Systematic Biology (2023), **0**, 0, *pp.* 1–46
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

The Alvinellidae are a family of worms that are endemic to deep-sea hydrothermal vents in the Pacific and Indian Oceans. These annelid worms, a sister group to the Ampharetidae, occupy a wide range of thermal habitats. The family includes 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 been studied extensively. It is, however, a complex case where molecular phylogenies give conflicting results, especially concerning 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 currently described or in description. The results obtained by the most popular phylogenetic inference models (gene concatenation with maximum likelihood, or coalescent-based methods from gene trees) are

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

compared using a series of ampharetid and terebellid outgroups.

Our study shows that a high number of gene trees support the hypothesis of the monophyly of the *Paralvinella* genus, as initially proposed by Desbruyères and Laubier, in which the species *Paralvinella pandorae* and *Paralvinella unidentata* are more closely related within the subgenus *Nautalvinella*. However, the global phylogenetic signal favors the hypothesis of paraphyly for this genus, with *P. pandorae* being sister species of the other Alvinellidae. Gene trees separated equally between these two hypotheses, making it difficult to draw conclusionsabout the initial split of the MCRA as different genomic regions seem to have very different phylogenetic stories. According to molecular dating, the radiation of the Alvinellidae was rapid and took place in a short period of time between 70 and 80 million years ago. This is reflected at the genomic scale by high rates of incomplete lineage sorting between the first ancestral lineages with probable gene transfers between the ancestors of *Alvinella*, *Nautalvinella*, and the rest of the *Paralvinella* lineages.

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

Introduction

After the discovery of hydrothermal vents and their associated communities on the Galapagos rift in the late 1970s, the Pompeii worm *Alvinella pompajena* Desbruyères and Laubier (1980), was one of the first emblematic species collected from the vent chimney environment following the sampling of a "black smoker" chimney at 21°N on the East Pacific Rise (EPR) (Monaco and Prouzet, 2015). Two morphological types living in close association (syntopy), initially viewed as ontogenic forms, were later found to represent two distinct species, *A. pompejana* and *Alvinella caudata* Desbruyères and Laubier (1986) (Jollivet and Hourdez, 2021). While the two *Alvinella* species were initially described as aberrant forms in the family Ampharetidae, some obvious specificities, such as the lack of

separation between the head and the rest of the body, a process of cephalization of the gills, uncini with a reduced number of teeth and notopodia with simple dorsal chaetae, led the authors to suggest the creation of the subfamily initially named Alvinellinae (Desbruyères and Laubier, 1980, 1982). Following the discovery of the species *Paralvinella grasslei* Desbruyères and Laubier (1982), at several sites in the EPR and the Guaymas basin (eastern Pacific), the Alvinellinae were split into two genera, *Alvinella* and *Paralvinella*. The distinction between the two genera was based in particular on the shapes of the mouth apparatus and of the secondary filaments of the gills, which are lamellar in *Alvinella* and filamentous in *Paralvinella* Desbruyères and Laubier (1982).

Subsequently, new species of Alvinellinae were discovered, all endemic to hydrothermal vents in the Pacific Ocean. In the south-east Pacific, two additional species were found and described as *Paralvinella pandorae irlandei* Desbruyères and Laubier (1986), (EPR) *P. p. irlandei* and *Paralvinella bactericola* Desbruyères and Laubier (1991) (Guaymas basin). In the north Pacific, several new species were described from the Juan de Fuca Ridge with *Paralvinella pandorae pandorae* Desbruyères and Laubier (1986), *Paralvinella palmiformis* Desbruyères and Laubier (1986), *Paralvinella dela* Detinova (1988), as well as *Paralvinella sulfincola* Desbruyères and Laubier (1993) found on top of vent chimneys. Finally, three additional species were added to the family from expeditions in the western Pacific: *Paralvinella hessleri* Desbruyères and Laubier (1989) is found in the Marianas back-arc basins and Okinawa Trough, while *Paralvinella fijiensis* Desbruyères and Laubier (1993) and *Paralvinella unidentata* Desbruyères and Laubier (1993) live in the south-west Pacific.

As the number of species in the subfamily increased, subsequent studies using both morphological characters and molecular data demonstrated that inside Terebelliformia, alvinellid worms form a monophyletic group, sister-group-to, but not nested within the Ampharetidae. The family was accordingly renamed Alvinellidae (Rousset et al., 2003; Glasby et al., 2004; Stiller et al., 2020). 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 expanded beyond Pacific waters, supporting the purported observation of a small population of alvinellid worms in the Solitaire field on the Central Indian Ridge, though they have beenneither sampled nor taxonomically resolved (Nakamura et al., 2012). Finally, another species, whose description is still in progress, was first discovered in 2011 from the Nifonea vent site along the volcanic arc of Vanuatu in the Manus Bassin. This species is referred to as *Paralvinella* sp. nov. in this study.

Alvinellid worms probably have a long evolutionary history of speciation, which could date back between 70 and 200 millions of years ago (Jollivet and Hourdez, 2021). Considering the amino acid composition of reconstructed ancestral protein sequences and the very low evolutionary rates of proteins in the thermophilic lineages (Fontanillas et al., 2017) led Fontanillas et al. to hypothesise that the ancestor of the Alvinellidae was a thermophilic species, suggesting an old colonization of the hydrothermal environment. Yet the family has now diversified to cope with a wider variety of vent conditions. Alvinellidae notably include two of the most thermotolerant animals known to date, namely *A. pompejana* and its ecological homolog *P. sulfincola* thriving between 40 and 50°C (Girguis and Lee, 2006; Ravaux et al., 2013), while species such as *P. grasslei* are comfortable in temperatures around 15°C (Cottin et al., 2008).

The ecological diversity of these closely-related worms is particularly intriguing. The question of the evolutionary path taken by the alvinellids naturally arises, as well as the different stages that allowed its dispersal 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. More specifically, the place of *P. p. irlandei* and *P. p. pandorae* in the phylogeny, which

determines the monophyly of the genus *Paralvinella*, remains uncertain. Resolving this question would represent a important step in our understanding of the relationships between the early lineages of the family.

Several partial phylogenies of the Alvinellidae have been proposed. Based on detailed observations 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 two 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 buccal tentacles of males are different (either three-lobed, coiled and tapered, or absent) (Jollivet and Hourdez, 2021). Other non-morphological differences between the two genera have been noted: *Alvinella* species secrete thick parchment-like tubes including inorganic material and have epibiotic bacteria, whereas *Paralvinella* species secrete mucus tubes or cocoons (with the exception of *P.* sp. nov., personal communication) and are devoid of filamentous epibiotic bacteria (Desbruyères and Laubier, 1991).

Within the *Paralvinella* species, Desbruyères and Laubier further proposed to group the species *P. p. pandorae*, *P. p. irlandei* and *P. unidentata* within the subgenus *Nautalvinella*, considering the comb-like insertion of the secondary filaments of the gills as well as the lack of 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).

Recent molecular phylogenies, on the other hand, have produced conflicting results regarding the subdivision of the family into two monophyletic genera. Jollivet et al. supported this view, as well as the existence of the *Nautalvinella* subgenus, based on

enzyme isoforms similarities (Jollivet et al., 1995). Another study by Jollivet and Hourdez, based on a set of 278 orthologous transcripts, came to the same conclusion, with *P. p. irlandei* being a sister species to the other *Paralvinella* species (Jollivet and Hourdez, 2021). Counter to this, a recent wide phylogeny for Terebelliformia by Stiller et al., based on several hundred orthologous genes derived from transciptomic datasets, proposed with high support that the *Paralvinella* genus is polyphyletic, with *P. p. irlandei* being sister to the other Alvinellidae (Stiller et al., 2020). Interestingly, this later study also fitted a previous mitochondrial sequence analysis done by (Vrijenhoek, 2013) but neither of them include sequences from *P. unidentata*, in contrast to studies suggesting the monophyly of the genus, which then anchor the clades *P. p. irlandei*, *P. p. pandorae* and *P. unidentata* together in the phylogeny. It appears that the phylogenetic relationship between the three *Nautalvinella* species and the *Alvinella* species is therefore key to understanding the deep phylogeny of the Alvinellidae.

Here, we propose a new molecular phylogeny of the family Alvinellidae based on several hundred 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 described species of Alvinellidae, with the exception of *P. p. pandorae*, and the rare species *P. dela* and *P. bactericola* for which no exhaustive molecular data or samples exist to date. This phylogeny is rooted with species of the sister group Ampharetidae, as well as more distantly-related species of Terebelliformia and the outgroup coastal pectinariid worm *Pectinaria gouldii*, in accordance with the currently accepted phylogeny of the Terebelliformia worms (Rousset et al., 2007; Stiller et al., 2020). Finally, the results obtained by different methods are compared, following 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). Two individuals from 9°50N/EPR were used to obtain the genome, one given by C.G. Cary and the other one collected during the Mescal 2012 French cruise with ROV *Victor6000* and RV L'Atalante. The software Augustus 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. (2010) for details). 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.

RNAseq reads for the species *P. gouldii*, *Anobothrus* sp., *Amphicteis gunneri*, *Amphisamytha carldarei* and *Hypania invalida* were downloaded from the NCBI SRA database (accession numbers SRR2057036, SRR11434464, SRR11434467, SRR11434468, SRR5590961) (Kocot et al., 2016; Stiller et al., 2020). Total RNA extractions from flash-frozen tissues 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 on board of the N/O L'Atalante and using 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 locations of the alvinellid worms collected along the East Pacific Rise are provided in Fontanillas et al. (2017). Other alvinellid species and the vent terebellid were sampled from 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 edwardsi*, 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 after tissue grinding using a ball mill. RNAseq libraries were produced at Genome Québec following a polyA purification of mRNAs, and sequenced using the Novaseq 6000 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 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 table 1 of the Supplementary data, together with the associated command lines. Transcriptomes are available from the NCBI BioProject repository with the following accession numbers (*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 (ORFs) were then identified with Transdecoder v.5.5.0, firstly trained with the 500 longest sequences accessible from each transcriptome. The predicted ORFs were retained, stripped of the 3' and 5' UTR regions. The metrics associated with these different steps are also presented in the table 1 of the Supplementary data.

In total, 25 transcriptomes were assembled and compared, from 19 different Terebelliformia species including 11 species of the family 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.

The nucleotide sequences were aligned for each gene with MACSE v.2.05, which is a codon-gaps aligner for coding sequences (Ranwez et al., 2011). The translated amino acid

sequences were aligned with Probcons v.1.12 (Do et al., 2005).

From these sequences alignments, we obtained the following datasets:

1. a set of 657 orthologous genes, in nucleotides, present in at least 20 transcriptomes that must include a gene ortholog for the species *P. p. irlandei* and *P. unidentata*, and, at least, a gene ortholog for one of the two *Alvinella* species (*A. pompejana* or *A. caudata*) . Aligned fragments shorter than 60 nucleotides between two subsequent gaps were discarded to improve the alignment reliability. The third nucleotide of each codon was also removed to reduce the saturation of the phylogenetic signal. Sequences shorter than 20% of the length of the total gene alignment were also discarded. Finally, each gene was tested to guarantee that the phylogenetic assumptions about residue homogeneity and stationarity were not violated (Naser-Khdour et al., 2019): for each gene, biased sequences were eliminated until the alignment did not fail this test at a 5% threshold. Uninformative sites at which at most one species had a residue were also removed;

2. a set of 699 orthologous amino-acid translated genes with constraints similar to those described for the nucleotidic sequences but with the removal of fragments shorter than 20 amino-acids when included between two subsequent gaps.

The composition and stationarity tests were performed with IQ-TREE 2.0.3 (Minh et al., 2020).

These sets were then concatenated into two supergenes. The first, in nucleotides (two first codon positions of the coding sequence), contained 657 genes and 499,036 sites, with *A. caudata* being the longest sequence with 461,042 sites, and *P. gouldii* the shortest with 245,846 sites. The second, in amino-acids, contained 699 genes and 277,900 sites, with *A. caudata* being the longest sequence (255,557 sites) and *P. gouldii* the shortest (139,067 sites).

Phylogenetic inference

Identifying well-resolved clades A global unrooted phylogeny was evaluated from the supergenes with the maximum likelihood (ML) approach implemented in IQ-TREE 2.0.3. The supergenes were first partitioned according to each gene, and we used the merging strategy implemented in IQ-TREE to retrieve relevant merged partitions (Chernomor et al., 2016). For each merged partition, ModelFinder, as implemented under IQ-TREE, was used to infer the best-fitting evolutionary model according to the minimized Bayesian Information Criteria (BIC) (Kalyaanamoorthy et al., 2017). For nucleotides and amino-acids sequences, heterogeneity of the evolutionary rate between sites was modelled by a Γ distribution with four discrete classes (Yang, 1994). For amino-acid sequences, equilibrium amino-acid frequencies were estimated from either empirical or ML frequencies, and mixture models were also added to the standard test of ModelFinder (Si Quang et al., 2008; Le et al., 2008, 2012). Finally, the robustness of the tree was inferred *via* 1,000 ultra-fast bootstraps (Hoang et al., 2018).

Testing alternative topologies In order to better resolve the relationships between the Alvinellidae species at the root of the family, we compared 15 alternative topologies one to each other. In these topologies, both the *Paralvinella* clade (excluding *P. p. irlandei* and *P. unidentata*) and the set of outgroup species were well resolved in the first step of the tree reconstruction, and subsequently constrained to remain identical among the 15 topologies. The outgroup species arrangement obtained from our own results was also consistent with the Terebelliformia phylogeny proposed by Stiller et al. (2020).

The 15 topologies were evaluated in a ML framework using IQ-TREE, using the GTR+Γ4+F model for the nucleotide-encoded supergene taken as a single partition, and the LG+Γ4+F model for the amino-acid encoded supergene also taken as one partition. The evolutionary model used for all supergenes corresponds to the most commonly chosen models obtained under a complete gene partition. The topologies were then tested against

the best topology obtained for each supergene using the AU test at a 5% threshold with 10,000 RELL replicates implemented in IQ-TREE (Shimodaira, 2002). This test gives the probability that a topology scores better than the reference topology after a site-to-site resampling.

The topologies were also tested against each separated gene in the coalescent framework. To this end, gene trees, for either nucleotide or amino-acid encoded sequence alignments, were first optimized using IQ-TREE. Substitution models were tested with the ModelFinder implementation of IQ-TREE, including mixture models for amino-acid sequences. Branches with low bootstrap support (<10%) were contracted, and species trees were then scored from the gene trees using ASTRAL v.5.7.8 (Zhang et al., 2018). The multiscale bootstrap algorithm underlying the AU test described in Shimodaira (2002) was recoded to compare the gene tree scores. Bootstraps consisted of 10,000 gene trees resamplings prior to the species tree reconstruction and scoring. The test's parameters were approximated using their maximum likelihoods estimates. Again, topologies were compared to the best-scoring topology obtained for nucleotide or amino-acid datasets at a 5% threshold.

Tree reconstruction with constrained gene trees It is known that coalescent methods, despite providing a more biologically realistic model, are sensitive to gene tree estimation error (GTEE) (Roch and Warnow, 2015; Nute et al., 2018). Therefore, we tried to estimate the species tree using constrained gene trees. This constraint aims to reduce the error in gene tree estimation. Indeed, incomplete lineage sorting (ILS) or species isolation with gene flow are less likely to occur between ancestral nodes separated by longer branch lengths as measured in coalescence units, as is the case for species of the genus *Paralvinella* (excluding *P. p. irlandei* and *P. unidentata*) and outgroup species (see Supplementary data Figure 3 and 4). Fixing species arrangement in these groups should therefore be closer to the true tree compared to free-running arrangements that can be biaised by GTEE, and improve the confidence on the probabilities for the nodes evaluated in the 15 topologies.

For each gene, the 15 candidate topologies were evaluated under ML using the best evolutionary model according to ModelFinder. Topology k was then given a Bayesian posterior probability from its likelihood L_k using the BIC approximation $P_k = e^{L_k} / \sum_{i=1}^{15} e^{L_i}$. The species tree scores were obtained with ASTRAL, with confidence intervals obtained by 10,000 bootstraps on each gene tree topology according to their Bayesian posterior probability, to take into account the confidence of gene tree topologies.

This procedure was conducted on both nucleotide-encoded genes (657) and amino-acid encoded genes (699). To summarize the information coming from the two types of sequences, a global species tree score was calculated. Nucleotide normalized quartet scores, N, were combined with amino-acid normalized quartet scores, A, by considering the Euclidean distance d between pairs of sequence datasets (N, A) and an hypothetical best topology associated with the pair $(1, 1)$: $d = \sqrt{(1-N)^2 + (1-A)^2}$. This topology would correspond to an ideal tree that would maximize the quartet scores both for nucleotide and amino-acid encoded genes. Thus, the lower the distance estimated, the better the topology.

The distance obtained for each candidate topology was then tested against the best-scoring topology with a procedure equivalent to the AU test (Shimodaira, 2002), using 10,000 multi-scaled bootstraps on genes and the sampled tree topologies obtained for each gene (derived from the topologies' Bayesian posterior probability) at a 5% p-value threshold.

Finally, gene histories were addressed by looking at the individual gene tree's probability over the 15 constrained topologies. To be able to distinguish between gene tree estimation error (GTEE), ILS and interspecific gene flow, genes were split into four categories according to the number of parsimony-informative sites for the nodes at the root of the family Alvinellidae. In case of high GTEE, one would expect the distribution of probabilities for the best topology to be more randomly distributed between the 15 constrained topologies for less informative genes compared to highly informative genes.

Topologies' probabilities were also mapped on a ternary plot, allowing us to

evaluate the number of genes strongly associated with particular topologies ($Pr > 0.8$). In that case, the 15 topologies were reduced to only three candidates corresponding to different quartets at the root of the Alvinellidae family:
((*P. p. irlandei*,*P. unidentata*),(*Alvinella*,*Paralvinella*)),
((*P. p. irlandei*,*Alvinella*),(*P. unidentata*,*Paralvinella*)),
((*Alvinella*,*P. unidentata*),(*P. p. irlandei*,*Paralvinella*)). In the case of high interspecific gene flow between specific ancestors in the early steps of the Alvinellidae radiation, one would expect to see one quartet supported by a majority of genes (corresponding to the species tree), one quartet supported by a high number of genes (corresponding to intraspecific gene flow) and a quartet supported by a lower number of genes (due to potential ILS).

Molecular dating of the family's radiation The Alvinellid phylogeny was dated with the software Phylobayes v.4.1. (Lartillot and Philippe, 2004; Lartillot et al., 2007). 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 concatenated amino acid-encoded genes for the two most relevant species tree topologies. 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 by a Gamma distribution *prior* with a mean of 330 million years and a standard deviation of 200 million years. This places 95% of the age distribution between 300 and 360 million years ago. We also used two calibration points corresponding to the divergence between the species *P. palmiformis* and *P. grasslei*, estimated to be 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* sampled in the Manus and Lau basins, estimated to first expand between 2 and 4 Ma (Desbruyères et al., 2006; Boulart et al., 2022).

Results

Global phylogenetic inference The global phylogeny was performed under a partition gene model on both the nucleotidic and amino-acid supergenes, or using the coalescent method on our individual gene datasets, based on either nucleotide or amino-acid encoded sequence alignments. Both approaches identify three main clades inside terebellomorph species, which are the Alvinellidae family, the Ampharetidae family comprising *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 as the melinnid species (*M. palmata*), as shown in Figure 1. *M. palmata* is a sister species of other terebellidae species for all methods but the phylogeny obtained from the nucleotide-encoded supergene, where it is the sister species to the three polychaete families (Alvinellidae, Terebellidae and Ampharetidae excluding the outgroup *P. gouldii*). *Anobothrus* sp. also represents a sister species to other Ampharetidae species in phylogenies using the amino-acid encoded sequences (82% bootstrap for the concatenation approach or 0.6 posterior probability in the coalescent method), but is sister to other Ampharetidae and Alvinellidae species using nucleotide-encoded sequences with a greater robustness (100% bootstrap or 1.0 posterior probability). All phylogenies with their node supports are included in the supplementary data, Figure 1 to 4. In further analyses, however, we chose to only consider phylogenies that group *M. palmata* in the Terebellidae clade and separate Alvinellidae and Ampharetidae as two monophyletic families (Fig. 1), according to current views on the Terebelliformia phylogeny (Stiller et al., 2020).

Within the Alvinellidae, all methods give a maximum bootstrap value to the grouping of the two *Alvinella* species. In the *Paralvinella* genus, with the exception of *P. unidentata* and *P. p. irlandei*, the phylogeny of species is also resolved without ambiguity, with a maximum of confidence on all nodes regardless of the phylogenetic method used. Conversely, the placement of the species *P. unidentata*, and more importantly *P. p. irlandei*, is not well supported. Depending on the number of genes

considered, the evolutionary model used, or the types of sequences, the placement of these two species in relation to the well-defined groups of the other *Paralvinella* species, on one hand, and the two species of the genus *Alvinella,* on the other hand, is fluctuating at the root of the family Alvinellidae. Two high-scoring alternative trees are shown in Figure 1a. and 1b. For this reason, we chose to address more specifically the resolution of these deep nodes with a topology fixed for the outgroups and the other *Paralvinella* species, as they were well resolved with all methods used.

Evaluating bifurcations at the root of Alvinellidae In order to solve the alvinellid family topology, we exhaustively compared the phylogenetic results obtained for all possible root topologies. The outgroup subtree (Terebellidae+Ampharetidae rooted with *P. gouldii*) as well as the *Paralvinella* subtree (excluding *P. p. irlandei* and *P. unidentata*) are considered resolved and constrained. Figure 2 lists all these topologies, numbered from 1 to 15.

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

- Topology 7, 8 and 9, in which *P. p. irlandei* is sister to the all other Alvinellidae. Topology 9 brings the *Alvinella* and *P. unidentata* lineages closer together, *A. pompejana*, *A. caudata* and *P. unidentata* being sister species. Topology 7 brings the *Alvinella* lineage closer to *P. p. irlandei*, and topology 8 brings *Alvinella* species closer to other *Paralvinella*. According to the AU test, the best-ranking topology obtained in the different phylogenetic method (T9 or T8) was not significantly better than these topologies, except T7 which was significantly worse than T9 in the nucleotide+ML analysis (p-value=0.3%). Following all analyses, we finally retained T9 as the most appropriate topology over the three and this topology is detailed in Figure 1a;

- Topology 6, in which the *Alvinella* lineage is sister to *Paralvinella*, which forms a monophyletic genus within which *P. p. irlandei* and *P. unidentata* are sister species. According to the AU test, the best-ranking topology T9 (using the ML inference on nucleotide sequences), or the topologies T9 and T8 (using the coalescence approach with both nucleotide and amino acid-encoded genes) were not significantly better than T6. However, T6 was significantly worse than T9 in the amino acid+ML approach (p-value=2.1%). This T6 topology is detailed in Figure 1b.

Phylogenetic discordance between genes We then focused on the evaluation of the phylogenetic signal at the gene level. We estimated the likelihood of the 15 constrained topologies for each gene separately, either on their coding sequences or their translated protein counterparts. As shown in Figure 3b. and 3d., genes provided very different phylogenetic signals. Interestingly, genes containing more parsimony-informative sites, and thus have more power to distinguish between the contrained topologies (by increasing the phylogenetic signal at the root of the family), were not able to exclude some topologies, even though the topologies T6 and T9 are supported by a relatively higher proportion of genes as the number of informative sites increases. As a consequence, the diversity of gene histories is apparently not correlated with a loss of phylogenetic signal at some slow-evolving genes in which a higher gene tree estimation error is prone.

In Figure 3a. and 3c., we reassigned the 15 topologies to three simpler scenarios labelled 1, 2 and 3, to identify the shared gene histories between the *Alvinella*, *Paralvinella*, *P. unidentata* and *P. p. irlandei* lineages. Considering nucleotide-encoded genes, 182 genes favor scenario 1 with a closer relationship between *Alvinella*+*P. p. irlandei* opposed to *P. unidentata* +*Paralvinella* (55 genes assigned with $Pr > 0.8$), 234 genes favor scenario 2 in which *Alvinella*+*P. unidentata* is opposed to *Paralvinella*+*P. p. irlandei* (91 with $Pr > 0.8$) and 241 genes are supporting scenario 3 where *Alvinella*+*Paralvinella* is opposed to *P. p. irlandei* +*P. unidentata* (94 with $Pr > 0.8$). For amino-acid encoded

	Concatenation		Coalescence	
	499,036 sites nucleotide	277,900 sites amino acid	657 genes nucleotide	699 genes amino acid
T1	- 419	- 582	- 2,743	- 16,170
T2	- 600	- 619	- 13,197	- 14,886
T3	- 343	- 410	- 6,822	- 6,420
T4	- 312	- 428	- 2,915	- 6,627
T5	- 271	- 480	- 2,304	- 16,085
T6	- 119	- 208	- 665	- 9,202
T7	- 167	- 88	- 2,234	- 1,019
T8	- 77	- 93	- 6,330	0
T9	0	0	0	- 802
T10	- 482	- 632	- 10,956	- 19,554
T11	- 380	- 521	- 10,896	- 13,790
T12	- 548	- 645	- 22,097	- 27,472
T13	- 443	- 408	- 9,572	- 16,903
T14	- 421	- 492	- 8,951	- 17,059
T15	- 644	- 703	- 14,128	- 21,391

Table 1: Scores obtained for the 15 Alvinellidae topologies under different phylogenetic models. First and second columns: maximum likelihood on the nucleotide or amino-acid encoded supergenes, run as one gene partition under the GTR+Γ or LG+Γ model. Third and fourth columns: quartet score of species trees established from gene trees, either from nucleotide or amino-acid encoded genes. For each method, the topology with the highest score is set to 0, and the difference between the tree score and the best-scoring topology is shown. According to the AU test, species trees for which the top ranked topology is not better at a 5% threshold are in grey.

genes, these quartets are respectively supported by 205 (55), 234 (72) and 260 (94) genes. Compared to a random distribution of genes between the three scenarios, as would be the case under strict incomplete lineage sorting, we observed that genes are globally less supportive the hypotheses of scenario 1 (nucleotide: $X \sim B(657, \frac{1}{3})$, $Pr(X \leqslant 182) = 0.001$; amino acid: $X \sim B(699, \frac{1}{3})$, $Pr(X \leqslant 205) = 0.013$). Genes are equally supportive of scenarios 2 and 3. Indeed, the distribution of genes strongly associated with one pole is the most biased between scenarios 2 and 3 for amino-acid encoded genes. In this case, 72 genes are associated with scenario 2 against 93 for scenario 3. Thus under the null hypothesis of an equal distribution between the two poles, $X \sim B(165, \frac{1}{2})$, and $Pr(X \leqslant 72) = 0.06$. The gene distribution indicated that a greater number of genes supports a closer relationship between *P. unidentata* and *P. p. irlandei*, but this was not significant at the 5% threshold.

Species tree reconstruction with constrained gene trees and the coalescent approach In our final approach, we attempted to reconstruct the species tree using the coalescent method from gene trees constrained to one of the 15 topologies. The goal was to reduce the overall gene tree estimation error (GTEE), assuming that incomplete lineage sorting (ILS) and inter-species gene flow were low at all nodes of the tree with the exception of the two nodes at the root of the family Alvinellidae. When evaluating the 15 topologies with this approach (see: Figure 4a), T7 had on average the highest quartet scores both for nucleotide-encoded genes and their translated amino-acid counterparts, although this result was sensitive to the resampling of genes and gene trees uncertainty. Figure 4b. displays the combined quartet scores obtained by each topology as the Euclidean distance to a hypothetical ideal topology at coordinates (1,1) that would maximize the quartet scores for both types of sequences. In this representation, T7 also obtains the best combined score, although not statistically better than T8 and T9 ($p - value = 0.23, 0.34$ respectively). In all these topologies, *P. p. irlandei* is paraphyletic with the other *Paralvinella* species. The relationship of *P. unidentata* with other alvinellid clades is however different in these three topologies, being sister species to *Paralvinella* in T7, to

Alvinella in T9 or to *Alvinella*+*Paralvinella* in T8. Compared to previous analyses for which T6 and T9 were preferred, these topologies are less supported here, even if the score of T7 is only significantly better than T6 ($p - value = 0.009$) but not T9. This may be explained by the fact that T7 represents an intermediate topology between T6 and T9, which are two most frequent co-occurring topologies in the gene trees sets (see Discussion, Monophyly of the genus *Paralvinella*).

Molecular dating of the alvinellid radiation We estimated the age of the alvinellid worm's radiation under the three main topology hypotheses. The three models used for molecular dating (Log-normal, CIR, and uncorrected Gamma multipliers) gave different results about the age of the radiation, ranging from 45 to 100 millions of years ago (Ma). However, the log-normal model appears to significantly underestimate ages, since the root of the tree is estimated to be between 79 and 89 millions of years old (My) depending on the tree topology, which is far from our expectations based on the age of the terebellomorph fossils from the Devonian-Carboniferous period (Sepkoski Jr., 2002). On the other hand, the uncorrelated Gamma multipliers model, which assumes that the evolutionary rates of branches are uncorrelated with time, places the age of the tree's root between 291 and 335 My (confidence intervals: 132-599 Ma), which is very close to the expected age. The radiation of the Alvinellidae is estimated between 92 and 99 Ma (47-180 Ma). Yet the 95% confidence intervals are still very large with this model, and the estimated ages are far from the ages with the model run under the *prior*, which dates the MCRA around 70-80 Ma (between 36 and 214 Ma). Finally, the CIR model proposed the most consistent estimates (see: Figure 5). The age of the terebelliform MCRA is 120-160 My (98-198 My). The age of the Alvinellidae radiation stands between 67 and 91 Ma (55-112 Ma). These ages also better agree with the model run under the *prior* (age of the root ranged between 34 and 210 My, with a mean between 75 and 81 Ma depending on the topology). In these two last models, the radiation of the Alvinellidae was a rapid event, probably occurring in the Late Cretaceous. The estimates obtained with the CIR model for

alternative topologies are given in the supplementary data, Figure 5.

Discussion

Monophyly of the genus Paralvinella

The initial description of the family by Desbruyères and Laubier subdivided the family into two genera, mainly based on the shape of the gills (four pairs of lamellar gills in *Alvinella*, *versus* simple cylindrical gills with no filaments at the tip in *Paralvinella*), the position and shape of the 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*) (Desbruyères and Laubier, 1986; Jouin and Gaill, 1990; Jollivet and Hourdez, 2021). Based on these morphological criteria, these authors therefore proposed a taxonomic arrangement in which *P. unidentata* was closer to the sibling pair of species *P. p. irlandei* and *P. p. pandorae* within the *Paralvinella* genus, and subsequently grouped into the Nautalvinella subgenus (Desbruyères and Laubier, 1993). This grouping proposal is equivalent to the T6 topology, which is one of the best topologies identified by our molecular phylogeny on supergenes, and which is also well supported by individual gene trees when constrained among the 15 most relevant tree topologies. The dichotomy between the genera *Alvinella* and *Paralvinella* was also suggested by Jollivet and colleagues who proposed the monophyly 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 of the species *P. unidentata*, which are mandatory to resolve the initial quartet at the root of the family clade.

However, although the alvinellid family appears to be monophyletic, the split of the first alvinellid ancestors is poorly resolved at the root of this clade. Other topologies, notably T7 and T9 are also well supported by numerous genes suggesting that the genus

Paralvinella could be paraphyletic. Topologies T7, T8 and T9, within which *P. p. irlandei* is a sister species to the other Alvinellidae, got very good scores and agree with previous studies proposing phylogenies based on either mitochondrial Cox1 sequences or sets of orthologous sequences derived from transcriptomes and combined with some morphological characters (Vrijenhoek, 2013; Stiller et al., 2020; Jollivet and Hourdez, 2021). In our analysis, the T9 topology - but also the T7 topology to a lesser extent - is well supported by the gene-to-gene approach (using a coalescent method similar to the one employed by Stiller et al. (2020)) or partitioned phylogeny. In the present study, we have taken care to ensure that each orthologous gene cluster contained a sequence for *P. p. irlandei*, but additional information from its sister species *P. p. pandorae* from the Juan de Fuca Ridge would have certainly help to improve the confidence at the first nodes of the family. These three topologies differ in the positioning of *P. unidentata* in relation to the rest of the Alvinellidae. In T7, *P. unidentata* is closer to other *Paralvinella* species, while it is closer to *P. p. irlandei* in T8, and closer to *Alvinella* in T9. It is worth noting that the grouping of *P. unidentata* and *Alvinella* (T9) is in agreement with some traits concerning the gills' morphology: in particular, the filaments of *P. unidentata* are flattened, which is unique compared to other *Paralvinella* species, and similar to the lamellar shape of the gills of the two *Alvinella* species. This topology thus raises the question of the monophyly of *Alvinella* in the traditional view that separates this genus from all other species of Alvinellidae. However, the overall structure of the gills of *P. unidentata*, 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). The study of the gills in Alvinellidae is a good example of the importance of morphological variations that can be observed within the Terebelliformia at the level of palps, chaetae and gills, as well as the shape and position of the uncini (Day, 1964; Stiller et al., 2020), which can be highly diverged even in closely-related species because of difference in their adaptive trajectories. To this extent, the shape of gills and their size are likely to be sensitive to natural selection

as they are crucial to enduring long periods of hypoxic conditions, while being able to rapidly capture massive amounts of oxygen in well-oxygenated water when the worm is no longer exposed to the fluid mixing(Jouin and Gaill, 1990). Depending on the species' ecology, these periods of oxygenation may greatly vary from very brief periods of times (*Alvinella*) to longer ones (*P. p. irlandei*).

From our results, the paraphyly of *Paralvinella* seems to be the preferred solution. It is noteworthy that this conclusion is more supported by amino-acid encoded genes when compared to nucleotide encoded genes. This could be partially due to long-branch attraction artifacts, given that the average amino acid composition of *P. p. irlandei* sequences deeply contrasts with the average composition of other alvinellid species (see Methodological control - Data quality). This bias in the amino-acid composition of *P. p. irlandei* is tricky but could be explained by the ecology of the worm, which lives under colder conditions and seems to be less adapted to hypoxia (i.e. comb-like gills). Nevertheless, we chose to retain T9, as shown in Figure 5, as the most reliable species tree topology for the Alvinellidae. This contrasts with the initial view of a family split into two genera with *Alvinella* species sister to *Paralvinella* species. However, most gene trees are inconsistent with this topology, even for the most phylogenetically-informative genes. The radiation of the Alvinellidae 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 in the timescale of the whole family, and possibly linked to some adaptive strategies to cope with different thermal habitats. This led to high incomplete lineage sorting and cross-species gene flow, possibly adaptive, which appears with branches shorter than 0.1 coalescent time units and a mixing of genes pointing to either the T6 or the T9 topologies. This also agrees well with the observation that the evolutionary rate may be slow in alvinellid worms, consequently maintaining ancestral polymorphisms for a very long time (Fontanillas et al., 2017). For example, in *P. unidentata* and *P. fijiensis*,

present-day populations that have been separated for 3 to 4 million years also display short coalescent times, between 0.05 and 0.35 (see supplementary data Fig. 1), and Jang et al. estimated the split of the the northern East Pacific Rise and the northeastern Pacific Antarctic Ridge populations to about 4.2 million years ago for *A. pompejana* (Jang et al., 2016), despite the maintenance of shared polymorphisms between the two metapopulations.

Finally, the distribution of gene tree topologies is not random, but biased toward two alternative scenarios where *P. unidentata* is either closer to the *Alvinella* species, or closer to *P. p. irlandei*. Such a distribution is expected if one of these topologies is reflecting the true species tree, while the other is the consequence of specific gene flow between the two incriminated lineages. Yet these two scenarios are not directly compatible with one another if we consider than T9 is the true species tree. On the contrary, T7 is an intermediate topology where asymmetric high gene flow from the ancestor to the *P. unidentata* lineage to the ancestor of *P. p. irlandei* would result in T6, while cross-species gene flow between ancestors of *P. unidentata* and the *Alvinella* lineage would result in T9. Thus, even though it is less well supported by the phylogenetic inferences from concatenated supergenes or unconstrained gene trees, T7 is an interesting topology that could reconcile the main phylogenetic relationships that appear in the multi-gene analysis. Lastly, while T8 is closer to the T9 topology, it is hardly compatible with T6 as the transition from T8 to T6 requires a higher number of tree rearrangements. This make T8 less likely to represent the true species tree.

Consequently, in the case of the Alvinellidae, the rapid radiation of the first ancestral lineages led to some probable introgressive gene flow between the ancestor of *P. unidentata* and both the ancestors of *P. p. irlandei* and the two *Alvinella* species which may have been coupled with the maintenance of ancestral polymorphisms due to a very high level of ILS. Following this view, the topologies T6, T7, and T9 appear to be equally likely to explain the evolutionary history of this peculiar and highly adapted family of worms. This casts doubt on the relevance of a binary tree of species at the root of the

family, as different genomic regions more likely have different phylogenetic relationships.

Division of the genus Paralvinella

Desbruyères and Laubier proposed to divide the genus *Paralvinella* 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 the species *P. hessleri*, *P. dela* and *P. bactericola*;
- *P. Nautalvinella*, which includes the species *P. unidentata*, *P. p. irlandei* and *P. p. pandorae*.

Several criteria were 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 were used to differentiate these subgenera despite being more variable, 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).

Han et al. also argued that the new species *P. mira* is related to the subgenus *Miralvinella* by considering the pointed shape of the sexual tentacles, the shape of gills and the presence of digitiform lobes on the notopodia. The authors still noted peculiarities, such as the first three setigerous segments, which are not fused in *P. mira*, and the

insertion of numerous slender oral tentacles on the buccal apparatus (Han et al., 2021). In our analyses, the phylogeny of the species which belong to the subgenera *Miralvinella* and *Paralvinella* is well resolved with a very high level of confidence. The species *P. mira*, as suggested by (Han et al., 2021), effectively groups with *P. hessleri*, which is the only other *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* when considering the shape of the buccal apparatus (stronger tentacles in *P. mira* and *P. hessleri*) and the position of the first uncinigerous neuropodial tori on chaetiger 16 or 18 in *P. mira* and *P. hessleri vs.* 32 for *P. dela* and *P. bactericola* (Han et al., 2021; Jollivet and Hourdez, 2021). The geographical distribution of the species also agrees with this view, as *P. mira* and *P. hessleri* are found in the Indian Ocean and south-west Pacific Ocean, while *P. dela* and *P. bactericola* are sister species inhabiting the Juan de Fuca Ridge and the Guaymas basin in the eastern Pacific Ocean (Desbruyères and Laubier, 1993; Jollivet and Hourdez, 2021). Unfortunately, molecular data are not available for the rarer species *P. dela* and *P. bactericola* to confirm the monophyly of the *Miralvinella* subgenus. The subgenus *Paralvinella* 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 a symplesiomorphy of the genus *Paralvinella*, derived later as pointed tentacles in *Miralvinella*.

Finally, the subgenus *Nautalvinella* as described by Desbruyères & Laubier is also no longer monophyletic. If T9 or T7 are kept as the most probable species trees, the species *P. p. irlandei* and *P. p. pandorae* are sister to the *Alvinella* and *Paralvinella* clades, and can be grouped under a reduced *Nautalvinella* genus as proposed by Stiller et al. (2020). It is noteworthy that in T9, *P. unidentata* is also a sister species to the two *Alvinella* species. This casts doubts on the definition of the *Alvinella* genus, which groups together two species with very little divergence ($< 1\%$) but not *P. unidentata* from which they greatly diverge. Yet single gene phylogenies are equally distributed between topologies that group

Alvinella and *P. unidentata* (mainly T9) and those which group *P. p. irlandei* and *P. unidentata* together (mainly T6, see Fig. 3). Thus, the acquisition of comb-like gills, the loss of the pair of sexual tentacles on the buccal apparatus of males and the loss of the digitiform notopodial lobes, which are the main elements in Desbruyères & Laubier's diagnosis of the subgenus *Nautalvinella* (Desbruyères and Laubier, 1993), may have arisen as a result of gene transfer between the *P. unidentata* and *P. p. irlandei* lineages.

The spread of alvinellid worms in and outside of the Pacific Ocean

Our molecular dating estimates rely on two calibration points. The first point corresponds to the recent opening of the Manus (3-4 Ma) and Lau (1-2 Ma) basins, which are assumed to form two disjoint ridge systems. The second refers to the subduction of the Farallon Plate beneath the North American Plate, which caused a vicariant event between hydrothermal vent communities of the Eastern Pacific between 34 and 23 My (Tunnicliffe, 1988). This vicariant event has already been used to date other phylogenies of vent species distributed between the EPR and JdF, if we consider that these species disperse mostly along the axial valley of oceanic ridges (Stiller et al., 2013; Jollivet and Hourdez, 2021). This last calibration point is particularly interesting in the case of the Alvinellidae, since it separates several pairs of alvinellid 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). Our estimates would certainly become more accurate through a more comprehensive sampling of the different species of the family, as 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 (Jollivet and Hourdez, 2021). Based on this sequence, the level of divergence estimated between *P. p. irlandei* and *P. p. pandorae* was nearly identical to the divergence estimated by Chevaldonné et al. for the species *P. palmiformis* and *P. grasslei*, and supports the *Cox1* divergence rate of 0.13%/My established by these

authors for the Alvinellidae (Chevaldonné et al., 2002). Following this rate, the estimates for dating the age of alvinellid radiation using our genomic datasets were not very different from the mitochondrial estimate with the same set of species (Jollivet and Hourdez, 2021).

Regardless of the topology chosen (T6, T7 or T9), most of the species in the family descending from the first ancestors (*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. This is consistent with the distribution of the sister family Ampharetidae, for which species diversity was the greatest in the eastern Pacific, which may suggest a common origin for both families in the now-extinct or subducted Pacific-Farallon and Mathematician ridges. These ridges have been active since the Early Jurassic (200 Ma) and have played a central role in the initial dispersal of the hydrothermal vent fauna in the Pacific (Mammerickx et al., 1980; Bachraty et al., 2009). The split between the Alvinellidae and Ampharetidae occurred between 100 and 130 Ma according to the CIR model, 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 ages (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 likely that the ancestor of the family emerged in the Pacific, possibly associated with vent sites.

Depending on the model used for the molecular dating, the subsequent radiation of the family Alvinellidae occurred between 92-99 Ma (uncorrelated gamma model, 47-180 Ma) or between 67-91 millions years ago (CIR model, 95% confidence interval 55-112 Ma) . These estimates suggest that alvinellid worms have a long history of speciation, dating back to the Cretaceous period (Haymon et al., 1984; Vrijenhoek, 2013). Using 278

orthologous genes or the *Cox1* mitochondrial gene, Jollivet and Hourdez proposed two dates for the radiation of the Alvinellidae, one at 198 Ma and the other at 72 Ma (Jollivet and Hourdez, 2021). The first dating, obtained from only one calibration point, is likely too old as 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) (Vrijenhoek, 2013). Moreover, the molecular dating did not take into account the upper age limit associated with the first occurrence of Terebelliformia fossils. Our estimates, under the hypotheses of the three tested topologies (T6, T7 and T9), are in line with a radiation dating back to 70-80 Ma. This is in agreement with the second dating of Jollivet and Hourdez at 72 Ma, established from *Cox1* with one calibration point for two pairs of sibling species. The radiation is therefore likely to follow the Cenomania/Turonian extinction phase, about 94 Ma. This major anoxic/dysoxic event has been argued to have created opportunities for modern taxa to invade deep environments (Thomas, 2007; Vrijenhoek, 2013).

Finally, considering the alternation 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. unidentata*, *P. hessleri*, *P.* sp. nov., *P. fijiensis*), the colonization of the Pacific Ocean must have occured several times (Desbruyères and Laubier, 1993). The divergence of the *Miralvinella* species would range between 35 and 65 My (95% confidence interval: 27-80 according to the CIR model), which corresponds to the spread and subduction of the Kula ridge (between 60 and 43 My, Smith (2003)). This ridge has been indeed proposed to represent a bridge between the western and eastern Pacific hydrothermal basins, which are now isolated from each other (Hessler and Lonsdale, 1991; Bachraty et al., 2009). The colonization of the Southwest Pacific region would be a second event between 45 and 22 My (55-18), potentially concurrent with the subduction of the Pacific plate under the Ontong-Java plateau

between 45 and 30 Ma and the subsequent opening of modern-day basins (Schellart et al., 2006). The recent description of *P. mira* in the Indian Ocean, at Wocan and Daxi vents, shows that the Alvinellid worms have also spread outside the Pacific Ocean. The divergence between *P. mira* and *P. hessleri* is dated back to 35-37 My (27-48) and could be an argument for the recent colonization of the Indian Ocean . 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 Philippines and the opening of the China Sea (Parker and Gealey, 1985; Moalic et al., 2012). Although the tectonic history of this region make it difficult to reconstruct ridge connectivity between oceanic basins, the colonization of the Indian Ocean by Alvinellidae was made possible through the Antarctic Ridge, whose spread started about 50 Ma and connects the Indian Ocean with the Pacific Ocean (Parker and Gealey, 1985). This ridge is indeed a link to most of the vent sites, from the EPR to the mid-Atlantic Ridge through the Indian Ocean (Moalic et al., 2012). Although Alvinellid worms are rare in the Indian Ocean, an undescribed species of *Paralvinella* was found at the Solitaire vent 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.

Methodological control of the phylogenetic reconstructions

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. Species associated with the subgenera *Paralvinella* and *Miralvinella* (without both *P. unidentata* and *P. p. irlandei*) were thus constrained, taking into consideration that this group was correctly resolved with reliable and consistent positioning of all species regardless of the method used (concatenation or coalescent-based approaches, on either 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 consistent with the phylogeny proposed by Stiller et al. and clusters the Melinnidae and the Terebellidae together with strong bootstrap support (Stiller et al., 2020). The species positioning in the family Ampharetidae is more questionable. *Anobothrus* sp. was indeed sister to the family Ampharetidae from analyses based on amino-acid encoded sequences (bootstrap=82% and posterior probability = 0.6), but the nucleotide sequence data suggested that *Anobothrus* sp. could be sister to both Alvinellidae and Ampharetidae. Here we chose to consider only the topology in which Ampharetidae species form a monophyletic group to reduce the variance of our tree likelihood estimates with another flexible species position, but this constraint is likely a simplification of the true gene topologies.

The question of whether the Alvinellidae form a sister family to the Ampharetidae or whether the family should be included within the Ampharetidae cannot be fully resolved without a more exhaustive sampling of species within the latter family. Both families have retractable buccal 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 a clear separation of the two families (Rousset et al., 2003; Stiller et al., 2020), Eilertsen et al. suggested on a broader ampharetid phylogeny based on classical phylogenetic markers (mtCOI, mitochondrial 16S rDNA, nuclear 18S and 28S rDNA) that Alvinellidae should be included within Ampharetidae (Eilertsen et al., 2017). The relationship between the two families is presented as a trifurcation between the Alvinellidae, *Amphysamytha* and [*Anobothrus*+*Amphicteis*] lineages. In contrast to this last result, our study concludes that the *Amphisamytha* species (*A. carldarei*) is sister to the *Amphicteis* species (*A. gunneri*). The unresolved trifurcation proposed by Eilertsen et al. is resolved in our analysis, with Alvinellidae being a monophyletic family, sister to the clade and *(Anobothrus*,(*Amphicteis*,*Amphysamytha*). As the question of polyphyly of the

Ampharetidae group was not the purpose of the present study, we therefore decided to follow the more consensual statement considering Ampharetidae and Alvinellidae as two separate monophyletic groups, in accordance with phylogenies including a broader range of Terebelliformia species (Stiller et al., 2020). Nevertheless, the monophyly of Ampharetidae with respect to Alvinellidae remains under question.

Data quality Phylogenetic reconstruction is sensitive to differences in species sequence composition. Indeed, compositional biases affecting divergent species in the same way may be explained by evolutionary convergence events that can be confused with synapomorphies. Conversely, the absence of a 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 contrasting thermal environments, which is known to have an effect on the amino-acid composition of proteins between mesophilic and thermophilic eukaryotes (Wang and Lercher, 2010; van Noort et al., 2013; Bock et al., 2014).

The evaluation of the amino-acid composition bias in the alvinellid orthologous gene clusters shows strong disparities between species (see supplementary data Fig. 6). In particular, *P. p. irlandei* shows very different values of amino-acid composition when compared to other species, with great differences in indices such as CvP bias, Serine bias and purine load: criteria which have been suggested to discriminate between warm and cold species (Hickey and Singer, 2004; Fontanillas et al., 2017). Therefore, we carefully filtered out genes that exhibited a high compositional heterogeneity in our phylogenetic analyses. We also ensured each gene alignment contained a sequence of *P. p. irlandei* in the analysis, although most of the *P. p. irlandei* sequences were on average shorter and more fragmented (158,783 total nucleotide sites and 292,610 amino acid sites). This limits the available phylogenetic information for this species compared to other alvinellid species

of the dataset, and may have played a role in the placement of this species in the alvinellid tree. One might indeed expect that the strong compositional bias found in *P. p. irlandei* would tend to make this species diverge more rapidly among the Alvinellidae in the phylogenetic reconstruction. Moreover, the average composition of *P. p. irlandei* is not only different from the other species of Alvinellidae, but also strangely differs from the species chosen as outgroup (Ampharetidae + Terebellidae + *P. gouldii*). The positioning of *P. p. irlandei* in the alvinellid tree could be therefore highly gene-dependent when estimating tree topologies (and to a lesser extent when fitting constrained gene trees). One must therefore be cautious with the good scores of the T7, T8 and T9 topologies, which place *P. p. irlandei* sister to other Alvinellidae species, as it may be a typical consequence of long-branch attraction.

The rapid radiation of the family probably occurred at the end of the Cretaceous period over a few million years, as shown by the short internal branches found at the root of the family. This is another difficulty for resolving tree bifurcations in this phylogeny. Rapid divergence can lead to a high rate of incomplete lineage sorting (ILS) and result in larger gene tree estimation errors (GTEE) (Molloy and Warnow, 2018). We must therefore be especially vigilant of 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 based on ODB9 (see Material & Methods), and paralogy should therefore not be a concern in our analysis. It has been suggested that in the case of very high ILS, coalescent-based methods give better results than 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 introgression or ILS. However, the superiority of coalescent methods depends on the confidence level put in gene tree estimation. For low or medium ILS or inter-species gene

flow, the concatenation method, which benefits from longer alignments, allows the reduction of GTEE, and is still more reliably in practice (Roch and Warnow, 2015; Molloy and Warnow, 2018). Generally speaking, in the case of high ILS or large amounts of inter-species gene flows, coalescent methods are preferable provided that GTEE remains low, particularly with the longest possible gene alignments (Nute et al., 2018). A simulation by Molloy and Warnow on 26 species and 1000 genes - conditions relatively similar to our study - further shows that if the ILS is high, increasing the number of genes is preferable over filtering out gene trees on the basis of GTEE. Nevertheless large GTEE severely limits the ability of both concatenation or coalescent methods to find true species trees (Molloy and Warnow, 2018).

To distinguish between GTEE and true gene tree variability at the root of the Alvinellidae family, we considered a second step in which ILS and gene flow were negligible at all nodes of the tree. Such an assumption was globally very well supported by all methods, except for the basal nodes of the Alvinellidae clade. We recomputed gene trees by reducing the topology space to trees T1 to T15, with the objective of limiting GTEE at these nodes by imposing a strong *prior* on these most probable topologies. 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 Fig. 3). If choosing between gene topologies was guided only by GTEE, one would expect to see certain topologies to prevail over others as the phylogenetic signal strengthens. This was not the case here, suggesting that other processes such as ILS and inter-species gene exchanges are causing the discordance between gene trees. Furthermore, unlike ILS which randomly segregates alleles between lineages at a given node, gene introgression occurs specifically between two lineages in some specific regions of their genome. Therefore, in the case of high gene flow from one lineage to another, one would expect to see some secondary topologies emerging to become more represented than others. In our case, *P. unidentata* was more frequently closely associated with *P. p. irlandei* and *Alvinella* than with other *Paralvinella* species,

suggesting that secondary gene exchanges occurred after the first speciation events with at least one of these two sister lineages.

After assessing the effect of GTEE on gene topologies, we also evaluated the species tree topology from the 15 constrained gene trees. This approach aimed at reducing the GTEE, which is an essential assumption of the coalescent-based methods. In this case, topologies T7, T8, and T9 were the three preferred solutions with very similar scores. These three topologies also obtained good scores from the phylogenies on supergenes (Table 1). In general, T9 was the most supported topology in all approaches. Yet, T7 draws our attention as it is the only species tree topology that can easily explain the high prevalence of both T6 and T9 at the gene level, with subsequent gene transfers between the *P. unidentata*, *P. p. irlandei* and *Alvinella* branches. 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. In the future, mapping the genes carrying differences in the phylogenetic signal onto the genome of one of these three species (and potentially enhancing the sequencing to non coding regions to improve genetic resolution) could enable us to see whether the genes associated with some specific topologies belong to specific regions of the genome. This would help to conclude whether our gene topology discrepancies were the result of ancestral gene transfers between these lineages and better quantify their relative impacts.

Conclusion

To conclude, our phylogenetic analysis of the alvinellid worms led to some uncertainty in choosing the right tree species topology between T6 (monophyly of the *Paralvinella* genus) and T9 (polyphyly of the *Paralvinella* genus with *P. p. irlandei* being sister to other alvinellid species). Different genes can have different phylogenetic histories, which is probably a consequence of the rapid radiation of several lineages from an ancestral worm population, resulting in high incomplete lineage sorting and intraspecific gene

introgression. Moreover, these two topologies can be equally supported by some morphological traits depending on whether some of them represent synapomorphies or symplesiomorphies. Some traits such as the gill shape or the amino-acid residue biases in proteins are likely to be the result of natural selection in the face of the contrasting thermal regime encountered by the worms.

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.

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.

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

Chernomor, O., Von Haeseler, A., and Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, 65(6):997–1008.

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*, 43(3-4):367–370.

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.

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.

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.

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.

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.

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.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, 35(2):518–522.

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.

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

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., 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.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

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.

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., Dang, C. C., and Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution*, 29(10):2921–2936.

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.

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.

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.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

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.

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., and Lanfear, R. (2019). The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biology and Evolution*, 11(12):3341–3352.

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.

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.

Schellart, W., Lister, G., and Toy, V. (2006). A Late Cretaceous and Cenozoic reconstruction of the Southwest Pacific region: Tectonics controlled by subduction and slab rollback processes. *Earth-Science Reviews*, 76(3-4):191–233.

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

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology*, 51(3):492–508.

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

Smith, A. D. (2003). A Reappraisal of Stress Field and Convective Roll Models for the Origin and Distribution of Cretaceous to Recent Intraplate Volcanism in the Pacific Basin. *International Geology Review*, 45(4):287–302.

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.

Thomas, E. (2007). Cenozoic mass extinctions in the deep sea: What perturbs the largest habitat on Earth? In *Large Ecosystem Perturbations: Causes and Consequences*. Geological Society of America.

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.

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.

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

Yang, Z. (1994). Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *Journal of Molecular Evolution*, 39(3):306–314.

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.

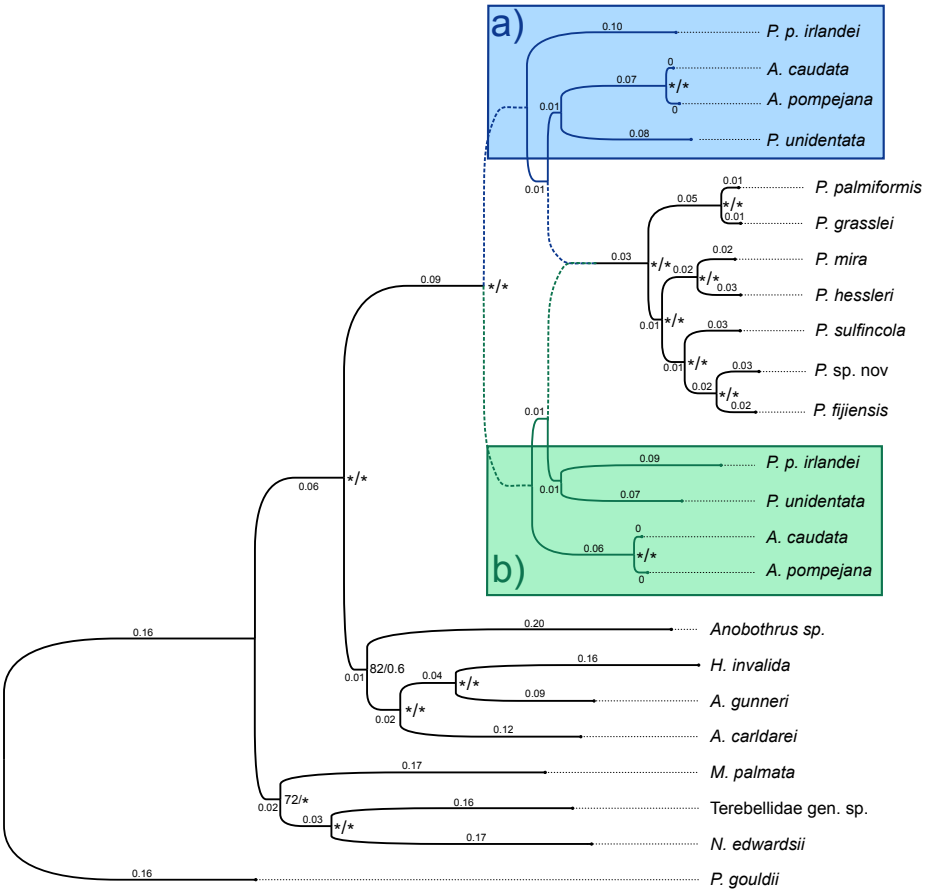

Fig. 1: High-scoring alternative topologies according to different phylogenetic inference methods for amino-acid encoded genes. (a) Main topology, with *P. p. irlandei* basal to the family Alvinellidae and clustering of *P. unidentata* with the *Alvinella* species (b) Alternative topology with clustering of all *Paralvinella* species together. Branch lengths are optimized under a partitioned model from concatenated amino-acid encoded genes. Node supports are given in the following order: Ultrafast bootstrap (partitioned model)/ Posterior probability (coalescent model). (*) indicates a bootstrap of 100% or posterior probability of 1.

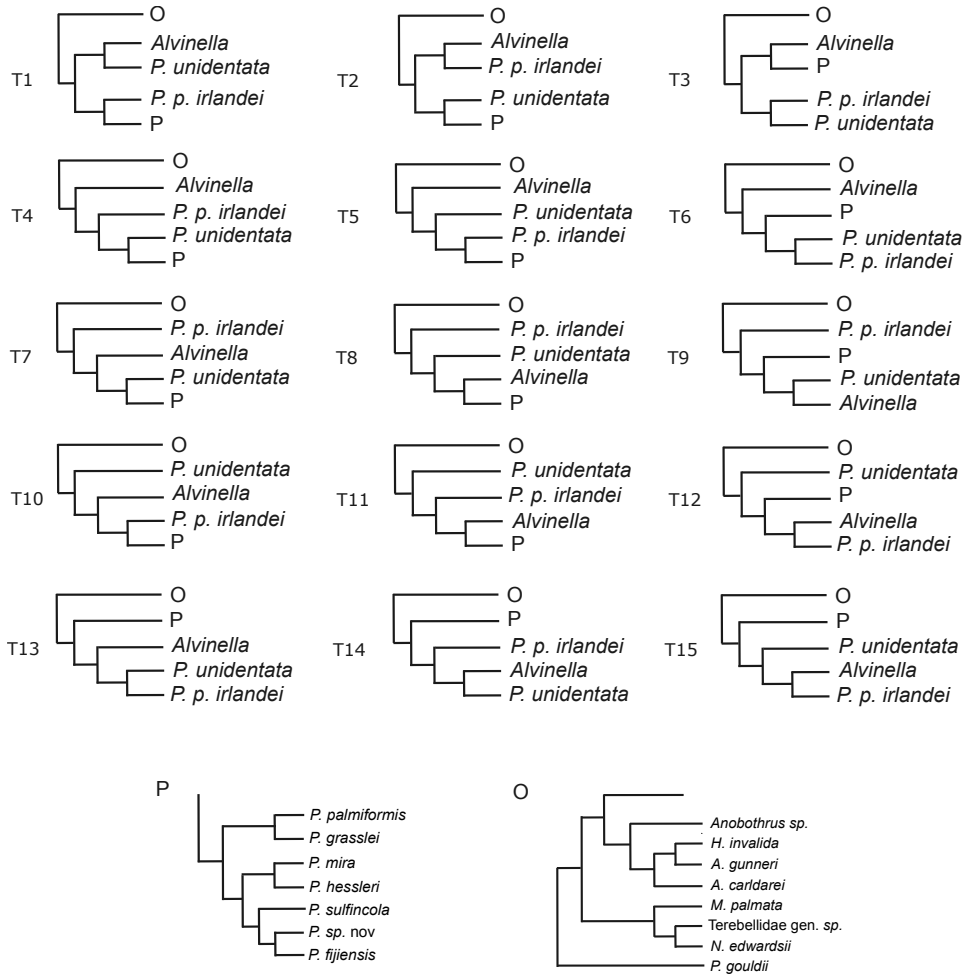

Fig. 2: Constrained topologies. The topologies only differ in the arrangement of the species *P. unidentata*, *P. p. irlandei*, when compared with the genera *Paralvinella* and *Alvinella* at the root of Alvinellidae. The *Alvinella* group comprises the species *A. pompejana* and *A. caudata*, and the *Paralvinella* group (P) comprises all species *Paralvinella* except *P. p. irlandei* and *P. unidentata*. Trees are rooted with a set of outgroups (O) with the same species arrangement comprising species from other closely-related families of Terebelliformia (Terebellidae+Ampharetidae+*P. gouldii*).

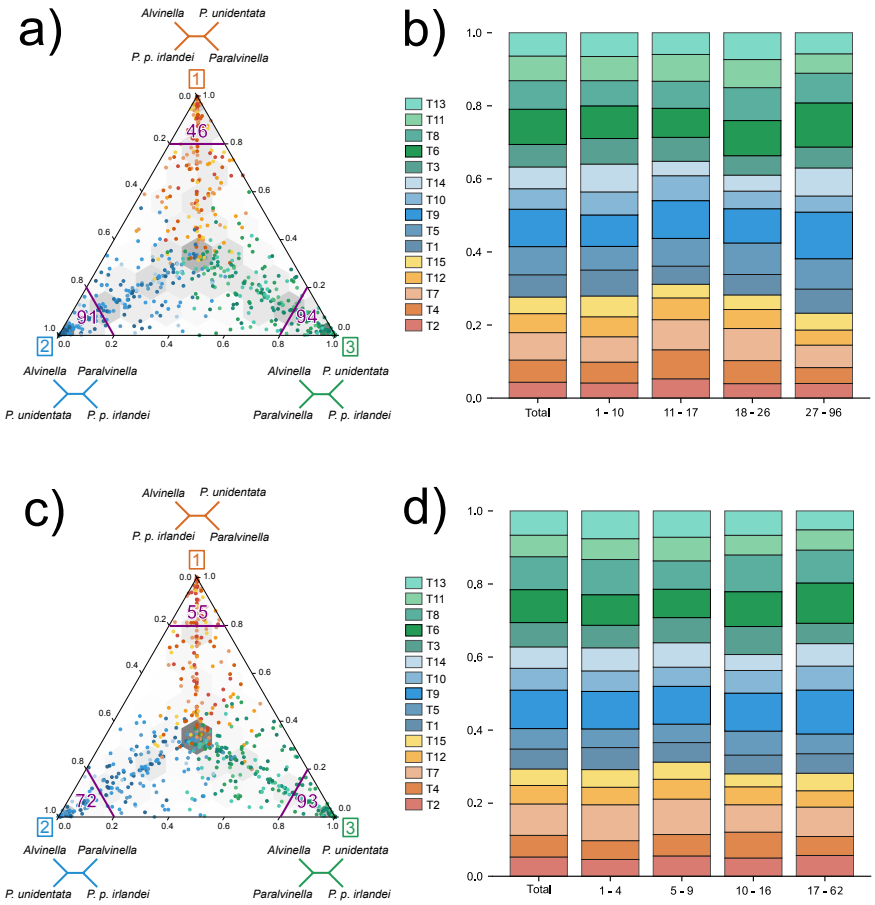

Fig. 3: Topology weight across genes. Topologies are colored according to the maximum probability of their associated lineages' quartet. (a) Nucleotide-encoded gene associations with the three possible scenarios labelled 1, 2 and 3 at the root of the family. Number of genes strongly associated with one scenario ($Pr > 0.8$) are indicated in each pole. (b) Distribution of genes' probabilities according to the number of parsimony-informative sites for nucleotide sequences. (c) Amino acid-encoded gene associations with the three possible scenarios labelled 1, 2 and 3 at the root of the family. Number of genes strongly associated with one scenario ($Pr > 0.8$) are indicated in each pole. (b) Distribution of genes' probabilities according to the number of parsimony-informative sites for amino-acid sequences.

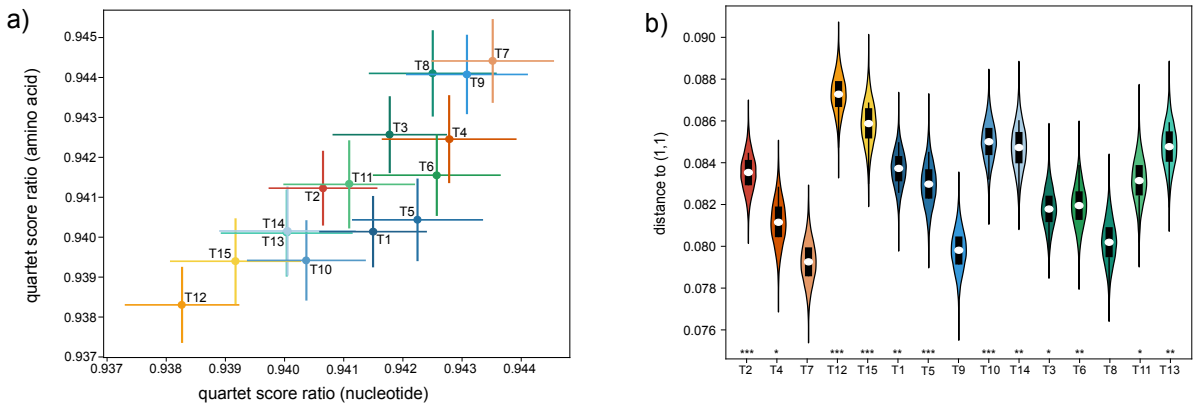

Fig. 4: Fitting scores of the species tree topology by ASTRAL from the 15 constrained topologies, with standard deviations. Each evaluation is the result of 10,000 resamplings of gene trees. (a) biplots of quartet scores (in %) obtained for each topology from the amino acid and nucleotide sequence datasets. A score closer to 1 indicates that the species-tree topology is more in agreement with the set of gene trees. (b) Euclidean distance d estimated for each topology to the point (1,1), which represents an ideal case where all gene trees agree with the proposed species tree topology. For each topology Tx, the hypothesis $d_7 < d_X$ is tested with a modified AU test. If the test fails, then T7 is not better than Tx: $*$: $p - value < 5\%$, $**$: $p - value < 1\%$, $***$: $p - value < 0.1\%$.

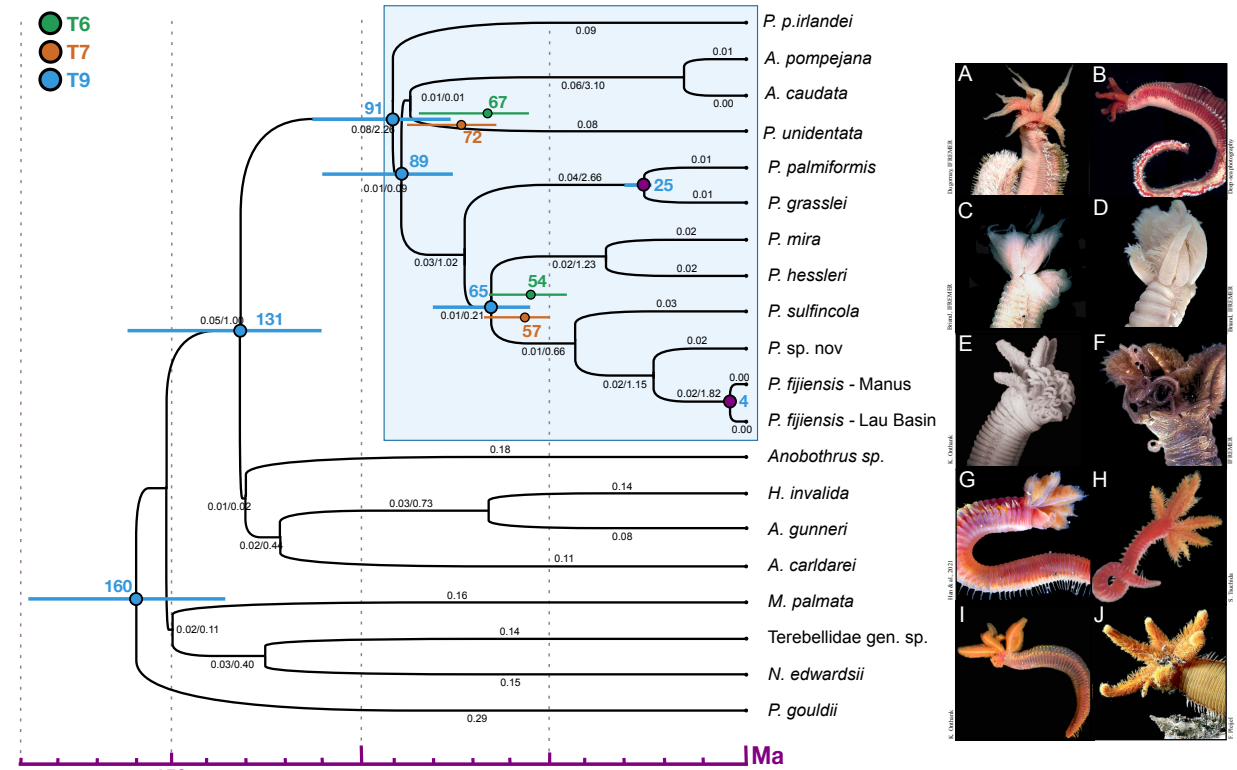

Fig. 5: Chronogram of the family Alvinellidae under the CIR model with two calibration points (in purple), assuming T9 topology. Branch lengths are reported in expected numbers of amino-acid substitutions (LG+Γ) and coalescent time unit (obtained from unconstrained gene trees). The ages of some nodes of interest with 95% confidence intervals are reported in millions of years ago (Ma). The ages of birth of the Alvinellidae family and its spread in the eastern Pacific are also reported assuming the T6 and T7 topologies with 95% confidence interval. 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*.