The genome of a vestimentiferan tubeworm (*Ridgeia piscesae*) provides insights into its adaptation to a deep-sea environment

Muhua Wang[1†], Lingwei Ruan[2†], Meng Liu[3†], Zixuan Liu[1], Jian He[1], Long Zhang[1], Yuanyuan Wang[1], Hong Shi[2], Mingliang Chen[2], Feng Yang[2], Runying Zeng[2], Jianguo He[1,4,5*], Changjun Guo[1,4*], Jianming Chen[2,6*]

[1]State Key Laboratory for Biocontrol, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), China-ASEAN Belt and Road Joint Laboratory on Mariculture Technology, School of Marine Sciences, Sun Yat-sen University, Zhuhai 519000, China.

[2]State Key Laboratory Breeding Base of Marine Genetic Resources, Key Laboratory of Marine Genetic Resources of Ministry of Natural Resources, Third Institute of Oceanography, Ministry of Natural Resources, Fujian Key Laboratory of Marine Genetic Resources, Xiamen 361005, China.

[3]Novogene Bioinformatics Institute, Beijing 100083, China.

[4]Guangdong Province Key Laboratory for Aquatic Economic Animals, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China.

[5]Maoming Branch, Guangdong Laboratory for Lingnan Modern Agricultural Science and Technology, Maoming 525435, China.

[6]Fujian Key Laboratory on Conservation and Sustainable Utilization of Marine Biodiversity, Fuzhou Institute of Oceanography, Minjiang University, Fuzhou, 350108, China.

† These authors contributed equally to this work.

*** Correspondence:**
Jianming Chen
chenjm@mju.edu.cn

Changjun Guo
gchangj@mail.sysu.edu.cn

Jianguo He (ORCID: 0000-0003-2460-7061)
lsshjg@mail.sysu.edu.cn

Abstract

Vestimentifera (Polychaeta, Siboglinidae) is a taxon of deep-sea worm-like animals living in the deep-sea hydrothermal vent and cold seep areas. The morphology and lifespan of *Ridgeia piscesae*, which is the only vestimentiferan tubeworm species found in the hydrothermal vents on the Juan de Fuca Ridge, vary greatly according to the endemic environments. Recent analyses have revealed the genomic basis of adaptation in three vent- and seep-dwelling vestimentiferan tubeworms. However, the evolutionary history and mechanism of adaptation in *R. piscesae*, a unique species in the family *Siboglinidae*, is remained to be investigated. Here we report a high-quality genome of *R. piscesae* collected at Cathedral vent of the Juan de Fuca Ridge. Comparative genomic analysis revealed that that the high growth rates of vent-dwelling tubeworms might derive from small genome size. The small genome sizes of these tubeworms are attributed to the repeat content but not the number of genes and intron sizes. Additionally, four genes involved in cell proliferation were subject to positive selection in the genome of *R. piscesae*, suggesting that, besides apoptosis, cell proliferation is important for regulating growth rate in this species.

Keywords

vestimentiferan tubeworm, *Ridgeia piscesae*, genome, deep-sea adaptation, lifespan,

Introduction

The discovery of deep-sea hydrothermal vents and cold seeps, as well as their associated ecosystems have revolutionized our view of biology and understanding of the energy sources that fuel primary productivity on Earth (Corliss et al., 1979; Paull et al., 1984; Petersen et al., 2011). Hydrothermal vents are areas on the ocean floor where hot, anoxic, chemical-rich water is expelled into the cold, oxygen-rich deep ocean (Von Damm, 1995). Cold seeps are areas where methane, hydrogen sulfide, and other hydrocarbons seep or emanate as gas from deep geologic sources (Suess, 2014). Both hydrothermal vents and cold seeps are characterized by high hydrostatic pressure, darkness, lack of oxygen and photosynthesis-derived nutrients, and high concentrations of toxic chemicals (Levin, 2005). Organisms inhabiting around hydrothermal vents and cold seeps develop unique characters to adapt to the adverse conditions (Grassle, 1985; Childress and Fisher, 1992). Due to the complete absence of light, hydrothermal vent and cold seep ecosystems are driven by the chemosynthesis instead of photosynthesis (Stewart et al., 2005; Vanreusel et al., 2009). The process is completed by the chemosynthetic microorganisms, which cooperate with a variety of macrobenthos to form chemosynthetic symbioses and contribute to the primary production supporting the ecosystem (Dick, 2019).

Vestimentifera (Polychaeta, Siboglinidae) is a taxon of deep-sea worm-like animals living in the deep-sea hydrothermal vent and cold seep areas (Bright and Lallier, 2010). The body of adult vestimentiferan tubeworm is enclosed in a chitinous tube that is closed at the posterior end. Vestimentiferan tubeworms lack a digestive tract and rely on symbiosis with chemoautotrophic microorganisms, which inhabit in an specialized internal organ, to derive their metabolic needs (Vrijenhoek, 2010). The first discovery of

chemoautotrophic symbiont in *Riftia pachyptila*, a vestimentiferan tubeworm inhabiting around hydrothermal vents on the East Pacific Rise (EPR), initiated the intensive study of these deep-sea tubeworms (Corliss et al., 1979). The body of adult tubeworm is comprised of four main parts. The anteriorly located branchial plume is the primary site of gas exchange with the environment. Below the plume is the vestimentum, where heart, gonopores and a simplified brain are located. The trophosome, which is primarily composed of symbiont-containing bacteriocytes and blood vessels, is located below vestimentum. And the segmented opisthosoma is located below vestimentum (Jones, 1981; Hand, 1987; Schulze, 2001).

Lifespan varies greatly between the vent- and seep-dwelling tubeworms (Lutz and Kennish, 1993). The lifespan of *R. pachyptila*, which thrives in relatively strong and continuous diffuse hydrothermal flow, was estimated to be less than 10 years (Urcuyo et al., 2007). In contrast, *Lamellibrachia luymesi*, which live around cold seeps in the Gulf of Mexico, can live up to 250 years (Bergquist et al., 2000). In general, vent-dwelling tubeworms grow faster than seep-dwelling tubeworms. The growth rate of *R. pachyptila* can reach up to about 160 cm yr^{-1}, while the growth rate of *L. luymesi* is only about 3 cm yr^{-1} (Shank et al., 1998; Thiebaut et al., 2002; Fisher et al., 2008). Previous analysis revealed that both vent- and seep-dwelling tubeworms have high rates of cell proliferation. Recent analysis suggested that the variation of growth rates is attributed to the variation of apoptosis between vent- and seep-dwelling tubeworms, where apoptosis is substantially downregulated in vent-dwelling species (Pflugfelder et al., 2009).

The Juan de Fuca Ridge in the northeast Pacific Ocean is characterized by a broad heterogeneity of chemical environments, ranging from vigorous, high-temperature vents to

95 diffuse flow (Tivey et al., 1999). The hydrothermal vents on the Juan de Fuca Ridge
96 provide numerous biotic habitats, which support the growth of a large quantity of endemic
97 organisms (Tsurumi and Tunnicliffe, 2003). Although the biomass of endemic
98 hydrothermal vent fauna is high, there is only a few macrofauna species dominating a
99 particular vent community (Lutz et al., 1998; Tunnicliffe et al., 1998). *Ridgeia piscesae*
100 Jones (1985) is the only vestimentiferan tubeworm species in the hydrothermal vents on
101 the Juan de Fuca Ridge. The tubeworm occurs at high density in most of the vents, acts as
102 an ecosystem-structuring species by providing habitats for several other organisms and
103 serving as a primary producer through chemosynthetic endosymbiosis (Childress and
104 Fisher, 1992; Urcuyo et al., 2003).

105 *Ridgeia piscesae* adopts different strategies to adapt to diverse environments in the
106 vents on Juan de Fuca Ridge. First, a strong phenotypic plasticity, where a given genotype
107 expresses different phenotypes in different ecological setting, was identified in *R. piscesae*
108 (Carney et al., 2007). Two extreme growth forms (morphotypes) of *R. piscesae*, "short-fat"
109 and "long-skinny", were discovered in the geologically and chemically diverse vent fields
110 (Southward et al., 1995). "Short-fat" *R. piscesae* prefers relatively high flow vent fluid of
111 high temperature (up to 30 ℃) and high concentrations of sulfide, while "long-skinny"
112 morphotype adapts to ambient temperature (2 ℃) and low concentrations of sulfide at areas
113 of diffuse hydrothermal fluids (Carney et al., 2007). These two morphotypes diverge in
114 several morphological characters (Jones, 1987). The tube of "short-fat" morphotype has a
115 generally constant diameter of 2-3 centimeters, while the tube diameter of "long-skinny"
116 morphotype reduces from the anterior to posterior. This morphotype acquires sulfide by
117 their buried posterior tube sections in areas where sulfide is generally not detectable around

118 their plumes (Urcuyo et al., 2003). Thus, unlike other vent-dwelling vestimentiferan
119 species, *R. piscesae* can thrive in the areas of diffuse vent fluids. Second, lifespan of *R.*
120 *piscesae* vary greatly according to the endemic environments (Urcuyo et al., 2007). The
121 species can grow with high growth rates ranging from 6 to 95 cm yr^{-1} under favorable
122 condition, and can grow very slowly when exposed to low levels of vent flow and sulfide
123 (Tunnicliffe et al., 1997; Sarrazin et al., 1998). Strong phenotypic plasticity and flexible
124 lifespan allows *R. piscesae* to adapt to diverse habitats and makes it a unique species in the
125 family *Siboglinidae*. Recent genomic analyses have revealed the genetic basis of adaptation
126 in three vent- and seep-dwelling vestimentiferan tubeworms (Li et al., 2019; Sun et al.,
127 2021; de Oliveira et al., 2022). Although its critical role in supporting the vent ecosystem,
128 the evolutionary history and mechanism of adaptation in *R. piscesae*, a unique deep-sea
129 tubeworm, is remained to be investigated.

130 Here, we assembled and annotated a high-quality genome of *R. piscesae* collected at
131 Cathedral vent of the Juan de Fuca Ridge. Evolutionary analysis inferred that two vent-
132 dwelling species (*R. piscesae* and *R. pachyptila*) started to diverge due to the separation of
133 the Gorda/Juan de Fuca/Explorer (GFE) ridge systems and the East Pacific Rise (EPR).
134 Comparative genomic analysis revealed that that the high growth rates of vent-dwelling
135 tubeworms might derive from small genome size. The small genome sizes of these
136 tubeworms are attributed to the repeat content but not the number of genes and intron sizes.
137 Additionally, four genes involved in cell proliferation were subject to positive selection in
138 the genome of *R. piscesae*, suggesting that, besides apoptosis, cell proliferation is important
139 for regulating growth rate in this species.

140

Results

Genome assembly and annotation

The samples of *R. piscesae* were collected from the deep-sea hydrothermal vent at Cathedral vent, Main Endeavor Field of the Juan de Fuca Ridge (47° 56' N, 129° 05' W, 2,181 m depth). Short-insert paired-end (180 bp, 300 bp and 500 bp) and long-insert mate-pair (2 kb, 5 kb, 10 kb and 15 kb) sequencing libraries were constructed and sequenced on the Illumina HiSeq 2000 platform. A total of 247.74 Gb sequencing data was generated (**Supplementary Table 1**). Based on the *k*-mer distribution of 180 bp paired-end Illumina read, the genome size of was estimated to be 694.79 Mb with a heterozygosity of 1.2% (**Supplementary Fig. S1**). The final assembly of *R. piscesae* genome was 574.96 Mb with a contig N50 size of 10.42 kb and a scaffold N50 size of 230.23 kb (**Supplementary Table S2**).

A total of 87.4% sequencing reads can be aligned unambiguously to the assembled *R. piscesae* genome sequence, covering 99.74% of the assembly (**Supplementary Table S3**). In addition, 99.63% of Trinity assembled sequences (Unigenes) could be aligned to the assembly (**Supplementary Table S4**). The integrity of genome assembly was further assessed using benchmarking universal single-copy orthologs (BUSCO) tools. The result showed that 897 of 978 (92.6%) single-copy metazoan genes (obd10) were captured in the assembly (**Supplementary Table S5**). These results demonstrated that the *R. piscesae* genome is of high quality, and comparable to the previously published tubeworm genomes (**Table 1**) (Li et al., 2019; Sun et al., 2021; de Oliveira et al., 2022).

Transposable elements (TEs) accounted for 30.17% in the *R. piscesae* genome assembly, with long interspersed elements (LINEs, 8.08%) as the most abundant class of TEs (**Supplementary Table S6**). The *R. piscesae* genome encodes 24,096 protein-coding genes, of which 95.54% are annotated based on known proteins in diverse public protein databases (**Supplementary Table S7**).

Phylogenomic analyses

To infer the evolutionary history of *R. piscesae*, a maximum-likelihood (ML) phylogenetic tree was constructed using single-copy orthologs of *R. piscesae* and 14 metazoans with *Adineta vaga* as an outgroup (**Fig. 1, Supplementary Fig. S2, Supplementary Table S8**). Two vent-dwelling tubeworms (*R. piscesae* and *R. pachyptila*) formed a clade. *Paraescarpia echinospica* and *L. luymesi* from cold seeps are basal to the vent clade. These results corroborate the view that vent-dwelling tubeworms might derived from their seep-dwelling relatives (Black et al., 1997; Halanych, 2005).

Ridgeia piscesae is endemic to the Gorda/Juan de Fuca/Explorer (GFE) ridge systems, and *R. pachyptila* is discovered on the East Pacific Rise (EPR). The subduction of the Farallon-Pacific Ridge separated GFE and EPR between 28.5-35 Ma. Molecular clock analysis revealed that *R. pachyptila* diverged from *R. piscesae* approximately 33.7 million years (Ma). This result suggests these two vent-dwelling species started to diverge due to adaptation to the newly formed ridge systems. The divergence time of *L. luymesi* and other three tubeworms was estimated to be approximately 65.1 Ma, corroborating the

183 view that modern vestimentiferan tubeworms started to diverge during the early Cenozoic
184 Era (Little and Vrijenhoek, 2003; Li et al., 2019; Sun et al., 2021).

185

186 **3.3 Genome evolution of vestimentiferan tubeworms**

187 It has been demonstrated that several factors, including repeat content, number of
188 genes, and the intron sizes, contributed to the variation of genome sizes among different
189 organisms (Lynch, 2007; Niu et al., 2022). In addition, previous reports proposed that the
190 differences of genome sizes among deep-sea tubeworms might be attributed to the numbers
191 of repetitive elements and genes (Sun et al., 2021; de Oliveira et al., 2022). The assembled
192 size of *R. piscesae* genome is closed to the size of *R. pachyptila*, but smaller than the sizes
193 of two seep-dwelling tubeworms (*L. luymesi* and *P. echinospica*) (**Table 1**). The genomes
194 of cold seep dwelling tubeworms have more transposable elements (TE), especially DNA
195 transposons, LINEs and LTR retrotransposons, than hydrothermal vent dwelling
196 tubeworms (**Fig. 2A**). TEs accounted for 38.2% and 55.1% of *L. luymesi* and *P.*
197 *echinospica* genomes, and they constituted 30.2% and 29.9% of the genomes of *R. piscesae*
198 and *R. pachyptila*. A strong positive correlation (R^2 = 0.98, P = 0.0052) was identified
199 between genome size and repeat content in these four species (**Fig. 2B**), suggesting that
200 TEs are a major contributor to the genome size evolution in vestimentiferans. Repeat
201 landscape plots indicate that TE activity is different between *R. pachyptila* and three other
202 tubeworm species (**Fig. 3**). There are recent expansions of TEs in the genomes of *L. luymesi*,
203 *P. echinospica*, and *R. piscesae*, but not in *R. pachyptila*. The main contributor to recent
204 TE expansions in *L. luymesi*, *P. echinospica*, and *R. piscesae* appear to have been LINEs

and DNA transposons. Nonetheless, only LINEs were expanded recently in the genome of *R. pachyptila* (**Fig. 3D**).

The number of annotated gene models in *R. piscesae* genome (24,096) is closed to the ones in the genomes of *R. pachyptila* (25,984) and *P. echinospica* (22,642) but less than the ones in *L. luymesi* genome (38,998). Introns account for 220.1 Mb and 204.7 Mb in the genomes of two seep-dwelling tubeworms (*L. luymesi* and *P. echinospica*), as well as 264.8 Mb and 234.5 Mb in the genomes of two vent-dwelling tubeworms (*R. pachtypila* and *R. piscesae*) (**Supplementary Table S9**). Average length of introns in the genome of *R. piscesae* is longer than introns of other three species. Additionally, two vent-dwelling tubeworms with smaller genome sizes have higher ratios of intron / exon length than the seep-dwelling tubeworms. Thus, gene number and intron size do not contribute to the differences of genome sizes between seep- and vent-dwelling tubeworms.

Previous study revealed that *R. pachyptila* experienced reductive evolution with more contracted than expanded gene families in the genome (de Oliveira et al., 2022). Gene-family analysis of four tubeworm species identified a core set of 10,225 gene families (**Fig. 4A**). In total, 601 and 279 lineage-specific gene families were identified in *R. piscesae* and *R. pachyptila*, respectively, which are much less than the ones in *L. luymesi* (1181) and *P. echinospica* (1045). Additionally, gene-family analysis of 12 lophotrochozoans revealed that the numbers of expanded gene families were substantially less than contracted gene families in the two vent-dwelling tubeworms, while more gene families were expanded than contracted in their seep-dwelling counterparts (**Fig. 4B**). These results indicate that the genomes of vent-dwelling tubeworms were characterized by gene loss.

227 *Hox* genes are a set of conserved regulators that specify regions of the body plan of
228 an embryo along anterior-posterior axis in metazoans (Pearson et al., 2005). One of the
229 Hox genes (*Antp*) plays a role in the development of posterior segment of several marine
230 annelids (Bakalenko et al., 2013). Loss of *Antp* was apparent across all four tubeworm
231 genomes (**Supplementary Fig. S3**), corroborating the view that the loss of *Antp*
232 contributes to the reduced segmentation of the posterior region of juvenile worms in
233 vestimentiferans (Sun et al., 2021). *Lox2* gene is missing in the genome of *L. luymesi* but
234 presented in the genomes of three other tubeworms, suggesting the loss of this gene might
235 be a lineage-specific event.

236

237 **3.4 Genomic basis of deep-sea adaptation**

238 Hemoglobins (Hbs) in vestimentiferan tubeworms, which bind oxygen and sulfide
239 simultaneously and provide substrate for chemosynthesis by the symbionts, facilitate the
240 adaptation of these species to deep-sea reducing environments. Four heme-containing
241 chains were identified (A1, A2, B1, B2) in hemoglobins of vestimentiferans (Zal et al.,
242 1997). To elucidate the evolution of Hbs in vestimentiferans, we identified Hb genes in the
243 genomes of four tubeworms species. A single copy of A2 and B2 Hb genes, as well as two
244 copies of A1 genes were identified in each of the tubeworm genomes (**Fig. 5**). Previous
245 studies found that the group of B1 Hbs were significantly expanded in *L. luymesi*, *P.*
246 *echinospica*, and *R. pachyptila* (Li et al., 2019; Sun et al., 2021; de Oliveira et al., 2022).
247 With 17 identified genes, the group of B1 Hbs was also expanded in the genome of *R.*
248 *piscesae*. The free cysteine residues in A2 and B2 chains contribute to the sulfide-binding

ability of the vestimentiferan Hbs (Bailly et al., 2002). Additionally, the free cysteine was identified at the same position in B1 Hb genes as in A2 Hb genes of *R. piscesae* (**Supplementary Fig. S4**), corroborating the view that the free cysteines might also contribute to sulfide binding in B1 hemoglobin chain of deep-sea tubeworms (Li et al., 2019).

Recent reports revealed that most enzymes related to amino acid biosynthesis were lost in *L. luymesi* and *R. pachyptila* (Li et al., 2019; de Oliveira et al., 2022). To gain better insight into nutrient dependence of endosymbionts in vestimentiferans, we identified key enzymes involved in amino acid biosynthesis in the genomes of tubeworms (**Fig. 6**). All four tubeworms (*L. luymesi* and *R. pachyptila*, *R. piscesae* and *P. echinospica*) lack most key enzymes related to amino acid biosynthesis, corroborating the view that vestimentiferan tubeworms mainly relied on endosymbionts for synthesizing amino acids (Li et al., 2019).

The expansion of gene families is considered as a major driver of adaptation and speciation (Sharpton et al., 2009). Thus, we performed gene family expansion and contraction analysis with 4 vestimentiferan tubeworms and 8 other lophotrochozoans (**Fig. 4B**). In total, 10 gene families were significantly expanded in the genomes of all four tubeworms compared to other 8 lophotrochozoans ($P < 0.05$) (**Supplementary Table S10**). Gene ontology analysis revealed the expanded gene families were involved in the process of chitin binding and innate immunity. Furthermore, 18 gene families were significantly expanded in the genomes of two cold-seep tubeworms. The expanded gene families were involved in the process of DNA repair, innate immunity, and protein stability (**Supplementary Table S11**).

In addition to expanded and contracted gene families, we identified positively selected genes (PSGs) in the genome of *R. piscesae*. Compared with other 11 lophotrochozoan species, 9 PSGs were identified in *R. piscesae* (**Supplementary Table S12**). Interestingly, four genes (alkB homolog 2, alpha-ketoglutarate-dependent dioxygenase, *ALKBH2*; Derlin-1, *DERL1*; Ras-related and estrogen-regulated growth inhibitor, *RERG*; AN1-type zinc finger protein 2B, *ZFAND2B*) involved in cell proliferation were subjected to positive selection in *R. piscesae*. Two of these genes (*ALKBH2* and *DERL1*) promote cell proliferation. ALKBH2 is responsible for protecting the genome from 1-meA damage by repairing the damage in double-stranded DNA(Aas et al., 2003). *ALKBH2* promotes cell proliferation and is overexpressed in several types of tumor cells (Wilson et al., 2018). DERL1 participates in endoplasmic reticulum (ER)-associated degradation response and unfolded protein response (UPR) (Eshraghi et al., 2014). *DERL1* is responsible for cell proliferation and promotes the progression of several types of cancers (Dong et al., 2013). Interestingly, two other genes (*RERG* and *ZFAND2B*) inhibit cell proliferation. Overexpression of *RERG*, a member of the RAS superfamily of GTPase, inhibits cell proliferation and tumor formation (Finlin et al., 2001; Ho et al., 2017). ZFAND2B reduces the abundance of IGF1R, a kinase that activates cell proliferation, in a proteasome-dependent manner (Osorio et al., 2016). These results suggest that, besides apoptosis, the regulation of cell proliferation also contributes to the variation of growth rates in *R. piscesae*.

Discussion

In many hydrothermal vent and cold seep ecosystems, vestimentiferan tubeworms are among the dominant megafauna at habitats where hydrogen sulfide is present (Cavanaugh et al., 1981; Tunnicliffe, 1992). These deep-sea tubeworms occupy a broad environmental gradient from areas where sulfide-rich fluids emanate as relatively vigorous and continuous diffuse vent flow to areas where vent fluids seep slowly from the seafloor (Urcuyo et al., 2007). Seep-dwelling tubeworms grow slowly as they live in environments where exposure to seep fluid is very low (Julian et al., 1999; Fisher et al., 2008). The lifespan of most vent-dwelling tubeworms is relatively short due to the habitats are short-lived. They thrive in strong and continuous diffuse hydrothermal flow and dies when vent flow subsides or sulfide levels decrease due to biotic or abiotic factors (Fisher et al., 1988; Shank et al., 1998). Great variations in lifespan and morphology were found among different morphotypes of *R. piscesae*, which is endemic to the highly heterogeneous chemical environments of hydrothermal vents on the Juan de Fuca Ridge. The "Short-fat" morphotype of *R. piscesae* that live around vigorous vents with high concentrations of sulfide grow rapidly, while the "long-skinny" morphotype inhabit around diffuse flow with low concentrations of sulfide have much slower growth rates (Southward et al., 1995; Carney et al., 2007). Unlike the "short-fat" morphotype, the tube diameter of "long-skinny" morphotype reduces from the anterior to posterior. As the concentration of sulfide is low in the diffuse vent flow, thin and transparent end facilitates the "long-skinny" morphotype to acquire sulfide by the posterior tube (Urcuyo et al., 2003). Thus, *R. piscesae* exhibit a greater tolerance to varying physicochemical conditions than any other known vestimentiferans (Tivey et al., 1999; Bright and Lallier, 2010). However, the evolutionary

316 history and underlying genetic mechanisms of adaptation in *R. piscesae* are remained to be
317 investigated.

318 Here we assembled and annotated a high-quality reference sequence of *R. piscesae*.
319 Phylogenomic analysis was performed among 4 vestimentiferan tubeworms and 11
320 metazoans using single-copy orthologs (**Fig. 1, Supplementary Fig. S2**). *Ridgeia*
321 *pachyptila* and *R. piscesae*, the two tubeworm species that are endemic to hydrothermal
322 vents of the eastern Pacific, formed a clade in the phylogenetic tree. Additionally,
323 *Lamellibrachia luymesi* from cold seep appeared sister to three other deep-sea tubeworms,
324 corroborating the view that this species derived early from other tubeworm (Black et al.,
325 1997). *Ridgeia piscesae* and *R. pachyptila* are endemic to the GFE ridge systems and the
326 EPR, respectively. The divergence time between these two vent-dwelling species was
327 estimated to be approximately 33.7 Ma. This result suggests that *R. piscesae* and *R.*
328 *pachyptila* started to diverge after the formation of GFE and EPR between 28.5-35 Ma.

329 It was proposed that natural selection and adaptive process shaped genome size
330 evolution (Gregory, 2001). Previous analyses revealed that genome sizes are correlated
331 with several phenotypic traits, including cell size, rates of metabolism and growth
332 (Cavalier-Smith, 1982; Vinogradov, 1995; Wyngaard et al., 2005; Wright et al., 2014). A
333 negative correlation between genome size and growth rate was identified in several species,
334 as organisms with smaller genomes might undergo more rapid replication time of their
335 genome (Wyngaard et al., 2005; Tenaillon et al., 2016). To investigate the genomic basis
336 of their high growth rates, we studied the factors which might contribute to the evolution
337 of genome sizes in four vestimentiferan tubeworms. Firstly, we studied the relationship
338 between repeat content and genome size. A strong positive correlation (R^2 = 0.98, P =

0.0052) between repeat content and genome size was identified in these tubeworm species (**Fig. 2B**), corroborating the view that repeat content contributes to the variation of genome sizes in tubeworm species (Sun et al., 2021). Secondly, we evaluated the number and size of genes in tubeworm genomes. *Lamellibrachia luymesi* genome has the most annotated gene models (38,998), following by *R. pachyptila* (25,984), *R. piscesae* (24,096), and *P. echinospica* (22,642). Average length of introns in the genome of *R. piscesae* is longer than introns of other three species. In addition, two vent-dwelling tubeworms with smaller genome sizes have higher ratios of intron / exon length than the seep-dwelling tubeworms. These results suggest that the variation of genome sizes in tubeworms is not attributed to gene number and intron length. Lastly, we performed gene family expansion and contraction analysis. Lineage-specific gene families of *R. piscesae* (601) and *R. pachyptila* (279) are much less than the ones of *L. luymesi* (1181) and *P. echinospica* (1045) (**Fig. 4A**). In addition, gene-family expansion and contraction analysis revealed that the numbers of expanded gene families were substantially less than contracted gene families in the two vent-dwelling tubeworms (**Fig. 4B**). These results indicate that, besides *R. pachyptila*, *R. piscesae* experienced reductive evolution. Taken together, these results indicate that the high growth rates of vent-dwelling tubeworms might derive from small genome size and protein family. The small genome sizes of these tubeworms are attributed to the repeat content but not the number of genes and intron sizes.

Previous immunohistochemical and ultrastructural cell cycle analyses revealed that cell proliferation activities of *L. luymesi* and *R. pachyptila* are as high as in tumor. The balanced activities of proliferation and apoptosis in the epidermis lead to the slow growth in *L. luymesi* from cold seeps, while apoptosis is substantially downregulated in this tissue

of *R. pachyptila* that maintains a high growth rate (Pflugfelder et al., 2009). Unlike other vestimentiferan tubeworms, growth rates vary greatly among different individuals of *R. piscesae* (Tunnicliffe et al., 1997; Sarrazin et al., 1998). Four genes involved in the regulation of cell proliferation were identified to be positively selected in *R. piscesae*. Interestingly, two of these genes promote cell proliferation, whereas two other genes inhibit cell proliferation. This result indicates that both cell proliferation and apoptosis involve in the regulation of growth in *R. piscesae*. Furthermore, up-regulation and downregulation of cell proliferation might play important role in the variation of growth rate in this species.

Materials and Methods

Sampling and sequencing

The samples of *R. piscesae* were obtained during *Alvin* dive 4243 from the deep-sea hydrothermal vent at Cathedral vent, Main Endeavor Field of the Juan de Fuca Ridge (47° 56' N, 129° 05' W, 2,181 m depth) on August 9, 2006. Genomic DNA (gDNA) was extracted from muscle of the specimen using a standard phenol/chloroform extraction protocol and broken into random fragments for whole-genome shotgun (WGS) sequencing. Agarose gel electrophoresis was used to check the quality of the gDNA, and Qubit system was used to quantify the gDNA. Short-insert paired-end libraries (180 bp, 300 bp and 500 bp) were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) according to the standard protocol, respectively. Large-insert mate-pair libraries (2 kb, 5 kb, 10 kb and 15 kb) were prepared following the Cre-lox recombination-based protocol (Van Nieuwerburgh et al., 2012). All DNA libraries were subjected to paired-end sequencing on the Illumina Hiseq 2000 platform (Illumina). Muscle samples were also collected for constructing RNA sequencing (RNA-seq) library. Total RNA was extracted with TRIzol reagent (Molecular Research Center). Paired-end library for RNA-seq was constructed using the Paired-End Sample Preparation Kit (Illumina) and sequenced on the Illumina Hiseq 2000 platform (Illumina).

Genome assembly

NGS QC toolkit (v2.1) (Patel and Jain, 2012) was used to evaluate the quality of raw sequencing reads and filter high-quality reads. High-quality reads were obtained by filtering out the following types of reads: (1) reads with >= 10% unidentified nucleotides (N); (2) reads with adaptor contamination; (3) reads with >= 20% bases having Phred quality score <= 5; (4) duplicated reads generated by PCR amplification during the library construction process.

The size and heterozygosity of *R. piscesae* genome were estimated using the high-quality short-insert paired-end reads (180 bp) by *k*-mer frequency-distribution method. The

399 number of *k*-mers and the peak depth of *k*-mer sizes at 19 was obtained using
400 GenomeScope2 (v1.0.0) (Ranallo-Benavidez et al., 2020). Due to the high heterozygosity
401 of *R. piscesae* genome, a modified version of SOAPdenovo (Wang et al., 2017) was
402 implemented for genome assembly. In brief, all short-insert paired-end reads were applied
403 for contig assembly. Depth of coverage was obtained for each contig using SOAPdenovo
404 with the parameters '-e 1 -M 0 -R', and the contigs with depth less than 60 were identified
405 as heterozygous contigs. All WGS reads were aligned to the heterozygous contigs using
406 SOAPdenovo. And links were generated between heterozygous contigs when supported by
407 a minimum of three read pairs. Heterozygous contigs were clustered into bubble clusters
408 based on the orientation and distance between heterozygous contigs. If two contigs
409 represented two potential haplotypes in a bubble structure, the longer one was retained to
410 ensure the integrity of contig assembly.

411 To scaffolding the contigs, all short-insert paired-end and long-insert mate-pair reads
412 were realigned onto the contig sequences using SOAPdenovo. Duplicated contigs that had
413 high depth of coverage and conflicting connections to the unique contigs were masked
414 during scaffolding. A hierarchical assembly strategy was used to construct contigs into
415 primary scaffolds by adding the ascending insert size reads gradually. Finally, all short-
416 insert reads were realigned onto the scaffold sequences to fill the gaps with the GapCloser
417 program implemented in SOAPdenovo (Luo et al., 2012).

418

419 **Genome quality assessment**

420 To assess the completeness of the *R. piscesae* genome, high-quality short-insert
421 paired-end reads were mapped to the genome assembly using Burrows-Wheeler Aligner
422 (BWA) (v0.7.17) (Li and Durbin, 2009) with parameters of '-o 1 -e 5 -t 8 -n 15'. In addition,
423 all RNA-seq reads were *de novo* assembled using Trinity (v2.9) (Grabherr et al., 2011).
424 The Trinity assembled sequences (Unigenes) with length >= 500 bp were mapped to the *R.*
425 *piscesae* genome using BLAT (v35.1) (Kent, 2002) with default parameters and an identify
426 cutoff of 90%. The completeness of the assembly was also evaluated using benchmarking

universal single-copy orthologs (BUSCO) (v3.1.0) (Simao et al., 2015) with 978 metazoa single-copy orthologous genes (obd10).

Genome annotation

Tandem repeats in the genome were predicted using the program Tandem Repeats Finder (TRF) (v4.09) (Benson, 1999) with default parameters. Transposable elements (TEs) were identified using the homology-based and *de novo* prediction approaches. For homology-based prediction, RepeatMasker (v4.1.0) (http://www.repeatmasker.org/) were conducted to identify repeat sequences against the Repbase library. For *de novo* prediction, RepeatModeler (v2.0.1) (http://repeatmasker.org/RepeatModeler.html), LTR-Finder (v1.0.7) (Xu and Wang, 2007), RepeatScout (v1.0.5) (Price et al., 2005) and Piler (v1.0) (Edgar and Myers, 2005) were used to construct *de novo* repeat libraries. RepeatMasker (v4.1.0) was run against these libraries to search repeat elements.

Protein-coding genes in *R. piscesae* genome were predicted with three approaches: homology-based prediction, *ab initio* prediction and RNA-seq-based prediction. Protein-coding sequences of *Lottia gigantea*, *Helobdella robusta*, *Capitella teleta*, *Schistosoma mansoni*, *Caenorhabditis elegans*, *Anopheles gambiae*, *Drosophila melanogaster* and *Homo sapiens* were aligned to the *R. piscesae* genome using tblastn with a cut off E-value of 1e-5. GeneWise (v2.4) (Birney et al., 2004) was employed to predict gene models. For *ab initio* prediction, Augustus (v3.3.2) (Stanke and Morgenstern, 2005), Genscan (Aggarwal and Ramaswamy, 2002), Geneid (v1.3) (Parra et al., 2000), GlimmerHMM (v3.0.4) (Majoros et al., 2004) and SNAP (Korf, 2004) were used to predict genes on the repeat-masked genome. For RNA-seq-based prediction, Trinity (v2.9) generated sequences (Unigenes) were aligned against the genome assembly with BLAT (v35.1) (identify >= 0.95 and align rate >= 0.95) (Kent, 2002). In addition, the RNA-seq reads from were aligned to the *R. piscesae* genome using Tophat (v2.1.1) (Trapnell et al., 2009). And gene structures were predicted using Cufflinks (v2.2.1) (Trapnell et al., 2010). EvidenceModeler (EVM) (v1.1.1) (Haas et al., 2008) was used to integrate all gene models derived from these three approaches into a non-redundant gene set.

Functional annotation was performed using BLASTP searches against SwissProt and TrEMBL databases (Bairoch and Apweiler, 2000) with a E-value cut-off of 1e-5. In addition, InterProScan (v5.4.0) (Mulder and Apweiler, 2007) was used to screen proteins against five databases (Pfam, PRINTS, PROSITE, ProDom and SMART) to determine protein domains and motifs. Gene Ontology (GO) annotation of each gene was retrieved from the corresponding InterPro entry. In addition, KEGG annotation was performed using GhostKOALA (Kanehisa et al., 2016).

Phylogenomic analysis

Protein sequences of 14 metazoan species (*Adineta vaga*, *Echinococcus multilocularis*, *Aplysia californica*, *Lottia gigantea*, *Octopus bimaculoide*, *Phoronis australis*, *Lingula anatina*, *Notospermus geniculatus*, *Capitella teleta*, *Helobdella robusta*, *Eisenia Andrei*, *Riftia pachyptila*, *Paraescarpia echinospica*, *Lamellibrachia lumysi*) were downloaded for gene family cluster analysis (**Supplementary Table S8**). The longest transcripts of each gene (more than 30 amino acids) were retained. All-to-all BLASTP was used to identify the similarities between retained protein sequences of these 14 metazoan species and *R. piscesae* (E-value threshold: 1e-7). OrthoFinder (v2.2.7) (Emms and Kelly, 2019) was used to identify and cluster gene families among 15 species with default parameters. Gene clusters with >100 gene copies in one or more species were removed. Protein sequences of all single-copy gene families were retrieved and aligned using MAFFT (v7.271) (Katoh et al., 2002). The alignments were trimmed using TrimAI (v1.2) (Capella-Gutierrez et al., 2009). The phylogenetic tree was reconstructed with the trimmed alignments using FastTree2 (v2.1.11) (Price et al., 2010) with *Adineta vaga* as outgroup.

To estimate the divergent time, the trimmed alignments of single-copy orthologs among the 15 metazoan species were concatenated using PhyloSuite (v1.2.2) (Zhang et al., 2020). MCMCtree module of the PAML package (v4.9) (Yang, 2007) was used to estimate the divergent time with the concatenated alignment. The species tree of the 15 metazoan species was used as a guide tree, and the analysis was calibrated with the divergent time obtained from TimeTree database (minimum = 470.2 million years and soft maximum = 531.5 million years between *P. australis* and *L. anatina*) (Kumar et al., 2017) and previous analyses (minimum = 470.2 million years and soft maximum = 531.5 million years

between *A. californica* and *L. gigantea*; minimum = 532 million years and soft maximum = 549 million years for the first appearance of Mollusca; minimum = 476.3 million years and soft maximum = 550.9 million years for the appearance of capitellid-leech clade; minimum = 550.25 million years and soft maximum = 636.1 million years for the first appearance of Lophotrochozoa and Ecdysoa) (Donoghue et al., 2009; Benton et al., 2015; dos Reis et al., 2015).

Gene family expansion and contraction analysis

r8s (v1.7) was applied to obtain the ultrametric tree of 12 lophotrochozoan species (*C. teleta*, *H. robusta*, *E. andrei*, *L. gigantea*, *A. californica*, *N. geniculatus*, *A. californica*, *P. australis*, *R. pachyptila*, *P. echinospica*, *L. lumysi*, *R. piscesae*), which is calibrated with the divergent time between *C. teleta* and *L. gigantea* (688 mya) obtained from TimeTree database. CAFÉ (v5) (De Bie et al., 2006) was applied to determine the significance of gene-family expansion and contraction among 12 lophotrochozoan species based on the ultrametric tree and the gene clusters determined by OrthoFinder (v2.2.7). Gene families that were significantly expanded in each of four tubeworm species (*R. pachyptila*, *P. echinospica*, *L. lumysi*, *R. piscesae*) ($P < 0.05$) were annotated using PANTHER (v16.0) with the PANTHER HMM scoring tool (pantherScore2.pl) (Mi et al., 2017).

Homeobox gene analysis

Homeodomain sequences, which were retrieved from HomeoDB database(Zhong and Holland, 2011), were aligned to *R. piscesae* genome assembly using tbalstn. Sequences of the candidate homeobox genes were extracted based on the alignment results. The extracted sequences were aligned against NCBI NR and HomeoDB database to classify the homeobox genes.

Hemoglobin gene family analysis

Protein sequences of hemoglobin A1, A2, B1, B2 chains of four tubeworm species were obtained with reference references using DIAMOND BLASTP (Buchfink et al., 2021) with a E-value cut-off of 1e-5. The sequences were annotated in NCBI NR database using BLASTP. And protein domains in these sequences were annotated by Pfamscan against Pfam-A.hmm database (Finn et al., 2014). Sequences that have almost full length protein domains were aligned using MAFFT (v7.271) (Katoh and Standley, 2013). The alignments were trimmed using TrimAI (v1.2) (Capella-Gutierrez et al., 2009). The phylogenetic tree was reconstructed with the trimmed alignments using a maximum-likelihood method implemented in IQ-TREE2 (v2.1.2) (Minh et al., 2020). The best-fit substitution model was selected by using ModelFinder algorithm (Kalyaanamoorthy et al., 2017). Branch supports were assessed using the ultrafast bootstrap (UFBoot) approach with 1,000 replicates (Hoang et al., 2018).

Identification of positively selected genes (PSGs)

We identified PSGs in the *R. piscesae* genome within single-copy orthologs among 12 lophotrochozoan species that were identified in gene-family expansion and contraction analysis. Protein sequences of all single-copy gene families were retrieved and aligned using MAFFT (v7.271) (Katoh et al., 2002). Phylogenetic tree of each family was reconstructed using IQ-TREE2 (v2.1.2) (Minh et al., 2020). PSGs were identified based on the phylogenetic trees using HyPhy (v2.5.30) with the adaptive Branch-Site Random Effects Likelihood (aBSREL) model (Pond et al., 2020).

Data Availability

Raw reads and genome assembly are accessible in NCBI under BioProject number PRJNA826206. Assembled genome sequences are accessible under Whole Genome Shotgun project number JALOCR000000000. Raw reads and genome assembly are also available at the CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0002911.

Reference

Aas, P.A., Otterlei, M., Falnes, P.O., Vagbo, C.B., Skorpen, F., Akbari, M., et al. (2003). Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature* 421(6925)**,** 859-863. doi: 10.1038/nature01363.

Aggarwal, G., and Ramaswamy, R. (2002). Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J Biosci* 27(1 Suppl 1)**,** 7-14. doi: 10.1007/BF02703679.

Bailly, X., Jollivet, D., Vanin, S., Deutsch, J., Zal, F., Lallier, F., et al. (2002). Evolution of the sulfide-binding function within the globin multigenic family of the deep-sea hydrothermal vent tubeworm Riftia pachyptila. *Mol Biol Evol* 19(9)**,** 1421-1433. doi: 10.1093/oxfordjournals.molbev.a004205.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1)**,** 45-48. doi: 10.1093/nar/28.1.45.

Bakalenko, N.I., Novikova, E.L., Nesterenko, A.Y., and Kulakova, M.A. (2013). Hox gene expression during postlarval development of the polychaete Alitta virens. *Evodevo* 4(1)**,** 13. doi: 10.1186/2041-9139-4-13.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2)**,** 573-580. doi: 10.1093/nar/27.2.573.

Benton, M.J., Donoghue, P.C.J., Asher, R.J., Friedman, M., Near, T.J., and Vinther, J. (2015). Constraints on the timescale of animal evolutionary history. *Palaeontol Electron* 18(1).

Bergquist, D.C., Williams, F.M., and Fisher, C.R. (2000). Longevity record for deep-sea invertebrate. *Nature* 403(6769)**,** 499-500. doi: 10.1038/35000647.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res* 14(5)**,** 988-995. doi: 10.1101/gr.1865504.

Black, M.B., Halanych, K.M., Maas, P.A.Y., Hoeh, J., Hashimoto, D., Desbruyeres, D., et al. (1997). Molecular systematics of vestimentiferan tubworms from hydrothermal vents and cold-water seeps. *Mar. Biol.* 130**,** 141-149.

Bright, M., and Lallier, F.H. (2010). The Biology of Vestimentiferan Tubeworms. *Oceanogr Mar Biol* 48**,** 213-265. doi: Book_Doi 10.1201/Ebk1439821169.

Buchfink, B., Reuter, K., and Drost, H.G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18(4)**,** 366-368. doi: 10.1038/s41592-021-01101-x.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15)**,** 1972-1973. doi: 10.1093/bioinformatics/btp348.

Carney, S.L., Flores, J.F., Orobona, K.M., Butterfield, D.A., Fisher, C.R., and Schaeffer, S.W. (2007). Environmental differences in hemoglobin gene expression in the

hydrothermal vent tubeworm, Ridgeia piscesae. *Comp Biochem Physiol B Biochem Mol Biol* 146(3), 326-337. doi: 10.1016/j.cbpb.2006.11.002.

Cavalier-Smith, T. (1982). Skeletal DNA and the evolution of genome size. *Annu Rev Biophys Bioeng* 11, 273-302. doi: 10.1146/annurev.bb.11.060182.001421.

Cavanaugh, C.M., Gardiner, S.L., Jones, M.L., Jannasch, H.W., and Waterbury, J.B. (1981). Prokaryotic Cells in the Hydrothermal Vent Tube Worm Riftia pachyptila Jones: Possible Chemoautotrophic Symbionts. *Science* 213(4505), 340-342. doi: 10.1126/science.213.4505.340.

Childress, J.J., and Fisher, C.R. (1992). The biology of hydrothermal vent animals: physiology, biochemistry, and autotrophic symbioses. *Oceanogr Mar Biol* 30, 337-441.

Corliss, J.B., Dymond, J., Gordon, L.I., Edmond, J.M., von Herzen, R.P., Ballard, R.D., et al. (1979). Submerine Thermal Springs on the Galapagos Rift. *Science* 203(16), 1073-1083. doi: 10.1126/science.203.4385.107.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22(10), 1269-1271. doi: 10.1093/bioinformatics/btl097.

de Oliveira, A.L., Mitchell, J., Girguis, P., and Bright, M. (2022). Novel Insights on Obligate Symbiont Lifestyle and Adaptation to Chemosynthetic Environment as Revealed by the Giant Tubeworm Genome. *Mol Biol Evol* 39(1). doi: 10.1093/molbev/msab347.

Dick, G.J. (2019). The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. *Nat Rev Microbiol* 17(5), 271-283. doi: 10.1038/s41579-019-0160-2.

Dong, Q.Z., Wang, Y., Tang, Z.P., Fu, L., Li, Q.C., Wang, E.D., et al. (2013). Derlin-1 Is Overexpressed in Non-Small Cell Lung Cancer and Promotes Cancer Cell Invasion via EGFR-ERK-Mediated Up-Regulation of MMP-2 and MMP-9. *Am J Pathol* 182(3), 954-964. doi: 10.1016/j.ajpath.2012.11.019.

Donoghue, P., Benton, M., Yang, Z.H., and Inoue, J. (2009). Calibrating and Constraining the Molecular Clock. *J Vertebr Paleontol* 29, 89a-89a.

dos Reis, M., Thawornwattana, Y., Angelis, K., Telford, M.J., Donoghue, P.C., and Yang, Z. (2015). Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Curr Biol* 25(22), 2939-2950. doi: 10.1016/j.cub.2015.09.066.

Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics* 21 Suppl 1, i152-158. doi: 10.1093/bioinformatics/bti1003.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20(1), 238. doi: 10.1186/s13059-019-1832-y.

Eshraghi, A., Dixon, S.D., Tamilselvam, B., Kim, E.J., Gargi, A., Kulik, J.C., et al. (2014). Cytolethal distending toxins require components of the ER-associated degradation

pathway for host cell entry. *PLoS Pathog* 10(7), e1004295. doi: 10.1371/journal.ppat.1004295.

Finlin, B.S., Gau, C.L., Murphy, G.A., Shao, H., Kimel, T., Seitz, R.S., et al. (2001). RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. *J Biol Chem* 276(45), 42259-42267. doi: 10.1074/jbc.M105888200.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue), D222-230. doi: 10.1093/nar/gkt1223.

Fisher, C.R., Childress, J.J., Arp, A.J., Brooks, J.M., Distel, D., Favuzzi, J.A., et al. (1988). Physiology, morphology, and biochemical composition of Riftia pachyptila at Rose Garden in 1985. *Deep Sea Res. Part I Oceanogr. Res. Pap.* 35(10-11), 1745-1758.

Fisher, C.R., Urcuyo, I.A., Simpkins, M.A., and Nlx, E. (2008). Life in the Slow Lane: Growth and Longevity of Cold-seep Vestimentiferans. *Mar Ecol* 18(1), 83-94.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7), 644-U130. doi: 10.1038/nbt.1883.

Grassle, J.F. (1985). Hydrothermal vent animals: distribution and biology. *Science* 229(4715), 713-717. doi: 10.1126/science.229.4715.713.

Gregory, T.R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc* 76(1), 65-101. doi: 10.1017/s1464793100005595.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* 9(1). doi: 10.1186/gb-2008-9-1-r7.

Halanych, K.M. (2005). Molecular phylogeny of siboglinid annelids (a.k.a. pogonophorans): a review. *Hydrobiologia* 535, 297-307. doi: DOI 10.1007/s10750-004-1437-6.

Hand, S.C. (1987). Trophosome Ultrastructure and the Characterization of Isolated Bacteriocytes from Invertebrate-Sulfur Bacteria Symbioses. *Biol Bull* 173(1), 260-276. doi: 10.2307/1541878.

Ho, J.Y., Hsu, R.J., Liu, J.M., Chen, S.C., Liao, G.S., Gao, H.W., et al. (2017). MicroRNA-382-5p aggravates breast cancer progression by regulating the RERG/Ras/ERK signaling axis. *Oncotarget* 8(14), 22443-22459. doi: 10.18632/oncotarget.12338.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35(2), 518-522. doi: 10.1093/molbev/msx281.

Jones, M. (1987). On the status of the phylum-name, and other names, of the vestimentiferan tube worms. *Proc. Biol. Soc. Wash.* 100, 1049-1050.

660 Jones, M.L. (1981). Riftia pachyptila Jones: Observations on the Vestimentiferan Worm
661 from the Galapagos Rift. *Science* 213(4505), 333-336. doi:
662 10.1126/science.213.4505.333.

663 Julian, D., Gaill, F., Wood, E., Arp, A.J., and Fisher, C.R. (1999). Roots as a site of
664 hydrogen sulfide uptake in the hydrocarbon seep vestimentiferan Lamellibrachia
665 sp. *J Exp Biol* 202(17), 2245-2257.

666 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S.
667 (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*
668 *Methods* 14(6), 587-589. doi: 10.1038/nmeth.4285.

669 Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA:
670 KEGG Tools for Functional Characterization of Genome and Metagenome
671 Sequences. *J Mol Biol* 428(4), 726-731. doi: 10.1016/j.jmb.2015.11.006.

672 Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for
673 rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids*
674 *Res* 30(14), 3059-3066. doi: 10.1093/nar/gkf436.

675 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software
676 version 7: improvements in performance and usability. *Mol Biol Evol* 30(4), 772-
677 780. doi: 10.1093/molbev/mst010.

678 Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* 12(4), 656-664.
679 doi: 10.1101/gr.229202.

680 Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinform* 5, 59. doi: 10.1186/1471-
681 2105-5-59.

682 Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for
683 Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* 34(7), 1812-1819. doi:
684 10.1093/molbev/msx116.

685 Levin, L.A. (2005). Ecology of cold seep sediments: Interactions of fauna with flow,
686 chemistry and microbes. *Oceanogr Mar Biol* 43, 1-46. doi: DOI
687 10.1201/9781420037449.ch1.

688 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-
689 Wheeler transform. *Bioinformatics* 25(14), 1754-1760. doi:
690 10.1093/bioinformatics/btp324.

691 Li, Y., Tassia, M.G., Waits, D.S., Bogantes, V.E., David, K.T., and Halanych, K.M. (2019).
692 Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm
693 Lamellibrachia luymesi. *BMC Biol* 17(1), 91. doi: 10.1186/s12915-019-0713-x.

694 Little, C.T.S., and Vrijenhoek, R.C. (2003). Are hydrothermal vent animals living fossils?
695 *Trends Ecol Evol* 18(11), 582-588. doi: 10.1016/j.tree.2003.08.009.

696 Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an
697 empirically improved memory-efficient short-read de novo assembler. *Gigascience*
698 1(1), 18. doi: 10.1186/2047-217X-1-18.

699 Lutz, R.A., Desbruyeres, D., Shank, T.M., and Vrijenhoek, R.C. (1998). A deep-sea
700 hydrothermal vent community dominated by Stauromedusae. *Deep Sea Res. Part*
701 *II Top. Stud. Oceanogr.* 45**,** 329-334.

702 Lutz, R.A., and Kennish, M.J. (1993). Ecology of deep ‐ sea hydrothermal vent
703 communities: A review. *Rev. Geophys.* 31(3)**,** 211-242.

704 Lynch, M. (2007). *The Origins of Genome Architecture.* Sunderland, MA: Sinauer
705 Associates.

706 Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two
707 open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16)**,** 2878-2879.
708 doi: 10.1093/bioinformatics/bth315.

709 Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., et al. (2017).
710 PANTHER version 11: expanded annotation data from Gene Ontology and
711 Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*
712 45(D1)**,** D183-D189. doi: 10.1093/nar/gkw1138.

713 Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler,
714 A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
715 Inference in the Genomic Era. *Mol Biol Evol* 37(5)**,** 1530-1534. doi:
716 10.1093/molbev/msaa015.

717 Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence
718 classification and comparison. *Methods Mol Biol* 396**,** 59-70. doi: 10.1007/978-1-
719 59745-515-2_5.

720 Niu, S., Li, J., Bo, W., Yang, W., Zuccolo, A., Giacomello, S., et al. (2022). The Chinese
721 pine genome and methylome unveil key features of conifer evolution. *Cell* 185(1)**,**
722 204-217 e214. doi: 10.1016/j.cell.2021.12.006.

723 Osorio, F.G., Soria-Valles, C., Santiago-Fernandez, O., Bernal, T., Mittelbrunn, M.,
724 Colado, E., et al. (2016). Loss of the proteostasis factor AIRAPL causes myeloid
725 transformation by deregulating IGF-1 signaling. *Nat Med* 22(1)**,** 91-96. doi:
726 10.1038/nm.4013.

727 Parra, G., Blanco, E., and Guigo, R. (2000). GeneID in Drosophila. *Genome Res* 10(4)**,**
728 511-515. doi: 10.1101/gr.10.4.511.

729 Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next
730 generation sequencing data. *PLoS One* 7(2)**,** e30619. doi:
731 10.1371/journal.pone.0030619.

732 Paull, C.K., Hecker, B., Commeau, R., Freeman-Lynde, R.P., Neumann, C., Corso, W.P.,
733 et al. (1984). Biological communities at the Florida escarpment resemble
734 hydrothermal vent taxa. *Science* 226(4677)**,** 965-967. doi:
735 10.1126/science.226.4677.965.

736 Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions
737 during animal body patterning. *Nat Rev Genet* 6(12)**,** 893-904. doi:
738 10.1038/nrg1726.

739 Petersen, J.M., Zielinski, F.U., Pape, T., Seifert, R., Moraru, C., Amann, R., et al. (2011).
740 Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* 476(7359)**,**
741 176-180. doi: 10.1038/nature10325.

742 Pflugfelder, B., Cary, S.C., and Bright, M. (2009). Dynamics of cell proliferation and
743 apoptosis reflect different life strategies in hydrothermal vent and cold seep
744 vestimentiferan tubeworms. *Cell Tissue Res* 337(1)**,** 149-165. doi: 10.1007/s00441-
745 009-0811-0.

746 Pond, S.L.K., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., et al.
747 (2020). HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing
748 Using Phylogenies. *Mol Biol Evol* 37(1)**,** 295-299. doi: 10.1093/molbev/msz197.

749 Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat
750 families in large genomes. *Bioinformatics* 21 Suppl 1**,** i351-358. doi:
751 10.1093/bioinformatics/bti1018.

752 Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-
753 likelihood trees for large alignments. *PLoS One* 5(3)**,** e9490. doi:
754 10.1371/journal.pone.0009490.

755 Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and
756 Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 11(1)**,**
757 1432. doi: 10.1038/s41467-020-14998-3.

758 Sarrazin, J., Robigou, V., Juniper, S.K., and Delaney, J.R. (1998). Biological and
759 geological dynamics over four years on a high-temperature sulfide structure at the
760 Juan de Fuca Ridge hydrothermal observatory. *Mar Ecol Prog Ser* 153(1)**,** 5-24.
761 doi: 10.3354/meps153005.

762 Schulze, A. (2001). Comparative anatomy of excretory organs in vestimentiferan tube
763 worms (Pogonophora, Obturata). *J Morphol* 250(1)**,** 1-11. doi: DOI
764 10.1002/jmor.1054.

765 Shank, T.M., Fornari, D.J., Von Damm, K.L., Lilley, M.D., Haymon, R.M., and Lutz, R.A.
766 (1998). Temporal and spatial patterns of biological community development at
767 nascent deep-sea hydrothermal vents (9° 50′ N, East Pacific Rise). *Deep Sea Res.*
768 *Part II Top. Stud. Oceanogr.* 45(1-3)**,** 465-515.

769 Sharpton, T.J., Stajich, J.E., Rounsley, S.D., Gardner, M.J., Wortman, J.R., Jordar, V.S., et
770 al. (2009). Comparative genomic analyses of the human fungal pathogens
771 Coccidioides and their relatives. *Genome Res* 19(10)**,** 1722-1731. doi:
772 10.1101/gr.087551.108.

773 Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M.
774 (2015). BUSCO: assessing genome assembly and annotation completeness with
775 single-copy orthologs. *Bioinformatics* 31(19)**,** 3210-3212. doi:
776 10.1093/bioinformatics/btv351.

777 Southward, E.C., Tunnacliffe, V., and Black, M. (1995). Revision of the species of Ridgeia
778 from northeast Pacific hydrothermal vents, with a redescription of Ridgeia piscesae
779 Jones (Pogonophora: Obturata=Vestimentifera). *Can. J. Zool.* 73**,** 282-295. doi:
780 10.1139/z95-033.

781 Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in
782 eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33**,** W465-W467.
783 doi: 10.1093/nar/gki458.

784 Stewart, F.J., Newton, I.L., and Cavanaugh, C.M. (2005). Chemosynthetic endosymbioses:
785 adaptations to oxic-anoxic interfaces. *Trends Microbiol* 13(9)**,** 439-448. doi:
786 10.1016/j.tim.2005.07.007.

787 Suess, E. (2014). Marine cold seeps and their manifestations: geological control,
788 biogeochemical criteria and environmental conditions. *Int J Earth Sci* 103(7)**,** 1889-
789 1916. doi: 10.1007/s00531-014-1010-0.

790 Sun, Y., Sun, J., Yang, Y., Lan, Y., Ip, J.C., Wong, W.C., et al. (2021). Genomic signatures
791 supporting the symbiosis and formation of chitinous tube in the deep-sea tubeworm
792 Paraescarpia echinospica. *Mol Biol Evol*. doi: 10.1093/molbev/msab203.

793 Tenaillon, M.I., Manicacci, D., Nicolas, S.D., Tardieu, F., and Welcker, C. (2016). Testing
794 the link between genome size and growth rate in maize. *PeerJ* 4**,** e2408. doi:
795 10.7717/peerj.2408.

796 Thiebaut, E., Huther, X., Shillito, B., Jollivet, D., and Gaill, F. (2002). Spatial and temporal
797 variations of recruitment in the tube worm Riftia pachyptila on the East Pacific Rise
798 (9 degrees 50 ' N and 13 degrees N). *Mar Ecol Prog Ser* 234**,** 147-157. doi: DOI
799 10.3354/meps234147.

800 Tivey, M.K., Stakes, D.S., Cook, T.L., Hannington, M.D., and Petersen, S. (1999). A
801 model for growth of steep-sided vent structures on the Endeavour Segment of the
802 Juan de Fuca Ridge: Results of a petrologic and geochemical study. *J Geophys Res-*
803 *Sol Ea* 104(B10)**,** 22859-22883. doi: Doi 10.1029/1999jb900107.

804 Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions
805 with RNA-Seq. *Bioinformatics* 25(9)**,** 1105-1111. doi:
806 10.1093/bioinformatics/btp120.

807 Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., et al.
808 (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated
809 transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5)**,**
810 511-515. doi: 10.1038/nbt.1621.

811 Tsurumi, M., and Tunnicliffe, V. (2003). Tubeworm-associated communities at
812 hydrothermal vents on the Juan de Fuca Ridge, northeast Pacific. *Deep Sea Res.*
813 *Part I Oceanogr. Res. Pap.* 50(5)**,** 611-629. doi: 10.1016/S0967-0637(03)00039-6.

814 Tunnicliffe, V. (1992). The nature and origin of the modern hydrothermal vent fauna.
815 *Palaios* 7(4)**,** 338-350. doi: 10.2307/3514820.

816 Tunnicliffe, V., Embley, R.W., Holden, J.F., Butterfield, D.A., Massoth, G., and Juniper,
817 S.K. (1997). Biological colonization of new hydrothermal vents following an
818 eruption on Juan de Fuca Ridge. *Deep Sea Res. Part I Oceanogr. Res. Pap.* 44(9-
819 10)**,** 1627-1644.

820 Tunnicliffe, V., McArthur, A.G., and McHugh, D. (1998). A Biogeographical Perspective
821 of the Deep-sea Hydrothermal Vent Fauna. *Adv. Mar. Biol.* 34**,** 353-442.

822 Urcuyo, I.A., Bergquist, D.C., MacDonald, I.R., VanHorn, M., and Fisher, C.R. (2007).
823 Growth and longevity of the tubeworm Ridgeia piscesae in the variable diffuse flow
824 habitats of the Juan de Fuca Ridge. *Mar Ecol Prog Ser* 344**,** 143-157. doi:
825 10.3354/meps06710.

826 Urcuyo, I.A., Massoth, G.J., Julian, D., and Fisher, C.R. (2003). Habitat, growth and
827 physiological ecology of a basaltic community of Ridgeia piscesae from the Juan
828 de Fuca Ridge. *Deep Sea Res. Part I Oceanogr. Res. Pap.* 50(6)**,** 763-780. doi:
829 10.1016/S0967-0637(03)00061-X.

830 Van Nieuwerburgh, F., Thompson, R.C., Ledesma, J., Deforce, D., Gaasterland, T.,
831 Ordoukhanian, P., et al. (2012). Illumina mate-paired DNA sequencing-library
832 preparation using Cre-Lox recombination. *Nucleic Acids Res* 40(3). doi:
833 10.1093/nar/gkr1000.

834 Vanreusel, A., Andersen, A.C., Boetius, A., Connelly, D., Cunha, M.R., Decker, C., et al.
835 (2009). Biodiversity of cold seep ecosystems along the European margins.
836 *Oceanography* 22(1)**,** 110-127. doi: 10.5670/oceanog.2009.12.

837 Vinogradov, A.E. (1995). Nucleotypic Effect in Homeotherms: Body-Mass-Corrected
838 Basal Metabolic Rate of Mammals Is Related to Genome Size. *Evolution* 49(6)**,**
839 1249-1259. doi: 10.1111/j.1558-5646.1995.tb04451.x.

840 Von Damm, K.L. (1995). "Controls on the Chemistry and Temporal Variability of Seafloor
841 Hydrothermal Fluids.," in *Seafloor Hydrothermal System: Physical, Chemical,*
842 *Biological, and Geological Interactions,* eds. R.A. Humphris, L.S. Zierenberg &
843 R.E. Thomson.), 222-247.

844 Vrijenhoek, R.C. (2010). Genetic diversity and connectivity of deep-sea hydrothermal vent
845 metapopulations. *Mol Ecol* 19(20)**,** 4391-4411. doi: 10.1111/j.1365-
846 294X.2010.04789.x.

847 Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., et al. (2017). Scallop genome
848 provides insights into evolution of bilaterian karyotype and development. *Nat Ecol*
849 *Evol* 1(5)**,** 120. doi: 10.1038/s41559-017-0120.

850 Wilson, D.L., Beharry, A.A., Srivastava, A., O'Connor, T.R., and Kool, E.T. (2018).
851 Fluorescence Probes for ALKBH2 Allow the Measurement of DNA Alkylation
852 Repair and Drug Resistance Responses. *Angew Chem Int Ed Engl* 57(39)**,** 12896-
853 12900. doi: 10.1002/anie.201807593.

854 Wright, N.A., Gregory, T.R., and Witt, C.C. (2014). Metabolic 'engines' of flight drive
855 genome size reduction in birds. *Proc Biol Sci* 281(1779)**,** 20132780. doi:
856 10.1098/rspb.2013.2780.

857 Wyngaard, G.A., Rasch, E.M., Manning, N.M., Gasser, K., and Domangue, R. (2005). The
858 relationship between genome size, development rate, and body size in copepods.
859 *Hydrobiologia* 532**,** 123-137. doi: 10.1007/s10750-004-9521-5.

860 Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-
861 length LTR retrotransposons. *Nucleic Acids Res* 35**,** W265-W268. doi:
862 10.1093/nar/gkm286.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8), 1586-1591. doi: 10.1093/molbev/msm088.

Zal, F., Suzuki, T., Kawasaki, Y., Childress, J.J., Lallier, F.H., and Toulmond, A. (1997). Primary structure of the common polypeptide chain b from the multi-hemoglobin system of the hydrothermal vent tube worm Riftia pachyptila: an insight on the sulfide binding-site. *Proteins* 29(4), 562-574. doi: 10.1002/(SICI)1097-0134(199712)29:4<562::AID-PROT15>3.0.CO;2-K.

Zhang, D., Gao, F., Jakovlic, I., Zou, H., Zhang, J., Li, W.X., et al. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol Ecol Resour* 20(1), 348-355. doi: 10.1111/1755-0998.13096.

Zhong, Y.F., and Holland, P.W. (2011). HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol Dev* 13(6), 567-568. doi: 10.1111/j.1525-142X.2011.00513.x.

878 ## Acknowledgements

879 This work was supported by National Key R&D Program of China (2018YFC0310702),
880 the Projects under Major State Basic Research Development Program of China (973
881 Program) (2015CB755906), National Natural Science Foundation of China (No.
882 31900309), GuangDong Basic and Applied Basic Research Foundation (No.
883 2019A1515011644), Innovation Group Project of Southern Marine Science and
884 Engineering Guangdong Laboratory (Zhuhai) (No. 311021006), and Project 2018N2001
885 from Department of Fujian Science and Technology and Program for Innovative Research
886 Team in Science and Technology in Fujian Province University. The funders had no role
887 in study design, data collection and analysis, decision to publish, or preparation of the paper.
888 We gratefully acknowledge the National Supercomputing Center in Guangzhou for
889 provision of computational resources.

890

891 ## Author contributions

892 J.M.C, C.J.G and J.G.H. conceived of the project and designed research; L.R., M.L., Z.L,
893 Y.W. assembled and annotated the genome; M.W, J.H., L.Z., H.S., M.C., Y.J., F.Y., and
894 R.Z. performed the evolutionary analyses; M.W., J.M.C., C.J.G, and J.G.H. wrote the paper
895 with contribution from all authors.

896

897 ## Ethics declarations

898 ### Conflict of interest

899 The authors declare that they have no conflict of interest or competing interests.

900

Animal and human rights statement

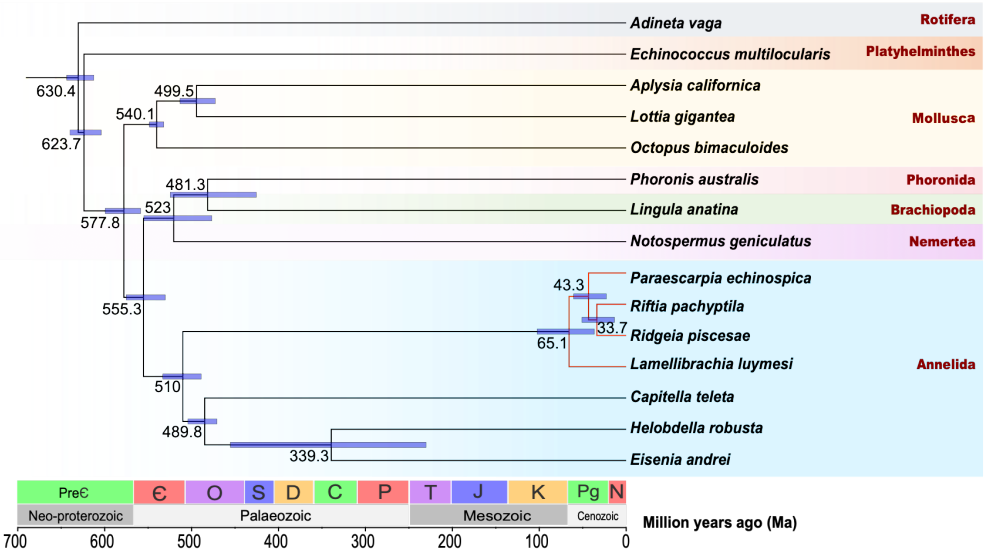
The tubeworm used in our study is an invertebrate, so the approval according to the regulations on the use of tubeworm is unnecessary.

905 **Table 1** Genome assembly statistics of four deep-sea vestimentiferan tubeworms

	Ridgeia piscesae	*Riftia pachyptila* (de Oliveira et al., 2022)	*Paraescarpia echinospica* (Sun et al., 2021)	*Lamellibrachia luymesi* (Li et al., 2019)
Assembled genome size (Mb)	574.9	560.7	1090.9	687.7
Contig N50 (kb)	10.4	2870.3	253.6	24
Scaffold N50 (kb)	230.2	-	67,235.3	372.9
BUSCO (%)	92.6	99.4	95.1	95.8
Repeat content (%)	30.2	29.9	55.1	38.2

906

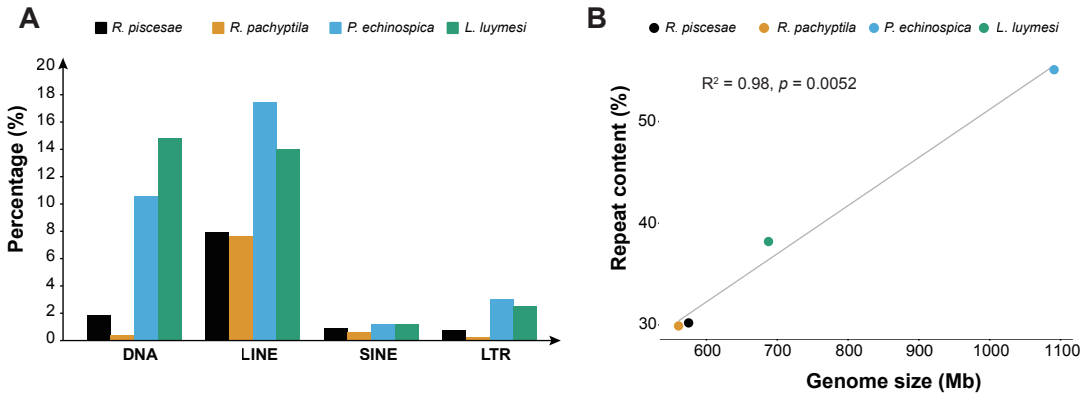
907 # Figures



908

909 **Figure 1 A species tree of *R. piscesae* and 14 metazoans.** Single-copy orthologs were
910 used to reconstruct the phylogenetic tree. The divergence time between species pairs was
911 listed above each node, and 95% confidence interval of the estimated divergence time was
912 denoted as blue bar. *R. piscesae* diverged from *R. pachyptila* approximately 33.7 million
913 years ago.

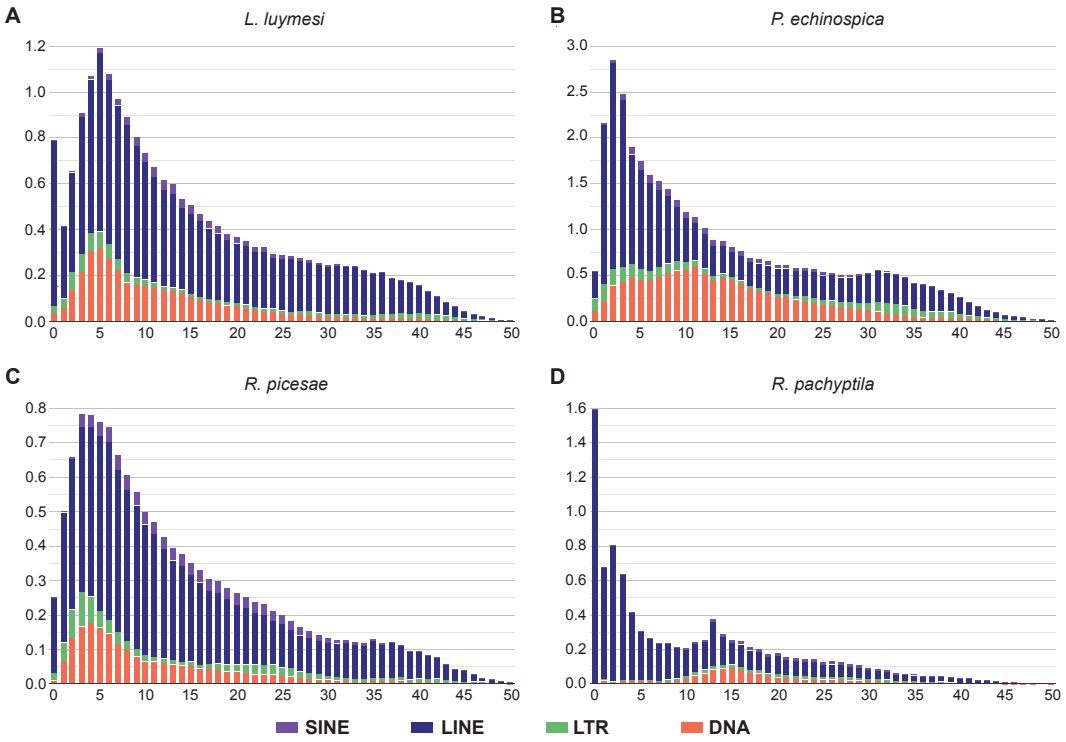
914



915

916 **Figure 2 Genome size evolution in four vestimentiferan tubeworms.** (**A**) Comparison
917 of the occurrence and composition of repetitive elements in the genomes of 4
918 vestimentiferan tubeworm. (**B**) The relationship between repeat content and genome size
919 in 4 vestimentiferan tubeworm. A strong positive correlation ($R^2 = 0.98$, $P = 0.0052$) was
920 identified between genome size and repeat content in these four species.

921



922

923 **Figure 3 Transposable element-accumulation profile in the genomes of four**
924 **vestimentiferan tubeworms.** There are recent expansions of TEs in the genomes of *L.*
925 *luymesi*, *P. echinospica*, and *R. piscesae*, but not in *R. pachyptila*.

926

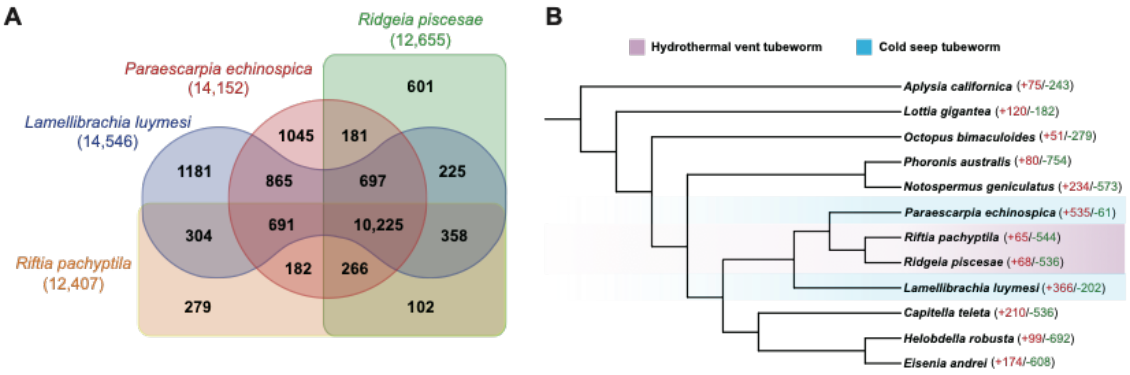
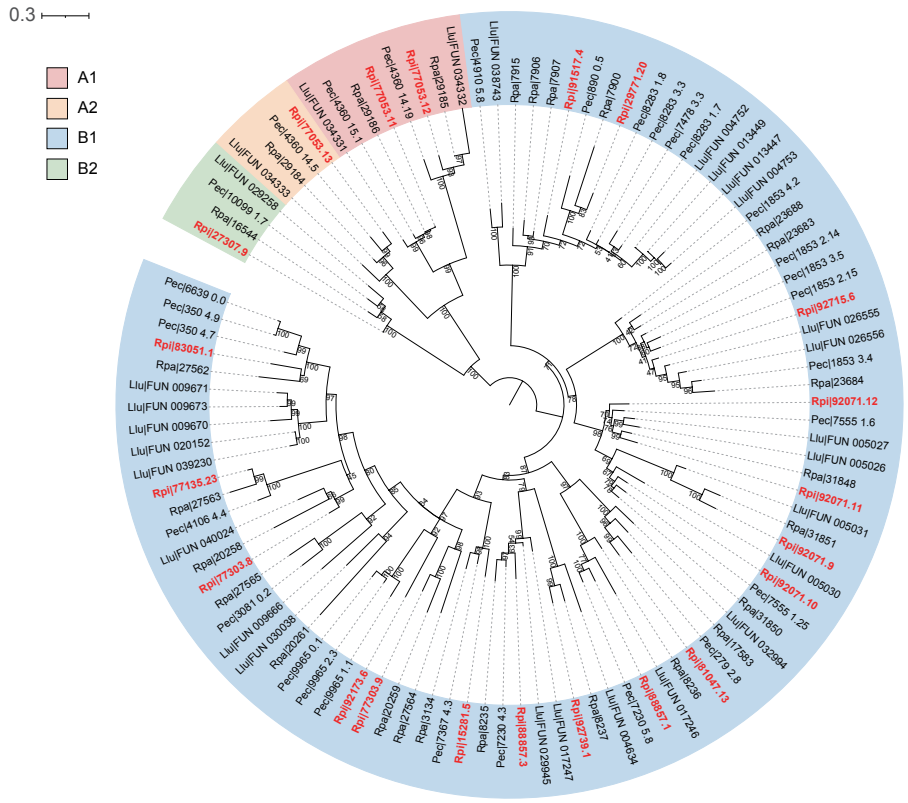


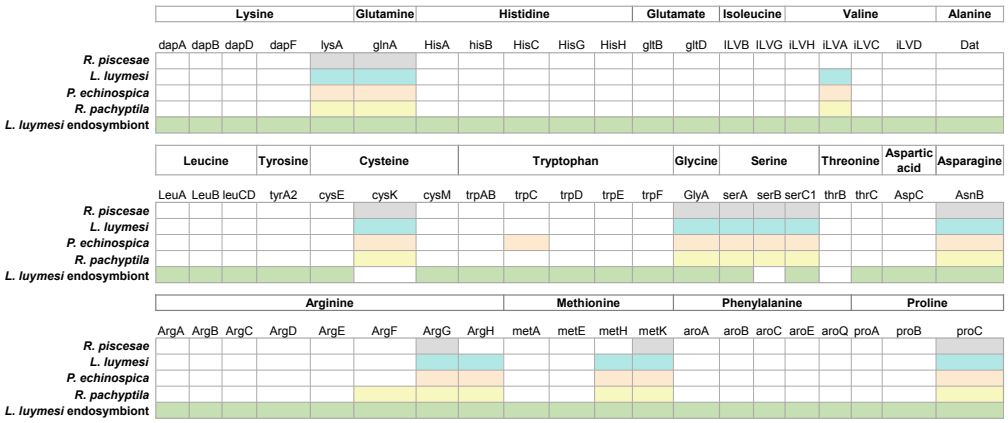
Figure 4 Protein family evolution in four vestimentiferan tubeworms. (**A**) Venn diagram of shared and unique gene families in four vestimentiferan tubeworm species. Lineage-specific gene families of *R. piscesae* and *R. pachyptila* are much less than the ones of *L. luymesi* and *P. echinospica*. (**B**) Gene family expansion/contraction analysis of 4 vestimentiferan tubeworms and 8 other lophotrochozoans.



934

935 **Figure 5 Gene tree of Hb subunits A1, A2, B1, and B2 from four vestimentiferan**
936 **tubeworms.** The values near the nodes are ultrafast bootstrap (UFBoot) values. *R. piscesae*
937 sequences are labeled red.

938



939

940 **Figure 6 Amino acid biosynthesis genes.** Most key genes associated with amino acid
941 biosynthesis are missing in the genomes of *L. luymesi*, *P. echinospica*, *R. pachyptila*, and
942 *R. piscesae*. These genes are presented in the genome of *L. luymesi* symbionts.