

Genomic Patterns of Divergence in the Early and Late Steps of Speciation of the Deep-Sea Vent Thermophilic Worms of the Genus *Alvinella*

Camille Thomas–Bulle (✉ camille.thomas-bulle@umt.edu)

Sorbonne Université, CNRS, UMR, Station Biologique de Roscoff

Denis Bertrand

Genome Institute of Singapore

Niranjan Nagarajan

Genome Institute of Singapore

Richard Copley

Villefranche Oceanographic Laboratory

Erwan Corre

CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680

Stéphane Hourdez

Observatoire Océanologique de Banyuls-sur-Mer

Eric Bonnivard

Sorbonne Université, CNRS, UMR, Station Biologique de Roscoff

Adam Claridge-Chang

Institute of Molecular and Cell Biology

Didier Jollivet

Sorbonne Université, CNRS, UMR, Station Biologique de Roscoff

Research Article

Keywords: Divergence, genome architecture, speciation, ecological species, habitat specialization, selection, hydrothermal vents

Posted Date: March 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1405281/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Genomic patterns of divergence in the early and late steps of speciation of the**
2 **deep-sea vent thermophilic worms of the genus *Alvinella*.**

3
4 Camille Thomas–Bulle^{1,2*}, Denis Bertrand³, Niranjana Nagarajan^{3,4}, Richard Copley⁵, Erwan
5 Corre⁶, Stéphane Hourdez⁷, Éric Bonnivard¹, Adam Claridge-Chang⁸, Didier Jollivet¹

6
7 ¹ Sorbonne Université, CNRS, UMR 7144 AD2M, Station Biologique de Roscoff, Place Georges
8 Teissier CS90074, 29688 Roscoff, France.

9
10 ² Emlen Lab, Division of Biological Sciences, University of Montana, Missoula, MT, USA

11
12 ³ Genome Institute of Singapore, 60 Biopolis Street, #02-01 Genome, Singapore 138672,
13 Singapore.

14
15 ⁴ National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077, Singapore.

16
17 ⁵ Sorbonne Université, CNRS, UMR 7009 LBDV, Laboratoire ‘Evolution des génomes et des
18 protéines animales’, Observatoire océanologique, Quai de la Darse, 06234, Villefranche/mer,
19 cedex, France.

20
21 ⁶ CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680,
22 France.

23
24 ⁷ Sorbonne Université, CNRS UMR 8222, Laboratoire d’Ecogéochimie des Environnements
25 Benthiques (LECOB), Observatoire Océanologique de Banyuls, Avenue du Fontaulé, 66650
26 Banyuls-sur-mer.

27
28 ⁸ Duke-NUS Medical School, Institute of Molecular and Cell Biology, 61 Biopolis Drive 8-05B,
29 138673 Singapore.

30
31 *Corresponding author: Camille Thomas-Bulle: camille.thomas-bulle@mso.umt.edu

32
33 **Keywords**

34 Divergence, genome architecture, speciation, ecological species, habitat specialization, selection,
35 hydrothermal vents

36

37 **Abstract**

38

39 **Background:** The transient and fragmented nature of the deep-sea hydrothermal environment
40 made of ridge subduction, plate collision and the emergence of new rifts is currently acting to the
41 separation of vent populations, local adaptation and is likely to promote bursts of speciation and
42 species specialization. The tube-dwelling worms *Alvinella pompejana* called the Pompeii worm
43 and its sister species *A. caudata* live syntopically on the hottest part of deep-sea hydrothermal
44 chimneys along the East Pacific Rise. They are exposed to extreme thermal and chemical gradients,
45 which vary greatly in space and time, and thus represent ideal candidates for understanding the
46 evolutionary mechanisms at play in the vent fauna evolution.

47 **Results:** In the present study, we explored genomic patterns of divergence in the early and late
48 steps of speciation of these emblematic worms using transcriptome assemblies and the first draft
49 genome to better understand the relative role of geographic isolation and habitat preference in their
50 genome evolution. Analyses were conducted on allopatric populations of *Alvinella pompejana*
51 (early stage of separation) and between *A. pompejana* and its syntopic species *Alvinella caudata*
52 (late stage of speciation). We first identified divergent genomic regions and targets of selection as
53 well as their position in the genome over collections of orthologous genes and, then, described the
54 speciation dynamics by documenting the annotation of the most divergent and/or positively
55 selected genes involved in the isolation process. Gene mapping clearly indicated that divergent
56 genes associated with the early stage of speciation, although accounting for nearly 30% of genes,
57 are highly scattered in the genome without any island of divergence and not involved in gamete
58 recognition or mito-nuclear incompatibilities. By contrast, genomes of *A. pompejana* and *A.*
59 *caudata* are clearly separated with nearly all genes (96%) exhibiting high divergence. This

60 congealing effect however seems to be linked to habitat specialization and still allows positive
61 selection on genes involved in gamete recognition, as a possible long-duration process of species
62 reinforcement.

63 **Conclusion:** Our analyses pointed out the non-negligible role of natural selection on both the early
64 and late stages of speciation in the emblematic thermophilic worms living on the walls of deep-
65 sea hydrothermal chimneys. They shed light on the evolution of gene divergence during the
66 process of speciation and species specialization over a very long period of time.

67

68 **Background**

69

70 The transient and fragmented nature of the deep-sea hydrothermal environment made of
71 ridge subduction, plate collision and the emergence of new rift systems likely led to the separation
72 of vent communities at a large spatial scale with transient events of spatial isolation and bursts of
73 speciation [1-4]. Most species inhabiting these unstable and harsh environments often display a
74 rapid growth, a fast maturation, a high investment in reproduction relative to survival and good
75 dispersal abilities enabling the colonization of vent emissions as soon as they appear [5-6]. Many
76 studies notably focused on how this highly specialized and endemic fauna disperse and colonize
77 new territories in the face of habitat fragmentation [7-8]. Given the dynamics of the vents
78 associated with the tectonic activity along oceanic ridges and its effect on the metapopulation
79 dynamics [9], such environment appears ideal to study the mechanisms by which speciation occurs
80 and to tease apart the relative role of geography and local adaptation to this extent. On one hand,
81 the vent environment is highly variable with abrupt chemical gradients leading to a mosaic of
82 habitats where species are spatially and temporally partitioned [10-12]. This may likely promote

83 ecological speciation when migration is not able to counter-balance local selection at some specific
84 locations. On the other hand, populations are connected within a linear network of habitats and
85 tectonic plates rearrangements are likely to result in either species spatial isolation or secondary
86 contact zones where hybridization can occur [3, 13-16]. In addition to isolation by distance
87 between vent populations, patterns of genetic differentiation are mainly explained by the presence
88 of physical barriers to dispersal such as transform faults, microplates, triple junction points or
89 zones of weaker hydrothermal activity [3,15-16]. These relatively dynamic geological features
90 may produce vicariant events in the faunistic composition of vent communities [17] that often
91 coincide with the appearance of barriers to gene flow and hybrid zones [14,16,18].

92 Early incipient species often represent groups of individuals that can still interbreed and
93 exchange genes despite the emergence of pre- or post-zygotic barriers to gene flow. During the
94 process of speciation and even at the latest steps of reinforcement, emerging species are thus not a
95 collection of impermeable taxonomic units with distinct morphologies but rather dynamic entities
96 gradually heightening their reproductive barriers in a more or less continuous way over long
97 periods of time [19]. This process depends on the divergence history of the populations and their
98 associated ecological changes. But it is worth mentioning that exceptions where speciation
99 happens relatively fast also exist, for example in the case of genomes under disruptive selective
100 conditions following changes in ploidy [20-21] or rapid shifts in reproductive timing/behaviour
101 due to the colonization of new territories/habitats [22-23]. In general, pre- and post-zygotic
102 mechanisms leading to reproductive isolation result in a gradual accumulation of genetic barriers
103 during the separation of groups of individuals in either space or time [24-31]. The point at which
104 speciation is initiated and then considered complete is however vague and varies with the species
105 concept used (*e.g.*, complete reproductive isolation, hybrid counter-selection, phylogenetic

106 isolation with complete sorting of allelic lineages between populations). In the initial part of the
107 speciation process, individuals are likely to still exchange genes across semi-permeable barriers
108 resulting in heterogeneous migration rates along the genome but could also endure local adaptation
109 resulting in a loss of genetic diversity over some portions of their interacting genomes [32].

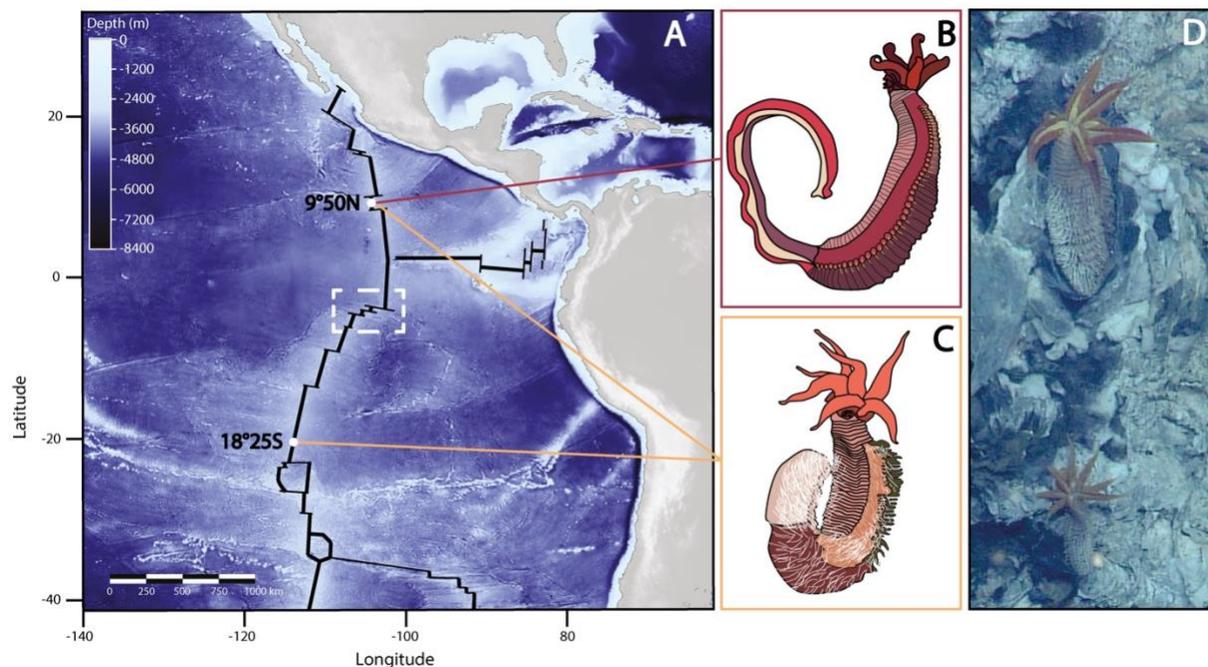
110 Recent studies on the genetic architecture of reproductive isolation across the entire
111 genome often found a genome-wide heterogeneity of genetic differentiation between populations,
112 ecotypes or even well-separated morphological species, with highly divergent regions of the
113 genome contrasting with regions of lower divergence rather than a homogeneous rate of
114 differentiation across the genome [30, 33-34]. To this extent, genome scans of differentiation and
115 genome-wide linkage maps based on polymorphic sites at multiple loci across the genome,
116 represent powerful tools to address the genomic landscape of population/species isolation. Regions
117 displaying high levels of differentiation between lineages, so-called “genomic islands of
118 speciation” were logically assumed to exhibit loci underlying reproductive isolation when the time
119 elapsed since isolation was long enough to create genetic incompatibilities [35]. Most of the time,
120 “islands of speciation” are believed to result from local barriers to gene flow at some specific genes
121 that rapidly led to allele fixation, gene hitchhiking and the subsequent accumulation of divergence
122 [36]. These islands differ from incidental islands of divergence resulting from accelerated rates of
123 lineage sorting within populations due to recurrent events of either selective sweeps or background
124 selection not necessarily related to reproductive isolation [37]. The genomic island metaphor has
125 proven popular and a wide array of studies searching for “islands of speciation” in multiple taxa
126 has been published over the last decade [30,34,35,38-42]. These studies identified outlier loci or
127 regions that stand out from the distribution of the multi-loci genetic differentiation (most often
128 through F_{ST} scans) expected under models of migration with or without isolation and

129 heterogeneous rates of migration across genomes (*e.g.*, assortative mating, hybrid counter-
130 selection, divergent selection). Such outlier genes usually fall into three categories: (1) outliers
131 resulting from a differential effect of purifying selection in non-crossover regions of the separating
132 genomes, (2) outliers involved in Dobzhansky-Muller incompatibilities that affect the fitness of
133 hybrids/heterozygotes, and (3) outliers resulting from differential local adaptation between the
134 splitting lineages. In the last category, a limited number of alleles involved in local adaptation have
135 been found to carry a selective value strong enough to initiate a barrier to gene flow primarily by
136 divergent selection. When considering speciation from both ends of the road, how do these patterns
137 of divergence differ between the early and the late stages of isolation?

138 To better understand the role of genomic architecture on speciation, the number, width, and
139 level of clustering of divergent regions within the genome of early incipient species are usually
140 compared with more divergent congeneric pairs. It is expected that early in speciation, regions of
141 divergence should be small and randomly scattered in the genome and possibly restricted to a few
142 non-recombining regions. These regions may correspond to the accumulation of genetic
143 incompatibilities or to loci under diversifying selection strong enough to overcome gene flow
144 which would be expected to homogenize the rest of the genome. Conversely, later in speciation,
145 when gene flow is greatly reduced between the two genomes by the cumulated effect of multiple
146 genetic barriers, divergent selection, physical linkage between targets of selection and
147 chromosomal rearrangements may create longer genomic regions of divergence. Such processes
148 acting over larger regions of the genome, followed by gene specialization within each genome in
149 response to local adaptation contribute to a genome-wide “congealing” (GWC) effect that finalizes
150 speciation with reproductive isolation [43-46]. GWC refers to the period along the speciation
151 continuum when the genomes of the splitting populations essentially become sealed and when

152 reproductive isolation becomes a property of the whole genome not just of a few loci [43].
153 Therefore, incomplete lineage sorting and semi-permeable barriers to gene flow (associated with
154 the early stages of speciation) and the wide synergetic effects of genome congealing due to
155 chromosomal linkage between diverging genes (associated with the latest stages) are likely to
156 result in very distinct patterns of divergence at a given stage of the speciation process. Moreover,
157 positive selection over the genome due to gene specialization under local conditions should affect
158 KEGG pathways differently over the course of the splitting process. Although fundamental to our
159 understanding of speciation, such comparisons between the early and late stages of speciation, and
160 the time of transition between these two stages have rarely been done when looking at the
161 relationship between the number of diverging genes in association (congealing effect) and their
162 accumulated divergence (time since species separation).

163



164

Figure 1: A) Bathymetric map of the East Pacific Ridge (in black), sampled populations sites along the ridge (white dots) and the geographical barrier between Southern and Northern populations (white dashed square); B) Drawing of a specimen of *Alvinella caudata*; C) Drawing of a specimen of *Alvinella pompejana*; D) *A. pompejana*, populations in situ (East Pacific Rise: 13°N, 2630 m); PHARE cruise ©Ifremer

165 In the present study, we explored genomic patterns of divergence in the early and late steps
166 of speciation of the emblematic *Alvinella* worms colonizing the hottest part of deep-sea
167 hydrothermal chimneys along the East Pacific Rise (EPR) (Figure 1) [47-48]. Using transcriptome
168 assemblies, we carefully examined the relative role of geographic isolation and habitat preference
169 in shaping gene divergence between newly isolated populations (in allopatry) of *Alvinella*
170 *pompejana* and between this latter species and its syntopic species *Alvinella caudata* (niche
171 specialization & reinforcement in sympatry). The tube-dwelling worms *Alvinella pompejana*
172 Desbruyères & Laubier 1980 [49] (Figure 1C and 1D) called the Pompeii worm and its sister
173 species *A. caudata* Desbruyères & Laubier 1986 [50] (Figure 1B) live on the wall of still-hot
174 chimneys of the EPR from 23°N (Guaymas basin) to 38°S (Pacific-Antarctic Ridge). The first
175 description of these two worms considered them as two morphologically distinct ontogenic forms
176 of a single species, one juvenile form with a long tail bearing filamentous campylobacteria on
177 modified notopodia (corresponding to *A. caudata*) and the adult reproductive form without tail but
178 bearing the same epibionts on the whole dorsal part of its body (corresponding to *A. pompejana*).
179 But enzyme polymorphism analyses led to the subdivision of this taxon into two distinct species
180 without any shared electromorphs between them [51-52]. For both species, a strong physical
181 barrier to gene flow was depicted at the Equator near 7°S/EPR and separates Northern and
182 Southern vent populations in two allopatric groups of putative cryptic species with differentiation
183 particularly marked in the Pompeii worm *Alvinella pompejana* [3,15]. Differences in the vent-site
184 turn-over from both sides of the EPR and its subsequent effect on local hydrothermal vent
185 conditions (*i.e.* changes in thermal conditions due to a highest proportion of still-hot chimneys and
186 fluid chemistry heterogeneity at both local and regional scales) could represent favorable
187 conditions for local adaptation and subsequent ecological filtering when migration is low or

188 episodic across the barrier. In addition, depth gradient along the ridge or between ridges, extreme
189 variations in the hydrothermal fluid composition and subsequent effects on symbiotic associations
190 with divergent bacterial strains are likely to induce divergent selection between vent populations
191 [53].

192 To compare both ends of the speciation process, we performed the first genome scans of
193 both synonymous and non-synonymous gene divergences using transcriptome assemblies of the
194 two allopatric forms of *A. pompejana* that have recently separated about 1.5 Mya (early stage of
195 speciation), and that of its co-occurring sister species *A. caudata*, known to exhibit almost
196 complete reproductive isolation with *A. pompejana* (late stage of speciation). We then identified
197 divergent genomic regions and targets of selection as well as their position in the genome over
198 collections of orthologous genes and, thus, described the speciation dynamics by documenting the
199 annotation of the most divergent and/or positively selected genes involved in the species isolation.

200

201 **Results**

202

203 **Assembly and annotation of the reference genome and associated transcriptomes**

204 The assembled genome of *A. pompejana* consisted in 3,044 scaffolds for a total size of
205 245Mb, which represents 61% of the total genome size (400 Mb). The mapping of both divergence
206 and associated d_N/d_S along scaffolds was however only performed on the 113 longest scaffolds (>
207 300,000 bp) to maximize the number of genes per scaffold.

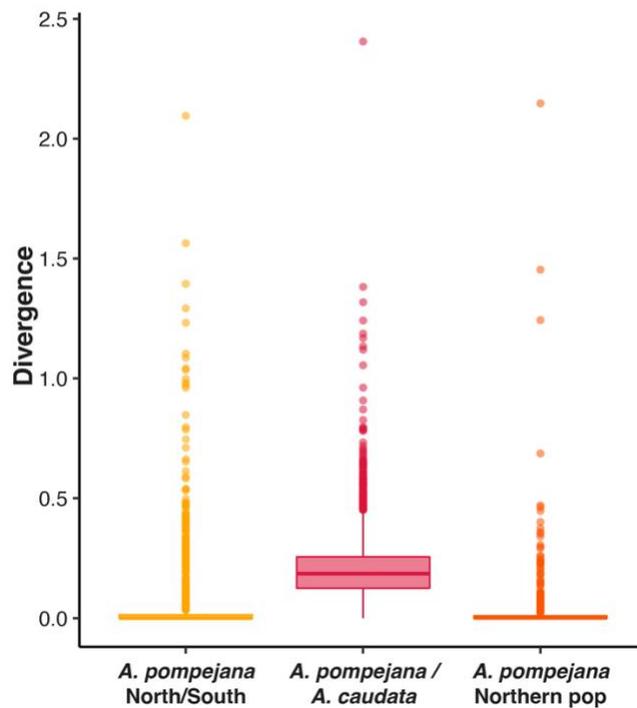
208

209 **Identification of orthologous genes**

210 Using the genome of *A. pompejana* as a reference database, we analyzed 94,006 pairwise
211 Blastn alignments between the Northern and Southern transcriptomes of *A. pompejana* and 68,963
212 pairwise Blastn alignments between the transcriptomes of *A. pompejana* and *A. caudata*. As these
213 putative genes contain paralogs, tandemly repeated paralogs, allelic forms of a gene in either genes
214 with or without introns, we filtered our datasets to only keep a subset of the most likely orthologous
215 genes. This reduces the number of pairwise alignments to 22,558 and 48,134 respectively.
216 Determination and translation of the CDS with a minimum threshold length of 300 bp and
217 sequence-similarity search against the Uniprot database reduced the two final datasets to 6,916
218 and 6,687 pairwise alignments of orthologous coding sequences on which further analyses were
219 performed.

220

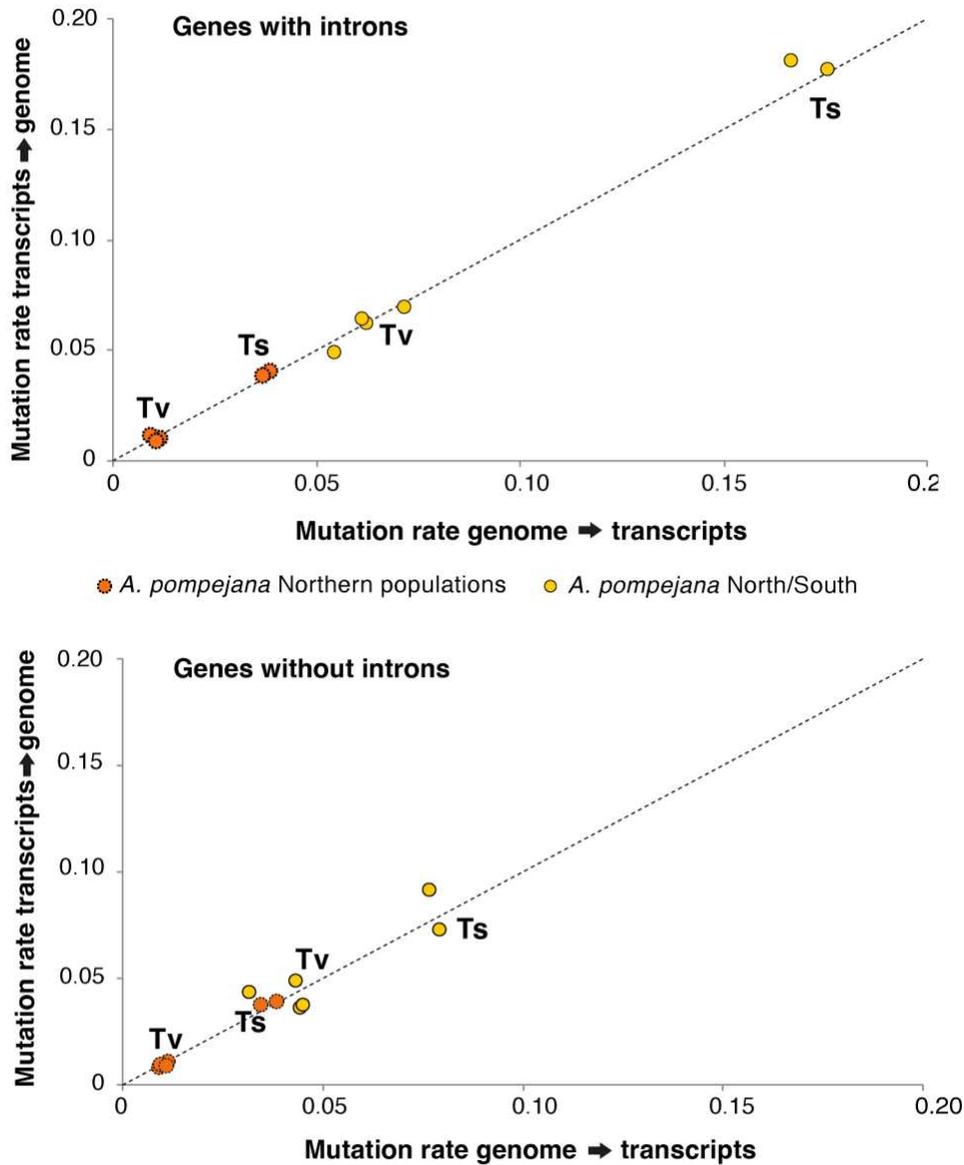
221 Distribution of genetic divergence and d_N/d_S values



222

223 **Figure 2:** Boxplots of the distribution of the divergence values found in different stages of speciation in
224 *Alvinella spp.* worms. The yellow color corresponds to the comparison *A. pompejana south/north*. The red
225 color corresponds to the comparison *A. caudata/A. pompejana*. The orange color corresponds to the
226 comparison *A. pompejana* Northern populations.

227 The evolutionary dynamics of genes during speciation was investigated by calculating
228 pairwise synonymous to non-synonymous substitution rates and the associated divergence
229 between orthologous transcripts of divergent individuals within and between closely-related
230 species (see the “Methods” section). Among the 6,916 orthologous genes studied between the two
231 Northern and Southern individuals of *A. pompejana*, 58% of the coding sequences are strictly
232 identical with no nucleotide differences, 26% of the genes display divergences between 0.5% and
233 1.75% ($0.01 > t > 0.05$), and only 4% of the genes represent outliers that diverge by more than 2%
234 ($t > 0.05$) along their sequence (Figure 2). As expected in early speciation steps, the overall mean
235 divergence of genes is low ($t = 0.016$) with values following a negative exponential distribution
236 with a long tail of outliers. However, the distribution has a slight bimodal shape, the second
237 unexpected mode being centered around a value ten times greater (i.e., about 0.020 substitution
238 per codon) (Figure 2). This bimodal shape was not encountered when assessing the intra-specific
239 allelic diversity between two Northern individuals (average diversity = 0.0096 ± 0.0051 with fewer
240 outliers ($t > 0.05 = 2\%$), see Figure 2). Analysis of both transition and transversion rates between
241 individuals from North and South were 4 times greater than those found by comparing two distinct
242 individuals of the same population in the northern part of the EPR (Figure 3).



243

244

245

246

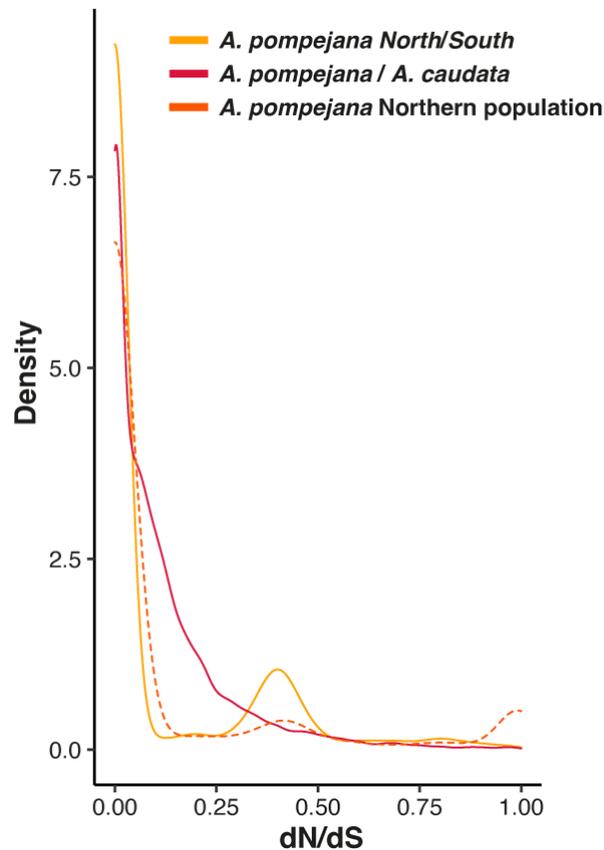
247

248

249

Figure 3: Transition and translation rates obtained between orthologous pairs of genes of *Alvinella pompejana* by comparing single-individual transcriptomes from the north (9°50N) and south (18°25S) EPR onto the genome of the worm (northern individual). Rates were obtained from coding sequences obtained from the orthologous gene datasets used for the d_N/d_S analysis by discriminating genes without introns and genes with introns (i.e. the exonic regions of genes obtained after mapping transcripts onto the genome).

250 On the opposite, the large majority of the 6,687 genes are highly divergent between *A.*
251 *pompejana* and *A. caudata*. Divergence follows a Gaussian distribution centered around 0.2
252 substitution per codon with an asymmetric tail of outliers (Figure 2). Only 0.4% of the sequences
253 are strictly identical with no nucleotide differences corresponding to only 33 pairwise alignments,
254 3% of the genes display divergence levels between 10% and 25% ($0.25 > t > 0.75$), and 96% of the
255 genes diverge by more than 25% ($t > 0.75$) along their sequence. Thus, the distribution of the
256 divergence between *A. caudata* and the 2 populations of *A. pompejana* is significantly different
257 (Figure 2; Wilcoxon test, $p < 2.2e-16$), but this distribution is not significantly different within *A.*
258 *pompejana* populations (Wilcoxon test, $p = 0.8266$).



259 **Figure 4:** Density distributions of d_N/d_S values estimated for each pairwise alignment of orthologous genes
260 in both the early and late states of speciation in *Alvinella spp.* worms using the method of Nielsen & Yang
261 (1998) implemented in PamlX. The yellow color corresponds to the comparison *A. pompejana south/north*.
The red color corresponds to the comparison *A. caudata/A. pompejana*. The orange color corresponds to the
comparison *A. pompejana* Northern populations.

262 Regarding the distribution of the d_N/d_S values between North and South individuals of *A.*
263 *pompejana*, most of the genes are non-divergent with a d_N/d_S ratio set to zero (Figure 4). Among
264 diverging genes, 44% of the genes are under strong purifying selection with d_N/d_S values lower
265 than 0.25. 36% of the d_N/d_S values are distributed in a small peak between 0.25 and 0.5, and 6%
266 of the genes are under positive selection with d_N/d_S values higher than 1 and represent about 3%
267 of the total number of genes examined (Figure 4). This peak of d_N/d_S was not encountered when
268 comparing the two northern *A. pompejana* transcriptomes although the number of outliers
269 ($d_N/d_S > 1$) was higher between sequences of these two specimens (about 12%). The absolute values
270 of d_N/d_S and π_N/π_S were greater than the 2% of genes under positive selection ($d_N/d_S > 1$) found
271 between *A. pompejana* and *A. caudata*, probably because of a sampling bias in estimating both
272 synonymous and non-synonymous rates from poorly intra-specific divergent sequences. The
273 remaining d_N/d_S values lesser than one between *A. pompejana* and *A. caudata* (98%) contains
274 about 30% of genes with d_N equal to zero (*i.e.*, “frozen” proteins), about 53% of genes with d_N/d_S
275 values ranging from 0 to 0.25 (genes under strong to moderate purifying selection) and 15% of
276 genes evolving under more relaxed conditions (close to one). The distribution of d_N/d_S values
277 within and between species was significantly different (Wilcoxon test, $p < 2.2e-16$) with a clear
278 bi- to tri-modal distribution of the d_N/d_S values in the specific case of the within-species
279 comparison. The average d_N/d_S estimated for both early and late steps of speciation was nearly
280 equal with values of 0.14 and 0.17, respectively. In other words, at least 86% and 84% of non-
281 synonymous mutations are deleterious and do not fix in the two *Alvinella* worms. This represents
282 a minimum estimate of the fraction of strongly deleterious mutations because even a small number
283 of advantageous mutations will contribute disproportionately to divergence [54].

284

285 **Functions of positively selected genes and genes with very high evolutionary rates.**

286

287 Early stage of speciation (*A. pompejana* south/north comparison)

288

289 Among the 41 genes under positive selection ($d_N/d_S > 1$) for which d_N/d_S values remained
290 significant after the resampling test of the d_N/d_S difference and the FDR procedure, only 12 of
291 them were annotated in the UniProt database. These genes are encoding proteins involved in (1)
292 transcription/translation/replication regulation and biosynthesis (ribosomal protein rpl19, GTP-
293 binding protein 1, elongation factor 1alpha, zinc finger protein GLI1) some of which being
294 involved in a wide variety of biological functions including spermatogenesis, (2) endocytosis and
295 immunity regulation, which may play a crucial role in the worm's epibiosis (Rab5 protein, CD209
296 antigen protein A), (3) neuro-transmission regulation (complexin) and the development of sensory
297 organs (protein mab-21), (4) development of nephridies, and possibly gonoducts in alvinellids
298 (actin-binding Rho protein), (5) mRNA methylation (pre-miRNA 5'-monophosphate
299 methyltransferase), and (6) setae/modified hook formation that should influence
300 reproductive/thermoregulatory behavior (chitin synthase).

301

302 Late stage of speciation (the *A. pompejana*/*A. caudata* comparison)

303

304 In the same way, among the 102 orthologous genes between *A. caudata* and *A. pompejana*
305 for which d_N/d_S values were greater than one, only 37 remained significant after the resampling
306 test of the d_N/d_S difference and the FDR procedure but, in this case, most them were annotated (31
307 genes) in the UniProt database. These genes encode for proteins involved in (1) carbohydrate

308 catabolism (deoxyribose-phosphate aldolase) , (2) immunity (innate and adaptive), viral responses,
309 apoptosis (fibrinogen-like protein, lysosomal protective protein, acetylcholine receptor), (3)
310 steroid/lipid metabolism and membrane composition (transmembrane protein, 3-oxo-5-alpha-
311 steroid 4-dehydrogenase, methylsterol monooxygenase, protein DD3-3-like), and tegument
312 structure (actin, collagen, epidermal growth factor named fibropellin, GTP-binding protein,
313 protein-methionine sulfoxide oxidase MICAL1), (4) sexual differentiation,
314 spermatogenesis/oogenesis, sperm adhesion (zonadhesin, innexin), (5) DNA damage repairs and
315 methylation (ATP-dependent RNA helicase DDX51 and DDX58, poly-(ADP-ribose) polymerase,
316 forkhead box G1 protein WD repeat-containing protein), oxidative stress response and protein
317 glycosylation (beta-1.3-galactosyltransferase 5-like, C19orf12), (6) cell proliferation/apoptosis,
318 mitosis, neuron formation (Tumor Necrosis Factor alpha, ras-related protein, APC/C activator
319 Cdc20 protein, tyrosine-protein kinase, protein FAM134C, zinc finger protein), and (7) signaling
320 mediated by metals (metabotropic glutamate receptor). Despite a suspected several millions of
321 years of species separation, it is worth-noting that positive selection is still acting on genes
322 involved in reproductive isolation, namely on zonadhesin and innexin, hence possibly reinforcing
323 species boundaries.

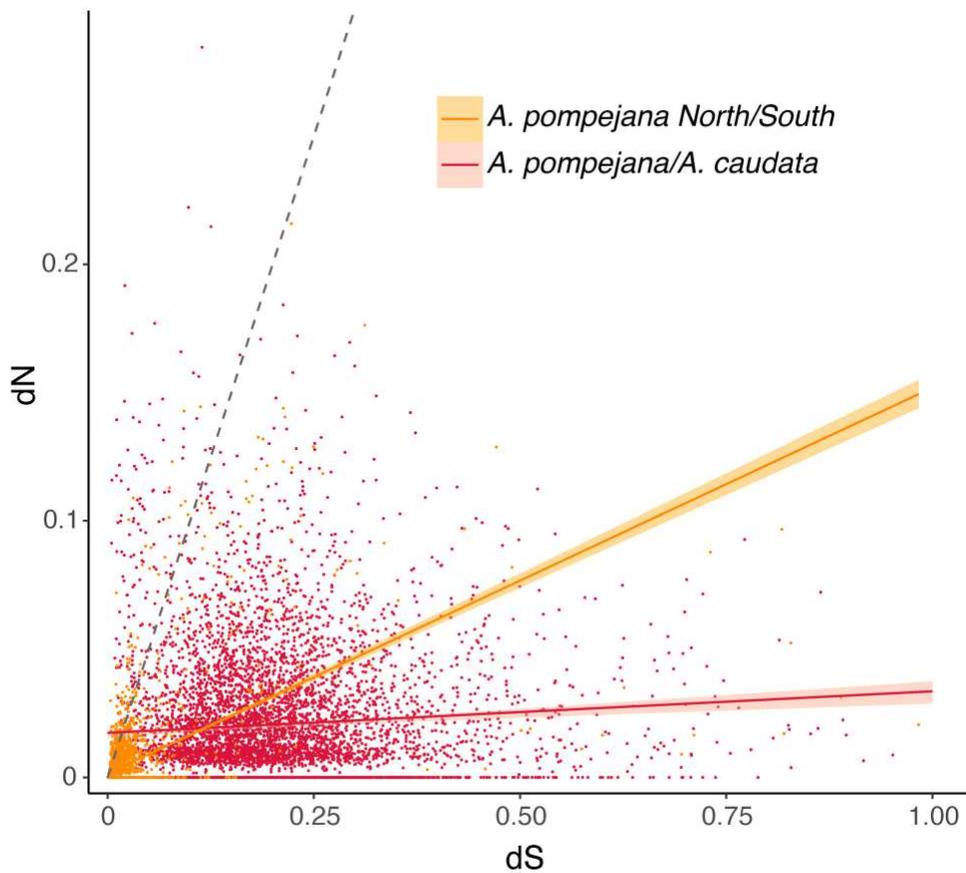
324

325 **Patterns of nucleotide substitutions and highly divergent genes**

326

327 Because natural selection usually operates mostly on the non-synonymous sites (but also
328 on translational accuracy [55]), the relative values of d_N and d_s may provide a key for making
329 inferences about the adaptive evolution of protein-coding genes. Usually, the synonymous rate is
330 three times greater than the rate of substitutions at nonsynonymous sites but should be equal under

331 neutral evolution. Thus, the depression in non-synonymous substitution rate is interpreted as being
332 caused by natural selection eliminating deleterious mutations to comply with both the 3D structure
333 and functions of proteins. The relationship between these two relative rates of substitutions was
334 assessed in the early and late steps of speciation thanks to a linear regression to take into account
335 the very high variance of these rates. The two comparisons greatly differ in the slope and intercept
336 of the linear regression between d_N and d_S (*A. pompejana* North/South: $y = 0.002631 + 0.040525x$;
337 *A. pompejana* / *A. caudata*: $y = 0.01818 + 0.01186x$) (Figure 5).
338



339
340 **Figure 5:** Rates of synonymous and non-synonymous divergences between *A. pompejana* North and South
341 populations (yellow) and *A. pompejana* and *A. caudata* (red). The grey dashed line represents the expected
342 linear relationship between d_S and d_N under neutral evolution.

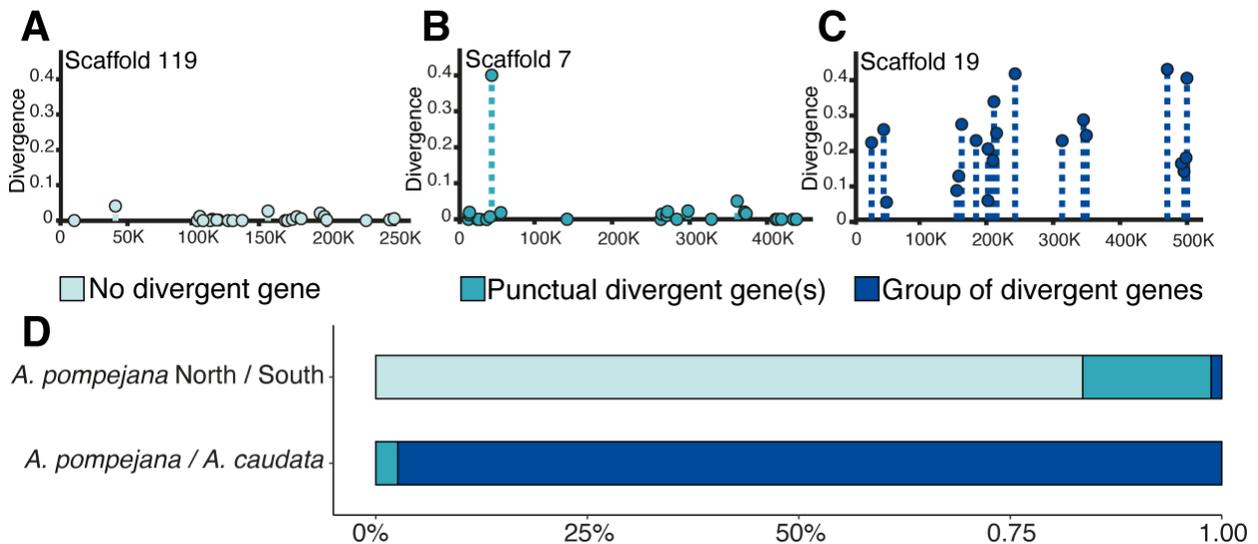
343 The number of highly divergent genes (i.e. in the tail distribution of the genes) is almost
344 five times greater (92 vs 20 annotated genes) and much more diversified when comparing the late
345 and early stages of speciation, respectively. Interestingly, most genes with saturated d_s rates in the
346 early step of speciation are either histone deacetylases or belong to the tubulin multigene family
347 (Supplementary Table 1). Most α - and β -tubulins are under strong purifying selection (indicated
348 by the relatively low rate of non-synonymous mutations) but are likely to contain at least one or
349 two non-synonymous substitutions in their divergence, suggesting that the family may have
350 endured a burst of duplications, and subsequent independent lineages sorting from both sides of
351 the East Pacific Rise barrier. In contrast, most of the 92 annotated genes with a high evolutionary
352 rate in the late step of speciation encode for a wide range of functions (e.g., replication, translation,
353 intracellular transport...), which are quite similar to those also depicted for positively-selected
354 genes in the Ap/Ac comparison. More specifically, highly divergent genes which also exhibited
355 fixed non-synonymous substitutions were also found in steroid/phospholipid metabolism
356 (phospholipid-transporting ATPase, sphingolipid delta(4)-desaturase, fatty acid-binding protein,
357 inositol polyphosphate 5-phosphatase), carbohydrate catabolism and the tricarboxylic cycle
358 implied in the worm's epibiosis with specific production of acetyl-CoA (carbohydrate
359 sulfotransferase4 dihydrolipoyllysine-residue acetyltransferase, acyl-CoA synthetase), cell
360 motility, spermatogenesis, and egg-sperm fusion (dynein, integrin, disintegrin, protocadherin),
361 embryonic and neuron development (methenyltetrahydrofolate synthase, protein sidekick),
362 membrane (nidogen-1, lipoprotein receptor1, EH domain-containing protein3, transmembrane and
363 coiled-coil domain-containing protein1, spectrin) and tegument organization (HHIP-like hedgehog
364 protein1, biotin, α -actinin) but also, proteins more targeted on putative "adaptive" molecular
365 pathways such as oxidative and thermal stress response (peroxidase, peroxiredoxin 4, glutathione

366 peroxidase, Hsp70, Hsp83), DNA repairs and methylation (ATP-dependent RNA helicase, DNA
 367 ligase1, RNA polymerase II transcription subunit, adenosylhomocysteinase B, BRCA1-associated
 368 RING domain protein1, histone deacetylase3), mRNA A-to-I editing (Double-stranded RNA-
 369 specific adenosine deaminase, xanthine dehydrogenase), protein glycosylation (Protein O-linked-
 370 mannose beta-1.2-N-acetylglucosaminyltransferase1), adaptation to hypoxia (Hypoxia-inducible
 371 factor1, carbonic anhydrase) and proteins involved in protein homo- heterodimerization
 372 (complement factor H-related protein2, BTB/POZ domain-containing protein7).

373

374 **Distribution of divergence and associated d_N/d_S along the genome**

375

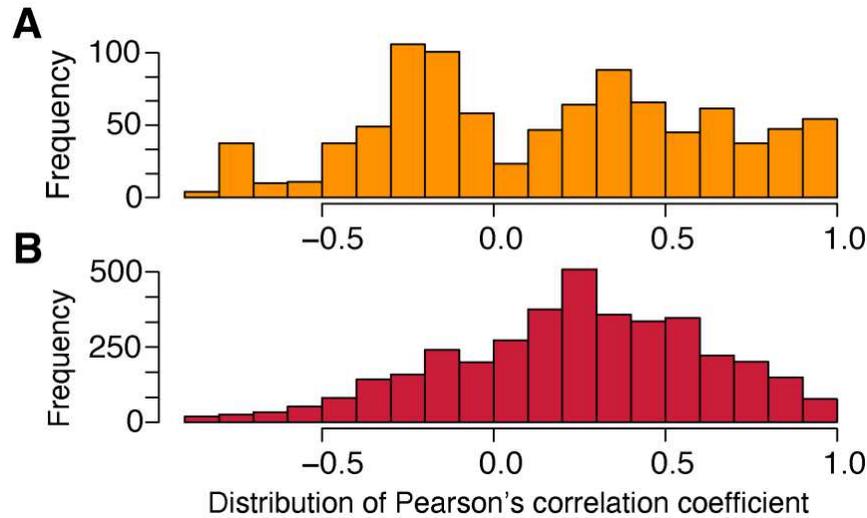


376

377

378 **Figure 6:** Genomic distribution of the divergence along the longest scaffolds (>400,000bp) *A. pompejana*
 379 draft genome (subset of genes representing about 1/10 of the genome). According to the index of distribution
 380 and our criterion of selection by eye, three distribution patterns of the divergence are found along scaffolds:
 (A) no divergent gene identified on the scaffold (in light blue), (B) sporadic presence of one divergent gene
 on a scaffold (in turquoise), and (C) group of spatially linked divergent genes (in dark blue). The relative
 proportion of each genomic pattern is shown in (D).

381 The distribution of gene divergence was examined over the 79 scaffolds exceeding 300,000
382 bp with the *A. pompejana* North/South comparison (Figure 6D). Because the global level of
383 divergence is low and close to zero, values of divergence are homogeneously distributed with a
384 few punctual observable outliers. As a consequence, about 83% of the examined scaffolds have no
385 divergent genes (Figure 6A and 6D). Scaffolds containing only one divergent gene (Figure 6B and
386 6D) account for 15% of the examined portion of the genome and only one of the largest scaffolds
387 displays an island of divergence pattern, and may be linked to sexual chromosome (Figure 6C and
388 6D). On the contrary, for the *A.caudata/A.pompejana* comparison, divergence is clearly structured
389 along the genome in large islands of highly divergent genes for 97% of the scaffolds (Figure 6D).
390 Each of the 76 scaffolds only carries very few genes with no divergence and we were not able to
391 identify any region or island of no divergence among the examined scaffolds. Here, divergence is
392 clearly structured in large blocks where all the genes and their putative exons are divergent. The
393 average values of d_N/d_S per scaffold are globally homogeneous between scaffolds but the
394 dispersion index of the d_N/d_S values around this mean fluctuates substantially. Gene density among
395 scaffold however greatly varies and therefore impacts the number of genes involved in the
396 genomes congealing.
397



398

399

Figure 7: Distribution of the Pearson's correlation coefficient between divergence and associated dN/dS values calculated for each scaffold containing more than 5 genes: (A) *A. pompejana* North/South in yellow and (B) *A. pompejana* / *A. caudata* in red.

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

For both the early and late stages of speciation, no correlation between d_S and d_N has been found over the whole set of orthologous genes (The North/South *A. pompejana* comparison: $r=0.18$, $p < 2.2e^{-16}$, The *A. caudata*/*A. pompejana* comparison: $r=0.07$, $p = 1.794e^{-08}$). We therefore estimated the Pearson's correlation coefficient within each scaffold to uncover putative islands of divergence and analyzed the distribution of this coefficient among scaffolds. The distribution of the Pearson's coefficient is almost bimodal and symmetrical around zero during the early stage of speciation (Figure 7A). Because most genes are weakly divergent (i.e., close to zero), this may indicate that scaffolds segregate into two distinct patterns: polymorphic genes for which allelic lineage sorting is incomplete and where the d_N/d_S is biased by excesses of deleterious/artifactual mutations and (2) genes for which allelic lineage sorting is completed with a fixed divergence where positive selection may have acted. In contrast, the distribution of the correlation coefficients displayed a clear shift towards a positive correlation between divergence and associated d_N/d_S values during the late stage of speciation (Figure 7B). Because most genes are highly divergent,

415 this overall increase of d_N/d_S with divergence therefore suggests that most of the genes have
416 evolved to adapt to a new ecological niche after the separation of the two sister species.

417

418 **Discussion**

419

420 Considering the speciation continuum from one end to the other is essential to shed light
421 on the mechanisms by which genomes separate, and more specifically the timing of reproductive
422 isolation and species specialization. In this study, we first examined an early stage of allopatric
423 speciation between populations of the Pompeii worm, *Alvinella pompejana* geographically
424 separated by a physical barrier to gene flow at the Equatorial triple junction of the East Pacific
425 Rise and the Galapagos rift [3,17]. This geographic separation of the Pompeii worm populations
426 is also associated with different venting dynamics related to different spreading rates of the ridge
427 segments that may play a crucial role in the proportion of still-hot chimneys and colder vent
428 habitats and thus the distribution of the microbiota/biota [8,56]. Then we also examined the
429 molecular evolution signature associated with a much more advanced stage of the speciation
430 process that separated *A. pompejana* and its syntopic sister species *A. caudata*, which exhibits a
431 slightly different niche. This event is relatively old and possibly predates the separation of the EPR
432 and the North East Pacific ridge (JdF) communities by vicariance that led for example to the
433 separation of sibling species like *Paralvinella pandorae pandorae* and *P. p. irlandei*, about 28
434 Mya [1]. Mitochondrial and nuclear divergences estimated between *A. pompejana* and *A. caudata*
435 are indeed greater than those estimated for these latter species (supplementary data in [57]).

436

437 To investigate the effect of “neutral” evolution and both “negative” and “positive”
438 selection along genomes of separating species, we estimated both divergence and d_N/d_S values
439 from a collection of orthologous genes by comparing transcriptomes of *A. pompejana* and *A.*
440 *caudata* against a first draft of *A. pompejana* genome. Although the d_N/d_S ratio is determined by
441 the combined effects of neutral, advantageous and deleterious mutations, it can tell us a lot about
442 the general impact of natural selection on the evolution of the coding portion of the genome. In
443 this context, a divergence scan on both synonymous and non-synonymous sites represents a very
444 powerful tool for tracing back the role of natural selection in the semi-permeability of diverged
445 genomes and the time at which these genomes started to congeal. It should be however noted that
446 d_N/d_S provides reliable information only for genes with complete lineage sorting where a
447 substantial amount of substitutions have become fixed between the separated genetic entities, and
448 caution must be taken during the early steps of speciation where most of loci are still polymorphic.
449 On the other side of the speciation spectrum, highly divergent genomes may also contain a large
450 number of saturated sites that leads to an overestimation of the d_N/d_S ratio in fast-evolving genes.

451

452 During the early steps of speciation, the large majority of genes (about 70%) are stacked
453 around null values of d_N and d_S but the relationship between the two rates for the remaining genes
454 is on average 5:1, one non synonymous substitution for five synonymous ones, probably due to
455 unfixed mutations in the worm’s polymorphism. Allele divergence and associated d_N/d_S
456 distributions between (North vs South individuals) and within (North individuals) the Pompeii
457 worm’s populations are indeed quite similar although a greater number of higher values has been
458 observed in the first comparison. This can be partially explained by the fact that the Equatorial
459 barrier to gene flow is not completely sealed [15] allowing rare episodes of allele introgression.

460 On the opposite, the slope of the regression line between d_N and d_S is close to zero during the late
461 steps of speciation and clearly highlights a large accumulation of synonymous substitutions close
462 to saturation for a large number of genes. This strongly suggests that even if a great number of
463 non-synonymous substitutions are positively-selected, most of this signal is likely to have been
464 erased in fast-evolving genes. Outlier genes associated with high evolutionary rate ($d_S > 0.5$) should
465 thus represent fast-evolving genes that may have endured duplications, selective losses or gains of
466 duplicates according to habitat/geographic isolation during the early steps of speciation, but also
467 genes that may have already lost their 'adaptive' signal in the face of saturation at synonymous
468 sites during the late steps of speciation.

469

470 As a general observation across many taxa, the average value of d_N/d_S over genes between
471 closely-related species is around 0.20 [58-60]. Multiple species comparisons suggest that more
472 than 70% of the whole set of mutations are strongly deleterious in most species and up to nearly
473 30% are slightly deleterious or neutral [61]. Information gathered with the present study is
474 consistent with nearly 87% of mutations being deleterious or slightly deleterious in *Alvinella*
475 worms (d_N/d_S mean = 0.10-0.15). This suggests that the majority of *Alvinella* genes are under
476 strong purifying selection probably as a consequence of the extreme thermal conditions
477 encountered by the worms [57]. Given such background selection, it is highly possible that many
478 proteins with d_N/d_S ranging from 0.5 to 1 share some history of positive selection, without any
479 possibility to test it.

480

481 **Early steps of speciation: lack of genes implicated in reproductive isolation despite**
482 **accumulation of divergence at many genes?**

483

484 Results from our genome-wide study are consistent with previous results relying on the
485 mitochondrial genome and some other genes in showing the geographic isolation of *A. pompejana*
486 populations across the Equator [15,62-63]. We indeed observed a clear signal of divergence at
487 nearly 30% of genes between the transcriptomes of the southern and northern individuals of *A.*
488 *pompejana*, despite a huge number of non-divergent genes (55% of the total number of genes
489 examined). This emerging divergence corresponds to a modal value of 0.025 substitution per
490 codon and represents a 4-times increase of both transitional and transversional substitutions when
491 compared with pairwise alignments of the same orthologous genes between two distinct northern
492 individuals from the same population. This emerging divergence at the genome scale confirms the
493 ongoing allopatric speciation occurring for these lineages along the EPR possibly reinforced by
494 local adaptation. Time since the population splitting does not seem to have been however sufficient
495 to produce a specific genomic architecture of divergence (*i.e.*, islands of speciation) as most of the
496 divergent genes are scattered among scaffolds. This may be explained by the fact that the predicted
497 North-to-South migration, though very low [15] could still be able to overcome population
498 differentiation over large portions of the genome and thus, would prevent genes from accumulating
499 fixed divergence. Alternatively, time since population splitting may still be too short to result in a
500 complete lineage sorting for most genes and the subsequent emergence of genetic
501 incompatibilities. According to Orr & Turelli [64] the strength of reproductive isolation (*i.e.* fitness
502 load of putative hybrids) increases as a function of the squared number of newly fixed mutations
503 in separated populations. The number of fixed substitutions (and hence associated genetic
504 incompatibilities) between the two isolated entities represents 15560 substitutions in coding
505 sequences (as the product of 2,075 300-codons genes with a divergence of 0.025 substitutions per

506 codon (30% of the investigated genes)) for approximately 1/10 of the whole genome. Compared
507 to other eukaryotic models (*e.g.*, about 70,000 changes accumulated between sub-species of the
508 *Drosophila simulans* complex over 0.25 My [65] for about the same number of coding sequences),
509 the number of accumulated changes (*i.e.* ~15,560) is quite low if the two populations separated 1.5
510 Mya [15], even if the generation time may be about 10 times greater in *Drosophila*. Such
511 discrepancy might be explained by a strong stabilizing selection associated with these extreme but
512 constant environmental conditions over space and time. Gene networks models indeed predict that
513 the hybrid incompatibility dynamics may be greatly reduced by stabilizing selection while
514 producing a “basin of attraction” for the optimal genotype on both sides of the barrier to gene flow
515 [66].

516

517 In this early stage of speciation, very few genes showed evidence of positive selection and
518 these genes do not relate to genes currently depicted as involved in reproductive isolation (*e.g.*,
519 genes involved in gamete recognition, pheromones, mito-nuclear incompatibilities, hybrid sterility
520 or immune incompatibility [67-70]. Neither sperm-egg binding or egg-fusion proteins nor nuclear
521 proteins transferred to the mitochondria displayed signatures of positive selection or strong
522 divergence. However, almost all genes showing high evolution rates in our data encode for
523 tubulins, which represents highly diversified multigene families and can have a great influence on
524 cell division (mitosis/meiosis) and the spermatozoon flagellum architecture (which is extremely
525 reduced as a byproduct of the internal fertilization of oocytes in *Alvinella* [71]). It should also be
526 noted that a meaningful proportion of positively-selected genes could indirectly participate to the
527 establishment of reproductive barriers as they encode for proteins involved in spermatogenesis,
528 the development of gonoducts, the sensory system and setae/hooks that may play a role in male

529 and female pairings (*i.e.*, pre-zygotic isolation). Other genes under divergent selection are mostly
530 unknown or involved in immunity and endocytosis and may affect the interaction of the worm
531 with the associated campylobacteria-dominated microbial assemblages covering its tube and its
532 dorsal epidermis (*i.e.*, trophic and detoxification role of the worm's epibiosis [72]). Microbial
533 communities are likely to be genetically different between vent fields as previously observed for
534 other vent taxa [73] and thus could induce co-evolutionary pathways of divergence in the worm's
535 immune system [74]. In the specific cases of sympatric or parapatric speciation, genes under
536 positive selection tend to be predominant among rapidly evolving genes [75]. Synergic effects
537 between positively-selected genes (linkage disequilibrium) are likely to produce local barriers to
538 gene flow if their number is high enough [30,46], but this is obviously not the case here. Analysis
539 of the correlation between divergence and d_N/d_S revealed a potential antagonistic action of
540 selection on divergent genes, one half of scaffolds producing a negative correlation between the
541 two parameters whereas the other half exhibited a positive correlation. Such correlations are
542 clearly biased by the very high number of genes with no divergence but negative correlations over
543 many scaffolds could be interpreted as an excess of deleterious mutations due to incompletely
544 sorted polymorphisms for many genes. Additional work on the genetic differentiation of the two
545 separating lineages is currently undergone at both the genome and population scales using ddRAD
546 markers to complement this present study.

547

548 **Late stage of speciation: role of gene specialization and gamete recognition genes in**
549 **species reinforcement after genome congealing**

550

551 As opposed to the recent history of allopatric speciation in *A. pompejana*, the scan of gene
552 divergence (t , d_N , d_S) between *A. pompejana* and its syntopic sister species *A. caudata* provides a
553 contrasting story of gene specialization after the congealing phase of the two corresponding
554 genomes. Our results confirm previous studies [51-52], who showed that these species did not
555 share any allele at enzyme loci and refuted the initial hypothesis that the two worms corresponded
556 to two morphological ontogenetic forms of the same species [49]. Both species occupy the same
557 hydrothermal vent habitat (hottest part of vent chimneys) and have exactly the same geographic
558 range from 21°N/EPR to 38°S/PAR (Pacific-Antarctic Ridge) leading to the possibility of
559 sympatric speciation. The two species indeed live in a spatially heterogeneous environment, which
560 greatly varies in time [76-77]. According to Gaill et al. [78], the parchment-like tube of the two
561 worms is highly thermostable, and the two species share many adaptations allowing them to thrive
562 in these waters loaded with heavy metals and sulfidic compounds (*e.g.*, similar epibiotic flora
563 composition: [79], similar lipids in the cuticle: [80], similar branchial crown, gaz transfer structure
564 and haemoglobins [81-82], and same reproductive mode: [71]). Our study of gene divergence
565 however reveals a very old separation of the two species and a relatively high proportion of genes
566 (2%) still showing evidence of positive selection when considering that the average gene
567 divergence is very high (*i.e.*, close to 0.2 substitution per synonymous site). Most of the positively-
568 selected genes are not necessarily those with the highest rate of evolution, thus indicating that the
569 signal of positive selection may have been erased by excesses of synonymous substitutions for
570 many genes. Distribution of gene divergence appears quite homogeneous among scaffolds in terms
571 of basal divergence, which is a strong indication that a genome-wide congealing effect occurred
572 much earlier. We estimated the time of genome congealing around 26.5 Mya by using the
573 substitution rate previously estimated by Chevaldonné *et al.* [83]. The first mapping of divergence

574 and associated d_N/d_S showed however a strong heterogeneity of these variables along the genome
575 with some scaffolds having relatively higher divergence values. Although we cannot rule out that
576 these scaffolds could represent either regions of lower recombination rates or a portion of sex-
577 linked chromosome, their high divergence could also trace back to the early steps of the speciation
578 process (i.e. primary islands of speciation), in this case suggesting that the separation of the two
579 species may have been initiated much more earlier.

580

581 Gene divergence is also slightly positively correlated with d_N/d_S within each scaffold. This
582 therefore suggests that the increase in gene divergence have been accompanied by the
583 accumulation of non-synonymous mutations. Assuming that d_N/d_S only increase when
584 synonymous sites become saturated, this trend could either indicate an adaptive divergence
585 associated with a long period of gene specialization for the two sister species regarding their habitat
586 after their separation, or that the time since species separation is so high that most genes reached
587 saturation but this later interpretation does not fit well with the presence of positively-selected
588 genes. The genes under positive selection ($d_N/d_S > 1$) are more specifically involved in the
589 repair/replication/biosynthesis of nucleic acids but also in several metabolic pathways usually
590 involved in the stress response, carbohydrate catabolism and lipid/steroid metabolisms associated
591 with membrane organization and tegument formation. Targets of positive selection seem to be
592 consistent with gene specialization in response to different thermo-chemical regimes encountered
593 by the worms, and more specifically high temperatures, natural doses of radioactivity and high
594 concentrations of sulfides and heavy metals producing reactive oxygen species (ROS) and reactive
595 sulfur species (RSS), and suggest that the two species do not share exactly the same niche.
596 Annotation of fast evolving genes also revealed that most of them belonged to exactly the same

597 metabolic pathways impacted by positive selection including membrane/tegument organization.
598 Regarding fast evolving genes, we noted a net enrichment in proteins involved in oxidative stress
599 response (peroxidase, Hsps), adaptation to hypoxia, DNA repairs and methylation, mRNA A-to-I
600 editing, and protein glycolysation. The two latter molecular processes represent a powerful way to
601 adapt to differing thermal regimes [84-85]. These results therefore indicate that the adaptive
602 evolution of the two species has been a sufficiently 'old' process to erase most signatures of
603 positive selection that has likely played a crucial role in their ecological isolation. As opposed to
604 *A. caudata*, *A. pompejana* is one of the first colonizer of newly-formed and still hot chimneys [86]
605 when these edifices are made of porous anhydrite (barium/calcium sulfates) with temperatures
606 above 100°C. Consistently, slight differences in terms of protein composition with more charged
607 residues in *A. pompejana* and more hydrophobic/aromatic ones in *A. caudata* have been described
608 [57]. Interestingly, when looking at the proportion of each species between chimneys at a local
609 spatial scale (*i.e.*, hundreds of meters), it can be observed that newly-formed anhydrite edifices
610 and 'white smokers' are mostly inhabited by *A. pompejana*, whereas older edifices and 'black
611 smokers' are mainly characterized by a greater abundance of *A. caudata* [7]. Accelerated
612 differentiation of genes involved in the adaptive response of the worms to environmental variations
613 are likely to explain their different ecologies and behaviors to cope with 'high' temperatures. For
614 instance, tubes of the two species are closely entangled in colonies but with differences in their
615 opening [72] (D. Jollivet, *pers. obs.*). While the aperture of the tube of *A. pompejana* is widely
616 opened like a funnel with septa, that of *A. caudata* is ended by a small hole. The shape of the tubes
617 is likely related to the thermoregulatory behavior of the worms that helps individuals to refresh
618 their microenvironment through water pumping [77,87]. While *A. pompejana* is widely observed
619 moving inside and outside its tube to refresh it with the cold surrounding waters, *A. caudata* usually

620 stays in its tube but exhibits a tail without organs used to cultivate its epibiotic flora that may be
621 viewed as a thermal sensor to position the worm at exactly the right temperature.

622

623 Despite this very long history of speciation and subsequent species specialization, one of
624 the most striking observations we made is that many genes involved in gamete recognition (*e.g.*,
625 zonadhesin, innexin) are under positive selection, suggesting that species sympatry may still play
626 a role in species reinforcement. These positively-selected proteins are accompanied by several
627 fast-evolving proteins involved in spermatogenesis, and egg-sperm fusion (*e.g.*, dynein, integrin,
628 disintegrin, protocadherin), and embryonic development (*e.g.*, methenyltetrahydrofolate
629 synthase). Moreover, specific pathways such as the Acetyl-CoA cycle (3 targets) together with a
630 lipid desaturase seems to become highly differentiated. In *Drosophila*, desaturase genes and the
631 Acetyl-CoA cycle are both involved in the production of cuticular hydrocarbons (CHCs) which
632 have a crucial role in adaptation to desiccation and also act as pheromones involved in mating
633 behavior [70]. It is therefore possible that these two worm species preliminary evolved in a
634 different geographical context for a long period of time and then met secondarily during the
635 recolonization of the EPR after the episodes of tectonic isolation that structured the hydrothermal
636 fauna between 11-18 Mya as recently suggested to explain the phylogeographic patterns of the
637 *Lepetodrilus elevatus* complex of limpet species [4]. This rather late evolution of prezygotic
638 mechanisms of isolation contrasts with the allopatric situation encountered between the northern
639 and southern populations of *A. pompejana* and supposes that these mechanisms could last for a
640 very long time even after a genetic barrier has been erected.

641

642

643

Conclusion

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

Material and methods

659

660

1) Animal sampling and genome/transcriptome sequencing

661

662

663

664

665

Our analyses pointed out the non-negligible role of natural selection on both the early and late stages of speciation in the emblematic thermophilic worms living on the walls of deep-sea hydrothermal chimneys. They shed light on the evolution of gene divergence during the process of speciation and species specialization over time. Due to habitat fragmentation, populations separate and start to accumulate putative adaptive mutations (and possibly genetic incompatibilities) in allopatry where they find differing local conditions but are likely to consolidate reproductive isolation in sympatry following secondary contacts during the specialization process, even after genomes diverged almost completely (congealing effect). These first analyses raise many further questions about the evolutionary mechanisms that led to the speciation of *Alvinella* spp. and their subsequent distribution within the spatial micro-mosaic of habitats typifying hydrothermal vents. Additional studies combining polymorphism and divergence are needed to better understand the respective roles of geographic and ecological histories of the worm's populations in speciation in a such fragmented and instable environment.

The genome of the Pompeii worm was assembled from a single individual collected at 9°50N/EPR. Several DNA purification methods using alcohol precipitation and/or column purification resulted in DNA samples that could not be effectively digested by restriction enzymes. To reduce mucopolysaccharides or similar co-precipitating contaminants, we purified the DNA from worm tissue with phenol extraction and isopropanol precipitation followed by a

666 clean-up step with cesium chloride gradient ultracentrifugation [88]. The DNA was used to prepare
667 paired-end shotgun and 5 kb mate-pair libraries for Illumina genome sequencing [89].

668

669 We used at least two additional individuals of *A. pompejana* and *A. caudata* coming from the same
670 locality to perform RNA sequencing and the subsequent assembly of transcriptomes for the
671 northern specimens of the two species [57]. In parallel, a Sanger sequencing project of a reference
672 and fully annotated transcriptome was performed using several individuals coming from a single
673 chimney of the hydrothermal vent field 18°25S [90], for additional information about this
674 annotated transcriptome) and, represented the third transcriptome used to characterize the
675 specimens of *A. pompejana* located further south to the Equatorial barrier to gene flow previously
676 described [3,15].

677

678 **2) Sequencing** (see Metrics in Supplementary Table 2)

679

680 ● Genome and transcriptomes

681 Genome of the Pompeii worm (about 370 Mb) was sequenced from a single individual using both
682 a shotgun and mate-pair 5 kb sequencing using the Illumina HiSeq2500 technology at the A*Star
683 Molecular and Cell Biology Institute (*Alvinella* genome project: coord. A. Claridge-Chang).
684 Paired-end reads libraries (insert size 300bp) were assembled using SOAPdenovo2 (v2.0, -K 41 -
685 d -R), contigs were then scaffolded with OPERA (v1.3.1) using both paired-end (insert size 300bp)
686 and mate pair libraries (insert size 5 kb). Finally, the assembly was gapfilled with Gapcloser
687 (v1.12) to subsequently produce about 3000 scaffolds with a size greater than 15 kb.

688 The annotated transcriptome of the southern form of the Pompeii worm collected at the vent field
689 18°25S has been already published in 2010 [90], and transcriptomes of the northern forms of *A.*
690 *pompejana* and *A. caudata* have been already published in 2017 [57] (see Availability of data and
691 materials). All transcriptomes were assembled *de novo* using the software Trinity [91], the
692 sequencing effort was a quarter of a lane (40million reads) per species on a HiSeq 2000 at Genome
693 Québec.

694

695 **3) Estimation of divergence and d_N/d_S for pairs of orthologous genes**

696

697 We designed a customized pipeline (described below) in order to calculate divergence and
698 estimate the selective pressures along the coding regions of assembled transcriptomes of closely
699 related species previously positioned on our reference genome, using a 24-columns MegaBLAST
700 outputs. The code to perform analyses for this study is available as a git-based version control
701 repository on GitHub (https://github.com/CamilleTB/dNdS_cartography_Alvinella_speciation)

702

- 703 • Megablast

704 Because coding sequences are compared between divergent individuals within a given species or
705 between closely-related species, we performed a sequence-similarity search using each collection
706 of transcripts obtained for *A. caudata* and the two geographic forms of *A. pompejana* against the
707 entire *Alvinella pompejana*'s genome with the software MegaBLAST implemented in NCBI
708 BLAST+ 2.2.30 nucleotide-nucleotide search, optimized for very similar sequences (set
709 expectation value cutoff = 10^{-15} , maximum hits to show = 1).

710

711 • Paralog

712 The 24-columns MegaBLAST output was then filtered using a custom-made Perl script (called
713 “Paralog”) to discriminate genes according to their position and occurrences in the genome. This
714 allowed us to rule out putative duplicated genes in our subsequent analyses. Based on the positions
715 of the hits on both scaffolds and transcripts, genes were separated into distinct categories, namely
716 scaffold paralog, tandemly repeated paralogs, allelic forms of a single gene in either genes with or
717 without introns. Scaffold paralogs correspond to coding sequences found on distinct scaffolds,
718 tandemly-repeated genes to sequences found at non-overlapping positions onto the same scaffold,
719 allelic forms of a gene corresponded to several transcripts matching exactly the same positions
720 onto a scaffold, and the two others categories were orthologous transcripts only found once into
721 the genome at a given scaffold position with a discrimination exon/uniq when the Megablast hits
722 corresponded or not to a succession of fragments with different positions both onto the scaffold
723 and the transcript (i.e. exonic regions). Paired orthologous sequences obtained in the two last
724 categories (exon and uniq) were then used to estimate divergence and associated d_N/d_S ratios in
725 the pipeline.

726

727 • Search for coding DNA sequences (CDS) in pairwise alignments

728 Nucleic acid sequences were translated in the six reading frames to select for the longest CDS
729 between two subsequent stop codons. Coding sequences containing gaps, undetermined nucleic
730 acids (N) or alternative frames without stop were removed. To avoid false positives and reduce the
731 risk of calculating erroneous d_N/d_S values from non-coding or incorrectly framed sequences, we
732 performed a sequence-similarity search against the UniProt database using BlastP from NCBI
733 BLAST+ 2.2.30 (set expectation value cutoff = 10^{-10} , maximum hits to show = 1). Sequences with

734 no annotation from UniProt with a CDS shorter than 300 nucleotides were removed from the
735 analysis.

736

737 • Divergence and d_N/d_S calculation

738 Estimation of divergence, synonymous and nonsynonymous substitution rates and detection of
739 positive selection in protein-coding DNA sequences were performed using the program yn00 from
740 pamlX 1.2 in batch [92-93]. This program implements the method of Yang & Nielsen
741 (implemented in [92]) to calculate d_S and d_N values from pairwise comparisons of protein-coding
742 DNA sequences. The weighing parameter that decides whether equal or unequal weighing will be
743 used when counting differences between codons was set to zero (= equal pathways between
744 codons) and we used the universal genetic code (icode = 0). Output from yn00 was then parsed
745 and subsequently filtered from infinite values (99) when $S.d_S$ was equal to zero. When both d_N and
746 d_S values were null (about 70% of values obtained when comparing the northern and southern
747 forms of *A. pompejana*), d_N/d_S was reset to zero. To avoid a division by zero, if d_N was different
748 from zero but d_S was zero, $S.d_S$ was reset to 1 assuming that at least one synonymous mutation
749 was missed by chance in the sequence.

750 To avoid any potential effect of greater divergence at non-synonymous sites possibly due to
751 sequencing errors when values of neutral divergence (d_S) departed from zero, the significance of
752 all d_N/d_S values greater than zero was tested using a custom-made pipeline. For that purpose, we
753 produced 1,000 random alignments of each pair of coding sequences using a bootstrap resampling
754 of codons (from the bootstrap option of codeML in PamlX 1.3.1). We then estimated the associated
755 d_N and d_S for each resampled alignment, and calculated the difference (D) between each d_N and d_S
756 resampled values in order to test whether the observed value of this difference was significantly

757 greater than zero (*i.e.*, fall outside the distribution of the pseudo-replicates). To this end, we
758 performed a unidirectional one-sample z -test for each resampled set of paired sequences and
759 adjusted p-values over the whole set of genes using the false discovery rate correction representing
760 the expected proportion of false positives [94].

761

762 **4) Identification of islands of divergence and/or adaptation**

763

764 To investigate the genomic architecture of molecular divergence of putative genes under
765 strong purifying (d_N/d_S close to zero) or positive selection ($d_N/d_S > 1$) along chromosomes, both the
766 divergence and their d_N/d_S associated values were positioned along the longest scaffolds ($>300,000$
767 bp with more than 5 genes per scaffold) of the draft genome of *A. pompejana* for each pairwise
768 comparisons (*i.e.*, AP north vs AP south, and AP vs AC). According to the gene distribution along
769 each scaffold, three different types of patterns were characterized depending on the number and
770 level of clustering of divergent genes (divergence > 0.1) among scaffolds: (1) scaffolds without
771 divergent genes, (2) scaffolds with one or two isolated divergent genes and, (3) scaffolds carrying
772 a genes cluster of divergence corresponding to a putative island of speciation. The proportion of
773 each type was then calculated for the two sets of analyses. In addition, we calculated the mean
774 scaffold distance (d) between two divergent genes and its standard deviation ($s^2(d)$) within each
775 scaffold. We then tested for contiguous distribution patterns of divergence along each scaffold
776 thanks to the Fisher's index of dispersion ($s^2(d)/\text{mean}(d)$). Divergent genes are randomly
777 distributed when the index is equal to one, over-dispersed for values lower than one and aggregated
778 for values greater than one.

779

780 We finally tested for a positive correlation between d_N/d_S and divergence (t) along scaffolds
781 (>300,000 bp with more than 5 genes per scaffold) using a home-made R script. Such a correlation
782 should indicate a genome-wide adaptive evolution of proteins assuming that, under neutral
783 evolution, d_N/d_S remains relatively constant before the time at which most of the synonymous sites
784 become saturated, and then slightly decreases with longer time of separate evolution.

785

786 **Abbreviations**

787 AP: *Alvinella pompejana*; AC: *Alvinella caudata*; EPR: East Pacific Rise; JDF: Juan de Fuca
788 Ridge; FDR: False discovery rate; MYA: Million years ago.

789 **Acknowledgments**

790 We would like to thank the captain and crews of the research vessel L'Atalante and the pilots and
791 associated teams of the manned submersible Nautile without which no specimen would have been
792 sampled. We are also very grateful to the chief-scientists of the oceanographic cruises Biospedo
793 2004 and Mescal 2010 for collecting *Alvinella* specimens and to S.C. Cary who kindly provided
794 in 2006 the AP specimen used to produce the genome. We thank Romain P. Boisseau for
795 considerable feedback on the manuscript and Ulysse Guyet for useful discussions on the statistical
796 analyses.

797 **Authors' contributions**

798 CTB and DJ performed analyses and wrote the manuscript. EB and DJ conceptualized, designed
799 and co-supervised the study. SH and DJ were in charge of the funded research project. ACC
800 supervised the AP genome project. DB, NN, RC and EC were involved in the AP genome
801 assembly, transcriptome assembly and mapping, and gene annotation. DJ, SH, and ACC edited

802 and contributed to the proofreading of the manuscript. All authors read and approved the final
803 manuscript.

804 **Funding**

805 This study was supported by the Total Fondation. The funders had no role in conceptualizations
806 and study design, data collection and analysis, decision to publish, or preparation of the
807 manuscript.

808 **Availability of data and materials**

809 The AP genome: Two replicate runs of Illumina were performed onto the same individual (2 sets
810 of R1 and R2 compressed files of 33 Go: cf. SRA project under experiment accession numbers
811 ERX5990023-25; ERX5990051 and, run accession number: ERR6358489-91; ERR6358517) and,
812 used in combination to generate the first genome draft (V1). The reads of transcriptomes are
813 available from the Sequence Read Archive (SRA) database under accession numbers SRP076529
814 (Bioproject number: PRJNA325100) and SPX055399-402 (Bioproject number: PRJNA80027)),
815 and from the EST database under accession numbers FP489021 to FP539727 and FP539730 to
816 FP565142.

817

818 **Declarations**

819 **Ethics approval and consent to participate**

820 Not applicable.

821 **Consent for publication**

822 Not applicable.

823 **Competing interests**

824 All the authors declare no competing interests.

825

826 **Authors details**

827 ¹Present address: Emlen Lab - BRB108, Division of Biological Sciences, The University of
828 Montana, Missoula, MT, USA

829 ²Sorbonne Université, CNRS, UMR 7144 AD2M, Station Biologique de Roscoff, Place Georges
830 Teissier CS90074, 29688 Roscoff, France

831

832

833 **References**

834

835 [1] Tunnicliffe, V The biology of hydrothermal vents: ecology and evolution. *Oceanogr Mar Biol*
836 *Annu Rev.* 1991;29:319-407.

837 [2] Van Dover CL, German CR, Speer KG, Parson LM, Vrijenhoek RC. Evolution and
838 biogeography of deep-sea vent and seep invertebrates. *Science.* 2002;295(5558):1253–57.

839 [3] Plouviez S, Shank TM, Faure B, Daguin-Thiebaut C, Viard F, Lallier FH, Jollivet D.
840 Comparative phylogeography among hydrothermal vent species along the East Pacific Rise
841 reveals vicariant processes and population expansion in the South. *Mol. Ecol.* 2009;18(18):3903–
842 17.

843 [4] Matabos M, Jollivet D. Revisiting the *Lepetodrilus elevatus* species complex (Vetigastropoda:
844 Lepetodrilidae), using samples from the Galápagos and Guaymas hydrothermal vent systems. *J*
845 *Moll Stud.* 2019;85(1):154–65.

846 [5] Lutz RA. Dispersal of organisms at deep-sea hydrothermal vents: a review. *Oceanol Acta.*
847 *Special Issue.* 1988;8:23-30.

- 848 [6] Tyler PA, Young CM. Reproduction and dispersal at vents and cold seeps. *J mar Biol Ass UK*.
849 1999;79(2):193-208.
- 850 [7] Jollivet D. Specific and genetic diversity at deep-sea hydrothermal vents: an overview. *Biodiv*
851 *Conserv*. 1996;5(12):1619–53.
- 852 [8] Vrijenhoek RC. Gene flow and genetic diversity in naturally fragmented metapopulations of
853 deep-sea hydrothermal vent animals. *J Hered*. 1997;88(4):285–93.
- 854 [9] Jollivet D, Chevalloné P, Planque B. Hydrothermal-vent alvinellid polychaete dispersal in
855 the Eastern Pacific. 2. a metapopulation model based on habitat shifts. *Evolution*.
856 1999;53(4):1128–42.
- 857 [10] Johnson KS, Childress JJ, Beehler CL. Short-term temperature variability in the Rose Garden
858 hydrothermal vent field: an unstable deep-sea environment. *Deep Sea Res Part A*. 1988;35(10-
859 11):1711-21.
- 860 [11] Le Bris N, Sarradin PM, Caprais JC. Contrasted sulphide chemistries in the environment of
861 13°N EPR vent fauna. *Deep Sea Res Part I*. 2003;50(6):737-47.
- 862 [12] Matabos M, Le Bris N, Pendlebury S, Thiébaud E. Role of physico-chemical environment on
863 gastropod assemblages at hydrothermal vents on the East Pacific Rise (13°N/EPR). *J mar Biol Ass*
864 *UK*. 2008;88(5):995–1008.
- 865 [13] O’Mullan GD, Maas PAY, Lutz RA, Vrijenhoek RC. A hybrid zone between hydrothermal
866 vent mussels (*Bivalvia*: *Mytilidae*) from the Mid-Atlantic Ridge. *Mol Ecol*. 2001;10(12):2819-31.
- 867 [14] Johnson SB, Won Y-J, Harvey JBJ, Vrijenhoek RC. A hybrid zone between *Bathymodiolus*
868 mussel lineages from Eastern Pacific hydrothermal vents. *BMC Evol Biol*. 2013;13(1):21.

869 [15] Plouviez S, Le Guen D, Lecompte O, Lallier FH, Jollivet D. Determining gene flow and the
870 influence of selection across the equatorial barrier of the East Pacific Rise in the tube-dwelling
871 polychaete *Alvinella pompejana*. BMC Evol Biol 2010;10(1):220.

872 [16] Plouviez, S, Faure B, Le Guen D, Lallier FH, Bierne N, Jollivet D. A new barrier to dispersal
873 trapped old genetic clines that escaped the Easter Microplate tension zone of the Pacific vent
874 mussels. PLoS One. 2013;8(12):e81555.

875 [17] Matabos M, Plouviez S, Hourdez S, Desbruyères D, Legendre P, Warén A, Jollivet D,
876 Thiébaud E. Faunal changes and geographic crypticism indicate the occurrence of a biogeographic
877 transition zone along the Southern East Pacific Rise. J Biogeogr. 2011;38(3):575–94.

878 [18] Won Y-J, Hallam SJ, O’Mullan GD, Vrijenhoek RC. Cytonuclear disequilibrium in a hybrid
879 zone involving deep-sea hydrothermal vent mussels of the genus *Bathymodiolus*. Mol Ecol.
880 2003;12(11):3185–90.

881 [19] Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding light on the grey
882 zone of speciation along a continuum of genomic divergence. PLoS
883 biology. 2016;14(12):e2000234.

884 [20] Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The
885 frequency of polyploid speciation in vascular plants. Proc Nat Acad Sci USA.
886 2009;106(33):13875–79.

887 [21] De Storme N, Mason A. Plant speciation through chromosome instability and ploidy change:
888 cellular mechanisms, molecular factors and evolutionary relevance. Current Plant Biol.
889 2014;1:10–33.

890 [22] Schöfl G, Heckel DG, Groot AT. Time-shifted reproductive behaviours among fall
891 armyworm (Noctuidae: *Spodoptera frugiperda*) host strains: evidence for differing modes of
892 inheritance. *J Evol Biol.* 2009;22(7):1447–59.

893 [23] Monteiro CA, Serrão EA, Pearson GA. Prezygotic barriers to hybridization in marine
894 broadcast spawners: reproductive timing and mating system variation. *PLoS One.*
895 2012;7(4):e35978.

896 [24] Drès M, Mallet J. Host races in plant-feeding insects and their importance in sympatric
897 speciation. *Phil Trans Roy Soc London, Ser B.* 2002;357(1420):471–92.

898 [25] Rundle HD, Nosil P. Ecological speciation. *Ecol Lett.* 2005;8(3):336–52.

899 [26] Butlin RK, Galindo J, Grahame JW. Sympatric, parapatric or allopatric: the most important
900 way to classify speciation? *Phil Trans Roy Soc B.* 2008;363(1506):2997–3007.

901 [27] Mallet J. Hybridization, ecological races and the nature of species: empirical evidence for the
902 ease of speciation. *Phil Trans Roy Soc B.* 2008;363(1506):2971–86.

903 [28] Berner D, Grandchamp A-C, Hendry AP. Variable progress toward ecological speciation in
904 parapatry: Stickleback across eight lake-stream transitions. *Evolution.* 2009;63(7):1740–53.

905 [29] Hendry AP, Bolnick DI, Berner D, Peichel CL. Along the speciation continuum in
906 Sticklebacks. *J Fish Biol.* 2009;75(8):2000–36.

907 [30] Nosil P, Funk DJ, Ortiz-Barrientos D. Divergent selection and heterogeneous genomic
908 divergence. *Mol Ecol.* 2009;18(3):375–402.

909 [31] Schluter D. Evidence for ecological speciation and its alternative. *Science.*
910 2009;323(5915):737–41.

911 [32] Gagnaire PA. Comparative genomics approach to evolutionary process connectivity. *Evol*
912 *Appl.* 2020;13(6):1320-34.

913 [33] Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL,
914 et al. Genomics and the origin of species. *Nature Rev Genet* 2014;15(3):176–92.

915 [34] Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of
916 speciation. *Nature Rev Genet.* 2016;18(2):87–100.

917 [35] Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in *Anopheles gambiae*.
918 *PLoS Biol.* 2005;3(9):e285.

919 [36] Via S, West J. The genetic mosaic suggests a new role for hitchhiking in ecological
920 speciation. *Mol Ecol.* 2008;17(19):4334-45.

921 [37] Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due
922 to reduced diversity, not reduced gene flow. *Mol Ecol.* 2014;23(13):3133-3157.

923 [38] Feder JL, Nosil P. The efficacy of divergence hitchhiking in generating genomic islands
924 during ecological speciation. *Evolution.* 2010;64(6):1729–47.

925 [39] Hofer T, Foll M, Excoffier L. Evolutionary forces shaping genomic islands of population
926 differentiation in humans. *BMC Genomics.* 2012;13(1):107.

927 [40] Larson WA, Limborg MT, McKinney GJ, Schindler DE, Seeb JE, Seeb LW. Genomic islands
928 of divergence linked to ecotypic variation in Sockeye salmon. *Mol Ecol.* 2017;26(2):554–70.

929 [41] Duranton M, Allal F, Fraïsse C, Bierne N, Bonhomme F, Gagnaire P-A. The origin and
930 remolding of genomic islands of differentiation in the European sea bass. *Nature. Comm.*
931 2018;9(1):1-11.

932 [42] Tavares H, Whibley A, Field DL, Bradley D, Couchman M, Copsey L, Elleouet J, et al.
933 Selection and gene flow shape genomic islands that control floral guides. *Proc Nat Acad Sci USA.*
934 2018;115(43):11006-11011.

935 [43] Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends Genet.
936 2012;28(7):342–50.

937 [44] Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. Genome-Wide
938 congealing and rapid transitions across the speciation continuum during speciation with gene flow.
939 J Hered. 2014;105(S1):810–20.

940 [45] Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the influence of
941 genomic architecture on the dynamics of speciation. Mol Ecol. 2014;23(16):4074–88.

942 [46] Tittes S, Kane NC. The genomics of adaptation, divergence and speciation: a congealing
943 theory. Mol Ecol. 2014;23(16):3938–40.

944 [47] Chevaldonné P, Desbruyères D, Childress JJ. ... and some even hotter. Nature.
945 1992;359(6396):593-594.

946 [48] Cary SC, Shank T, Stein J. Worms bask in extreme
947 temperatures. Nature. 1998;391(6667):545-546.

948 [49] Desbruyères, D, Laubier L. *Alvinella pompejana* gen. sp. nov., ampharetidae aberrant des
949 sources hydrothermales de la ride Est-Pacifique. Oceanol Acta. 1980;3(3):267-274.

950 [50] Desbruyères, D, Laubier L. Les Alvinellidae, une famille nouvelle d'annélides polychètes
951 inféodées aux sources hydrothermales sous-marines: systématique, biologie et écologie. Can J
952 Zool 1986;64(10):2227–45.

953 [51] Autem M, Salvidio S, Pasteur N, Desbruyères D, Laubier L. Mise en évidence de l'isolement
954 génétique de deux formes sympatriques d'*Alvinella pompejana* (Polychaeta: Ampharetidae),
955 annélides inféodées aux sites hydrothermaux actifs de la dorsale du Pacifique Oriental. C R Acad
956 Sci Paris. 1985;301:131–35.

957 [52] Jollivet D, Desbruyères D, Ladrat C, Laubier L. Evidence for differences in the allozyme
958 thermostability of deep-sea hydrothermal vent polychaetes (Alvinellidae): a possible selection by
959 habitat. *Mar Ecol Prog Ser.* 1995;123(1–3):125–136.

960 [53] Faure B, Jollivet D, Tanguy A, Bonhomme F, Bierne N. Speciation in the deep sea: multi-
961 locus analysis of divergence and gene flow between two hybridizing species of hydrothermal vent
962 mussels. *PLoS One.* 2009;4(8):e6485.

963 [54] Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of
964 molecular evolution. *Nature.* 1977; 267(5608):275–76.

965 [55] Stoletzki N, Eyre-Walker A. Synonymous codon usage in *Escherichia coli*: selection for
966 translational accuracy. *Mol Biol Evol.* 2007;24(2):374–81.

967 [56] Carbotte S, Macdonald K. East Pacific Rise 8°–10°30'N: Evolution of ridge segments and
968 discontinuities from SeaMARC II and three-dimensional magnetic studies. *J Geophys Res: Solid*
969 *Earth.* 1992;97(B5):6959–82.

970 [57] Fontanillas E, Galzitskaya OV, Lecompte O, Lobanov MY, Tanguy A, Mary J, Girguis PR,
971 Hourdez S, Jollivet D. Proteome evolution of deep-sea hydrothermal vent alvinellid polychaetes
972 supports the ancestry of thermophily and subsequent adaptation to cold in some lineages. *Genome*
973 *Biol Evol* 2017;9(2):279-296.

974 [58] Mikkelsen PM. Speciation in modern marine bivalves (Mollusca: Bivalvia): insights from the
975 published record. *Am Malacol Bull.* 2011;29(1–2):217–45.

976 [59] Rat Genome Sequencing Project Consortium. Genome sequence of the brown Norway rat
977 yields insights into mammalian evolution. *Nature.* 2004;428(6982):493–521.

978 [60] Gaut B, Yang L, Takuno S, Eguiarte LE. The patterns and causes of variation in plant
979 nucleotide substitution rates. *Ann Rev Ecol Evol Syst.* 2011; 42(1):245–66.

- 980 [61] Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. Quantifying the slightly deleterious
981 mutation model of molecular evolution. *Mol Biol Evol.* 2002; 19(12):2142–49.
- 982 [62] Hurtado LA, Lutz RA, Vrijenhoek RC. Distinct patterns of genetic differentiation among
983 annelids of Eastern Pacific hydrothermal vents. *Mol Ecol.* 2004;13(9):2603–15.
- 984 [63] Jang S-J, Park E, Lee W-K, Johnson SB, Vrijenhoek RC, Won Y-J. Population subdivision
985 of hydrothermal vent polychaete *Alvinella pompejana* across Equatorial and Easter Microplate
986 boundaries. *BMC Evol Biol.* 2016;16(1):235.
- 987 [64] Orr HA, Turelli M. The evolution of postzygotic isolation: accumulating Dobzhansky-Muller
988 incompatibilities. *Evolution.* 2001;55(6):1085–94.
- 989 [65] Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC.
990 Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res.*
991 2012;22(8):1499–1511.
- 992 [66] Palmer ME, Feldman MW. Dynamics of hybrid incompatibility in gene networks in a constant
993 environment. *Evolution.* 2009;63(2):418–31.
- 994 [67] Palumbi SR. Genetic divergence, reproductive isolation, and marine speciation. *Ann Rev Ecol*
995 *Syst.* 1994;25(1):547-572.
- 996 [68] Orr HA. The genetic basis of reproductive isolation: insights from *Drosophila*. *Proc Nat Acad*
997 *Sci USA.* 2005;102(1):6522–26.
- 998 [69] Presgraves DC. The molecular evolutionary basis of species formation. *Nature Rev Genet.*
999 2010;11(3):175–80.
- 1000 [70] Castillo, DM, Barbash DA. Moving speciation genetics forward: modern techniques build on
1001 foundational studies in *Drosophila*. *Genetics.* 2017;207(3):825–42.

1002 [71] Jouin-Toulmond C, Mozzo M, Hourdez S. Ultrastructure of spermatozoa in four species of
1003 Alvinellidae (Annelida: Polychaeta). Cah. Biol. Mar. 2002;43(3/4):391-394.

1004 [72] Desbruyères D, Chevaldonné P, Alayse A-M, Jollivet D, Lallier FH, Jouin-Toulmond C, Zal
1005 F, et al. Biology and Ecology of the ‘Pompeii Worm’ (*Alvinella pompejana* Desbruyères and
1006 Laubier), a normal dweller of an extreme deep-sea environment: a synthesis of current knowledge
1007 and recent developments. Deep Sea Res Part II: Topical Studies in Oceanography.
1008 1998;45(1):383–422.

1009 [73] Breusing C, Castel J, Yang Y, Broquet T, Sun J, Jollivet D, Qian P-Y, Beinart RA. Global
1010 16S rRNA diversity of provannid snail endosymbionts from Indo-Pacific deep-sea hydrothermal
1011 vents. Environ Microbiol Reports, 2021;in press.

1012 [74] Papot C, Massol F, Jollivet D, Tasiemski A. Antagonistic evolution of an antibiotic and its
1013 molecular chaperone: how to maintain a vital ectosymbiosis in a highly fluctuating habitat.
1014 Scientific Reports. 2017;7(1):1–14.

1015 [75] Swanson, WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nature Rev Genet.
1016 2002;3(2):137–44.

1017 [76] Chevaldonné, P, Desbruyères D, Le Haître M. Time-series of temperature from three deep-
1018 sea hydrothermal vent sites. Deep Sea Res Part A. 1991;38(11):1417–30.

1019 [77] Le Bris N, Gaill F. How does the annelid *Alvinella pompejana* deal with an extreme
1020 hydrothermal environment? Rev Environ Sci Biotechnol. 2006;6(1):197.

1021 [78] Gaill F, Wiedemann H, Mann K, Kühn K, Timpl R, Engel J. Molecular characterization of
1022 cuticle and interstitial collagens from worms collected at deep sea hydrothermal vents. J Mol Biol.
1023 1991;221(1):209–23.

1024 [79] Cary, SC, Cottrell MT, Stein JL, Camacho F, Desbruyères D. Molecular identification and
1025 localization of filamentous symbiotic bacteria associated with the hydrothermal vent annelid
1026 *Alvinella pompejana*. Appl Environ Microbiol. 1997;63(3):1124–30.

1027 [80] Phleger, CF, Nelson MM, Groce AK, Cary SC, Coyne K, Gibson JAE, Nichols PD. Lipid
1028 biomarkers of deep-sea hydrothermal vent polychaetes - *Alvinella pompejana*, *A. caudata*,
1029 *Paralvinella grasslei* and *Hesiolyra bergi*. Deep Sea Res Part I. 2005;52(12):2333–52.

1030 [81] Toulmond A, Slitine FEI, De Frescheville J, Jouin C. Extracellular hemoglobins of
1031 hydrothermal vent annelids: structural and functional characteristics in three alvinellid species.
1032 Biol Bull. 1990;179(3):366–73.

1033 [82] Jouin-Toulmond C, Augustin D, Desbruyères D, Toulmond A. The gas transfer system in
1034 alvinellids (Annelida Polychaeta, Terebellida). anatomy and ultrastructure of the anterior
1035 circulatory system and characterization of a coelomic, intracellular haemoglobin, Cah Biol
1036 Mar. 1996;37(2):135-152.

1037 [83] Chevaldonné P, Jollivet D, Desbruyères D, Lutz RA, Vrijenhoek RC. Sister-species of Eastern
1038 Pacific hydrothermal vent worms (Ampharetidae, Alvinellidae, Vestimentifera) provide new
1039 mitochondrial COI clock calibration. Cah Biol Mar. 2002;43(3–4):367–70.

1040 [84] Sicot F-X, Mesnage M, Masselot M, Exposito J-Y, Garrone R, Deutsch J, Gaill F. Molecular
1041 adaptation to an extreme environment: origin of the thermal stability of the Pompeii worm
1042 collagen. J Mol Biol. 2000;302(4):811–20.

1043 [85] Garrett SC, Rosenthal JJC. A role for A-to-I RNA editing in temperature adaptation.
1044 Physiology. 2012;27(6):362–69.

1045 [86] Pradillon F, Zbinden M, Mullineaux LS, Gaill F. Colonisation of newly-opened habitat by a
1046 pioneer species, *Alvinella pompejana* (Polychaeta: Alvinellidae), at East Pacific Rise vent
1047 sites. *Mar Ecol Prog Ser.* 2005;302:147-157.

1048 [87] Chevaldonné P, Jollivet D. Videoscopic study of deep-sea hydrothermal vent alvinellid
1049 polychaete populations: estimation of biomass and behaviour. *Mar Ecol Prog Ser.* 1993;95(3):251-
1050 262.

1051 [88] Lee PN, McFall-Ngai MJ, Callaerts P, de Couet HG. Preparation of genomic DNA from
1052 Hawaiian bobtail squid (*Euprymna scolopes*) tissue by cesium chloride gradient
1053 centrifugation. *Cold Spring Harbor Protocols.* 2009;11,pdb-prot5319.

1054 [89] Yao B, Zhao Y, Zhang H, Zhang M, Liu M, Liu H, Li J. Sequencing and *de novo* analysis of
1055 the Chinese Sika deer antler-tip transcriptome during the ossification stage using Illumina RNA-
1056 Seq technology. *Biotechnol lett.* 2012;34(5):813-822.

1057 [90] Gagnière N, Jollivet D, Boutet I, Brélivet Y, Busso D, Da Silva C, Gaill F, et al. Insights into
1058 metazoan evolution from *Alvinella pompejana* cDNAs. *BMC Genomics.* 2010;11(1):634.

1059 [91] Bankar KG, Todur VN, Shukla RN, Vasudevan M. Ameliorated *de novo* transcriptome
1060 assembly using Illumina paired end sequence data with Trinity Assembler. *Genomics*
1061 *data.* 2015;5:352-359.

1062 [92] Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol.*
1063 2007;24(8):1586–91.

1064 [93] Xu B, Yang Z. PamlX: a graphical user interface for PAML. *Mol Biol Evol.* 2013;30
1065 (12):2723–24.

1066 [94] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
1067 approach to multiple testing. *J Roy Stat Soc, ser B.* 1995;57(1):289–300.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupTable1Genesannotation.xlsx](#)
- [SuppTable2metrics.xlsx](#)
- [Supplementarytableslegends.docx](#)