microfabricated High-Density Picolitre Reactors. *Nature*, 437(7057):376–380.

Mary, J., Rogniaux, H., Rees, J.-F., and Zal, F. (2010). Response of *Alvinella pompejana* to Variable Oxygen Stress: A Proteomic Approach. *Proteomics*, 10(12):2250–2258.

Matabos, M., Plouviez, S., Hourdez, S., Desbruyères, D., Legendre, P., Warén, A., Jollivet, D., and Thiébaut, E. (2011). Faunal Changes and Geographic Crypticism Indicate the Occurrence of a Biogeographic Transition Zone Along the Southern East Pacific Rise: Biogeographic Transition Zone Along the Southern EPR. *Journal of Biogeography*, 38(3):575–594.

McHugh, D. (1989). Population Structure and Reproductive Biology of Two Sympatric Hydrothermal Vent Polychaetes, *Paralvinella pandorae* and *P. palmiformis*. *Marine Biology*, 103(1):95–106.

Moalic, Y., Desbruyères, D., Duarte, C. M., Rozenfeld, A. F., Bachraty, C., and Arnaud-Haond, S. (2012). Biogeography Revisited with Network Theory: Retracing the History of Hydrothermal Vent Communities. *Systematic Biology*, 61(1):127.

Molloy, E. K. and Warnow, T. (2018). To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2):285–303.

Monaco, A. and Prouzet, P. (2015). Hydrothermal Vents: Oases at Depth. In Monaco, A.

and Prouzet, P., editors, *Marine Ecosystems*, pages 225–292. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Nakamura, K., Watanabe, H., Miyazaki, J., Takai, K., Kawagucci, S., Noguchi, T., Nemoto, S., Watsuji, T.-o., Matsuzaki, T., Shibuya, T., Okamura, K., Mochizuki, M., Orihashi, Y., Ura, T., Asada, A., Marie, D., Koonjul, M., Singh, M., Beedessee, G., Bhikajee, M., and Tamaki, K. (2012). Discovery of New Hydrothermal Activity and Chemosynthetic Fauna on the Central Indian Ridge at 18°–20°S. *PLoS ONE*, 7(3):e32965.

Nute, M., Chou, J., Molloy, E. K., and Warnow, T. (2018). The Performance of Coalescent-Based Species Tree Estimation Methods under Models of Missing Data. *BMC Genomics*, 19(S5):286.

Parker, E. and Gealey, W. K. (1985). Plate Tectonic Evolution of the Western Pacific-Indian Ocean Region. *Energy*, 10(3-4):249–261.

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R. S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., and Niehuis, O. (2017). Orthograph: a Versatile Tool for Mapping Coding Nucleotide Sequences to Clusters of Orthologous Genes. *BMC Bioinformatics*, 18(1):111.

Piccino, P., Viard, F., Sarradin, P.-M., Le Bris, N., Le Guen, D., and Jollivet, D. (2004). Thermal Selection of PGM Allozymes in Newly Founded Populations of the Thermotolerant Vent Polychaete *Alvinella pompejana*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1555):2351–2359.

Pradillon, F., Le Bris, N., Shillito, B., Young, C. M., and Gaill, F. (2005). Influence of Environmental Conditions on Early Development of the Hydrothermal Vent Polychaete *Alvinella pompejana*. *Journal of Experimental Biology*, 208(8):1551–1561.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE: Multiple

Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS ONE*, 6(9):e22594.

Ravaux, J., Hamel, G., Zbinden, M., Tasiemski, A. A., Boutet, I., Léger, N., Tanguy, A., Jollivet, D., and Shillito, B. (2013). Thermal Limit for Metazoan Life in Question: *In Vivo* Heat Tolerance of the Pompeii Worm. *PLoS ONE*, 8(5):e64074.

Roch, S. and Warnow, T. (2015). On the Robustness to Gene Tree Estimation Error (or Lack thereof) of Coalescent-Based Species Tree Methods. *Systematic Biology*, 64(4):663–676.

Rousset, V., Pleijel, F., Rouse, G. W., Erséus, C., and Siddall, M. E. (2007). A Molecular Phylogeny of Annelids. *Cladistics*, 23(1):41–63.

Rousset, V., Rouse, G. W., Feral, J.-P., Desbruyères, D., and Pleijel, F. (2003). Molecular and Morphological Evidence of Alvinellidae Relationships (Terebelliformia, Polychaeta, Annelida). *Zoologica Scripta*, 32(2):185–197.

Röhl, U. and Ogg, J. G. (1996). Aptian-Albian Sea Level History from Guyots in the Western Pacific. *Paleoceanography*, 11(5):595–624.

Savin, S. M. (1977). The History of the Earth's Surface Temperature During the Past 100 Million Years. *Annual Review of Earth and Planetary Sciences*, 5(1):319–355.

Sepkoski Jr., J. J. (2002). A Compendium of Fossil Marine Animal Genera. *Bulletins of American Paleontology*, pages 1–560.

Si Quang, L., Gascuel, O., and Lartillot, N. (2008). Empirical Profile Mixture Models for Phylogenetic Reconstruction. *Bioinformatics*, 24(20):2317–2323.

Simmons, M. P. and Ochoterena, H. (2000). Gaps as Characters in Sequence-Based Phylogenetic Analyses. *Systematic Biology*, page 13.

Somero, G. N., Childress, J. J., and Anderson, A. E. (1989). Transport, Metabolism, and Detoxification of Hydrogen Sulfide in Animals from Sulfide-Rich Marine Environments. *Critical Reviews in Aquatic Sciences*, 1(4):591–614.

Stiller, J., Rousset, V., Pleijel, F., Chevaldonné, P., Vrijenhoek, R. C., and Rouse, G. W. (2013). Phylogeny, Biogeography and Systematics of Hydrothermal Vent and Methane Seep *Amphisamytha* (Ampharetidae, Annelida), with Descriptions of Three New Species. *Systematics and Biodiversity*, 11(1):35–65.

Stiller, J., Tilic, E., Rousset, V., Pleijel, F., and Rouse, G. W. (2020). Spaghetti to a Tree: A Robust Phylogeny for Terebelliformia (Annelida) Based on Transcriptomes, Molecular and Morphological Data. *Biology*, 9(4):73.

Thorne, J. L., Kishino, H., and Painter, I. S. (1998). Estimating the Rate of Evolution of the Rate of Molecular Evolution. *Molecular Biology and Evolution*, 15(12):1647–1657.

Tunnicliffe, V. (1988). Biogeography and Evolution of Hydrothermal-Vent Fauna in the Eastern Pacific Ocean. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 233(1272):347–366.

Tunnicliffe, V., Desbruyères, D., Jollivet, D., and Laubier, L. (1993). Systematic and Ecological Characteristics of *Paralvinella sulfincola* Desbruyères and Laubier, a New Polychaete (Family Alvinellidae) from Northeast Pacific Hydrothermal Vents. *Canadian Journal of Zoology*, 71(2):286–297.

van Noort, V., Bradatsch, B., Arumugam, M., Amlacher, S., Bange, G., Creevey, C., Falk, S., Mende, D. R., Sinning, I., Hurt, E., and Bork, P. (2013). Consistent Mutational Paths Predict Eukaryotic Thermostability. *BMC Evolutionary Biology*, 13(1):7.

Vrijenhoek, R. C. (2013). On the Instability and Evolutionary Age of Deep-Sea Chemosynthetic Communities. *Deep Sea Research Part II: Topical Studies in Oceanography*, 92:189–200.

Wang, G.-Z. and Lercher, M. J. (2010). Amino Acid Composition in Endothermic Vertebrates is Biased in the Same Direction as in Thermophilic Prokaryotes. *BMC Evolutionary Biology*, 10(1):263.

Wang, H.-C., Li, K., Susko, E., and Roger, A. J. (2008). A Class Frequency Mixture Model that Adjusts for Site-Specific Amino Acid Frequencies and Improves Inference of Protein Phylogeny. *BMC Evolutionary Biology*, 8(1):331.

Watanabe, H. and Beedessee, G. (2015). Vent Fauna on the Central Indian Ridge. In Ishibashi, J.-i., Okino, K., and Sunamura, M., editors, *Subseafloor Biosphere Linked to Hydrothermal Systems*, pages 205–212. Springer Japan, Tokyo.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. *Genome Biology*, 15(3):R46.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Zal, F., Jollivet, D., Chevaldonné, P., and Desbruyères, D. (1995). Reproductive Biology and Population Structure of the Deep-Sea Hydrothermal Vent Worm *Paralvinella grasslei* (Polychaete: Alvinellidae) at 13°N on the East Pacific Rise. *Marine Biology*, 122:637–648.

Zdobnov, E. M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., and Kriventseva, E. V. (2021). OrthoDB in 2020: Evolutionary and Functional Annotations of Orthologs. *Nucleic Acids Research*, 49(D1):D389–D393.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees. *BMC Bioinformatics*, 19(S6):153.

Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K. M., and Struck, T. H.

(2011). Detecting the Symplesiomorphy Trap: a Multigene Phylogenetic Analysis of Terebelliform Annelids. *BMC Evolutionary Biology*, 11(1):369.

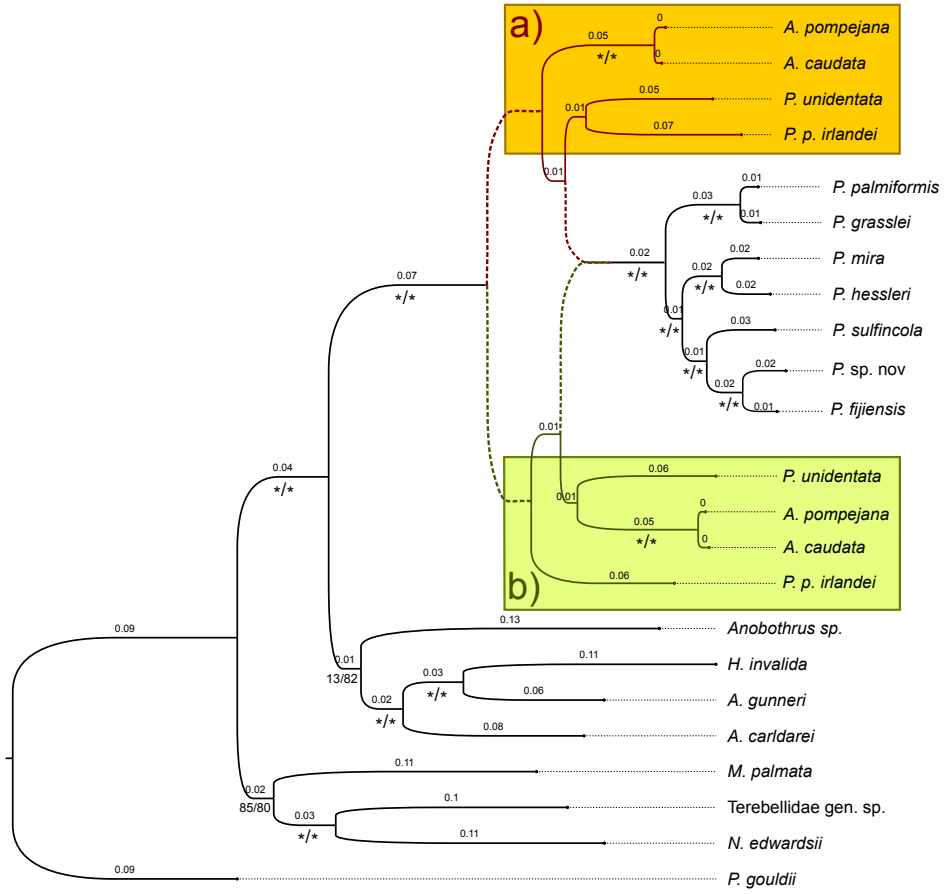

Fig. 1: Best ranking topologies according to different phylogenetic inference methods. (a) clustering of all *Paralvinella* species together, (b) *P. p. irlandei* basal to the family Alvinellidae and clustering of *P. unidentata* with the *Alvinella* species. Branch length are optimized under the JTT+G model. Boostraps are given in the following order: GTR+G/JTT+G. (*) indicates a bootstrap of 100.

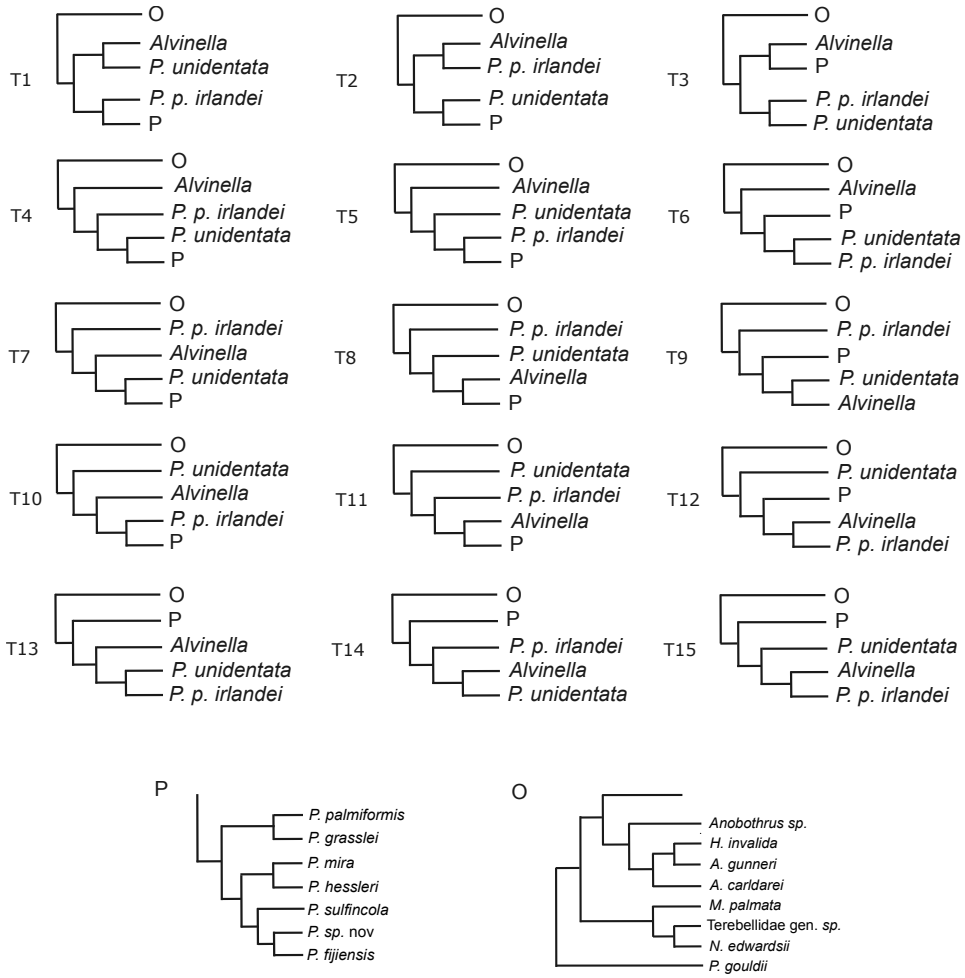

Fig. 2: Constrained topologies used as user trees under different phylogenetic inference methods. The topologies differ only in the arrangement of the species *P. unidentata*, *P. p. irlandei*, and the genera *Paralvinella* and *Alvinella* at the root of Alvinellidae. The *Alvinella* group comprises the species *A. pompejana* and *A. caudata*, and the *Paralvinella* group (P) comprises all species *Paralvinella* except *P. p. irlandei* and *P. unidentata*. Trees are rooted with a set of outgroups (O) which comprises species from other closely-related families of Terebelliformia (Terebellidae+Ampharetidae+*P. gouldii*).

	BaseML 895 genes	CodeML 385 genes	Phylobayes 385 genes	RAxML + MP-EST 966 genes	RAxML + MP-EST 473 genes	Phylobayes +ASTRAL 478 genes	GAP ML soft filter 2050 indels	GAP ML hard filter 1416 indels
T1	-1623	-457	-161±23	-274	-39	0	-44	-13
T2	-1922	-367	-149±21	-534	-73	-22048	-43	-29
T3	-1330	-251	-71±22	-595	-97	-12618	-34	-32
T4	-734	-251	-107±22	-534	-73	-15869	-43	-37
T5	-386	-221	-90±22	-584	-92	-12732	-32	-38
T6	0	-45	0±22	-584	-92	-10414	-53	-31
T7	-593	0	-23±23	-265	-45	-17729	-22	-32
T8	-655	-16	-12±22	-265	-51	-15839	-18	-40
T9	-537	-60	-23±22	0	0	-1888	0	-16
T10	-1164	-328	-127±23	-595	-98	-14983	-23	-30
T11	-1328	-318	-118±23	-595	-98	-17598	-22	-37
T12	-1520	-318	-124±22	-595	-98	-22400	-44	-33
T13	-1669	-306	-74±23	-584	-84	-1990	-40	0
T14	-1882	-420	-128±23	-274	-39	-1235	-20	-6
T15	-2406	-455	-153±22	-584	-84	-6131	-46	-7

Table 1: Scores obtained for the 15 Alvinellidae topologies under different phylogenetic models. Maximum likelihood: BaseML (concatenation of nucleotide transcripts), CodeML (concatenation of translated proteins), RAxML+MP-EST 997 genes (single-gene coalescent method on nucleotide transcripts), RAxML+MP-EST 473 genes (single-gene coalescent method on translated proteins), GAP ML 'soft' and 'hard' filtering (on indels of a concatenated set of transcripts), Bayesian likelihood: Phylobayes (concatenation of translated proteins). Quartet score (Zhang et al., 2018): Phylobayes+ASTRAL (single-gene coalescent method on translated proteins). For each method, the topology with the highest score is set to 0. The difference between the tree score and this maximum is shown for the other topologies.

filter	r	$l_1(0)$	$l_2(0)$
L ⩽ 40 , n ⩽ 10, n/L ⩽ 2	0.35	2.0	2.39
L ⩽ 10, n ⩽ 4, n/L ⩽ 2	0.67	2.25	0.53
L ⩽ 5, n ⩽ 1, n/L ⩽ 1	1.44	2.36	2.66

a) ML optimization (Nelder-Mead) of the parameters r, $l_1(0)$ and $l_2(0)$

	T	I	σ	π	ML
T1 hard filter	7.0×10^{-3}	0	0	0.05	-15,426.82
T6 hard filter	7.0×10^{-3}	0	0	0.40	-15,445.04
T9 hard filter	7.0×10^{-3}	0	0	0.58	-15,430.46
T1 soft filter	1.3×10^{-3}	0.92	1.23	0.39	-27,996.08
T6 soft filter	1.3×10^{-3}	0.92	1.23	0.42	-28,041.82
T9 soft filter	1.4×10^{-3}	0.92	1.23	0.38	-27,989.00

b) ML optimization (Truncated Newton algorithm) of the parameters T, I, σ and π

Table 2: ML optimization of the indel model of evolution on different topologies and with different filtering levels. L is the maximum allowed length of gaps, and n is the maximum number of insertions and deletions allowed to explain the formation of a gap parsimoniously. To be included, a gap must meet the filtering criteria for at least one of the 15 topologies shown in Figure 2.

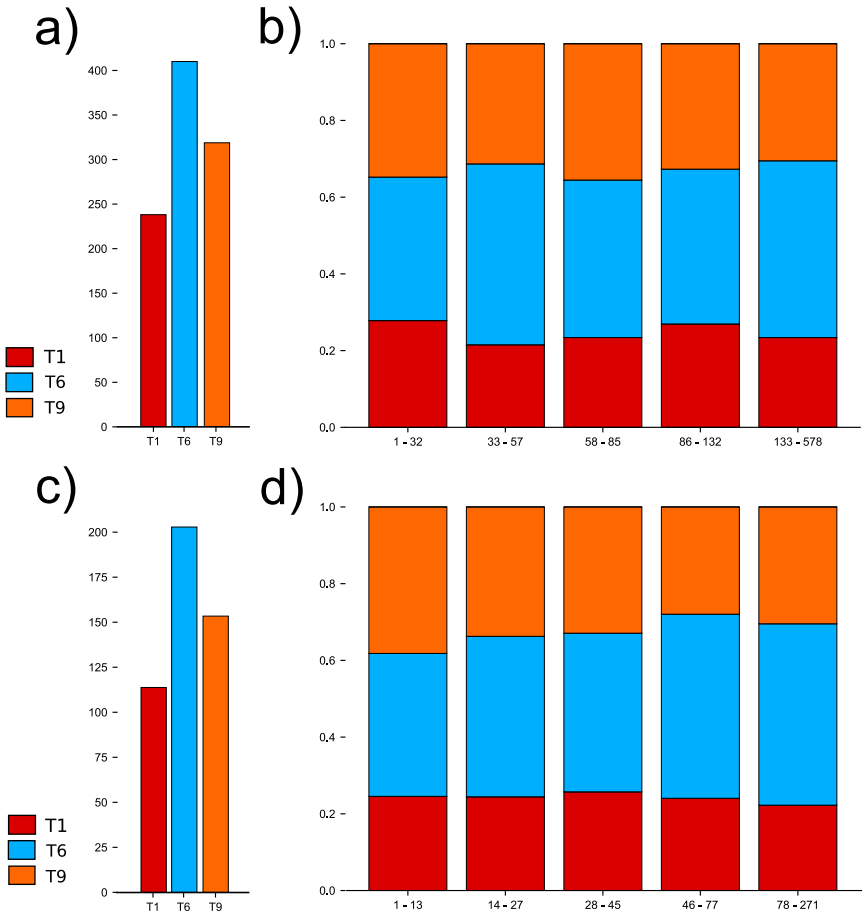

Fig. 3: Number of best-ranked genes in topologies T1, T6 and T9. For each gene, the best-ranked tree is deduced from its likelihood score according to RAxML-ng. (a) Total number of best-ranked CDS for topologies T1, T6 or T9. (b) Distribution of genes according to the number of phylogenetically informative sites for coding sequences.(c) Total number of best-ranked translated proteins for topologies T1, T6 or T9. (d) Distribution of genes according to the number of phylogenetically informative sites for translated protein sequences.

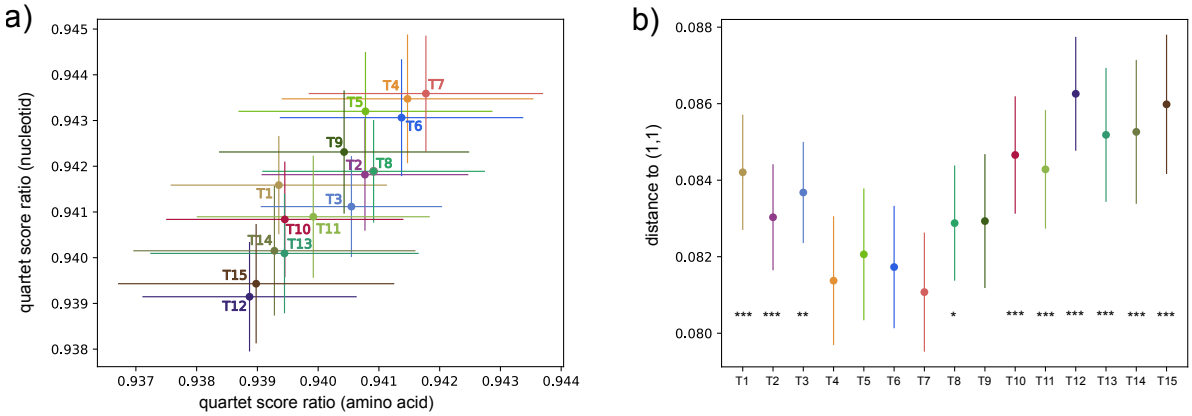

Fig. 4: Evaluation of species tree topologies by ASTRAL for the 15 topologies, with a 95% confidence interval. Each evaluation is the result of 100 replicates on gene resampling in the gene sets. (a) biplots of quartet scores (in %) for each topology from the protein and coding sequence datasets. A higher proportion indicates a species-tree topology more in agreement with our set of gene trees. (b) Euclidean distance of each topology to the point (1,1), which should represent the 'true' species-tree topology that verifies all quartet trees both for the two sequence datasets. A smaller distance indicates that the topology is closer to the ideal tree. T7 is the topology closest to (1,1). The distance to (1,1) relative to T7 is tested for every other topology: *: $p-value < 5\times10^{-2}$, **: $p-value < 1\times10^{-2}$, ***: $p-value < 1\times10^{-3}$.

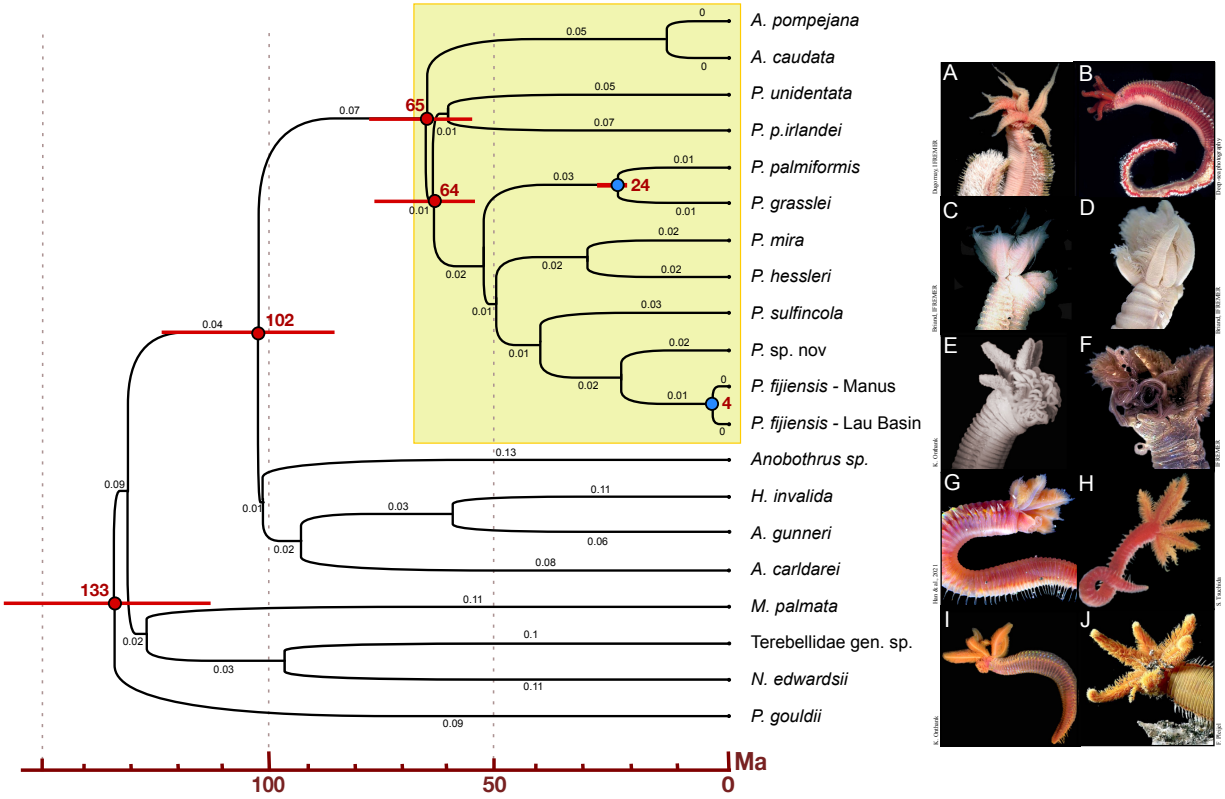

Fig. 5: Chronogram of the family Alvinellidae according to the main topology T6 under the CIR model and two calibration points (in blue). Branch lengths obtained with CodeML as expected numbers of amino-acid substitutions are shown, as well as the ages of some nodes of interest with 95% confidence intervals, millions of years ago (Ma). Illustrations: (A) *A. pompejana*, (B) *A. caudata*, (C) *P. unidentata*, (D) *P. p. irlandei*, (E) *P. palmiformis*, (F) *P. grasslei*, (G) *P. mira*, (H) *P. hessleri*, (I) *P. sulfincola*, (J) *P. fijiensis*.