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[1,2*], Denis Bertrand[3], Niranjan Nagarajan[3,4], Richard Copley[5], Erwan Corre[6], Stéphane Hourdez[7], Éric Bonnivard[1], Adam Claridge-Chang[8], Didier Jollivet[1]

[1] Sorbonne Université, CNRS, UMR 7144 AD2M, Station Biologique de Roscoff, Place Georges Teissier CS90074, 29688 Roscoff, France.

[2] Emlen Lab, Division of Biological Sciences, University of Montana, Missoula, MT, USA

[3] Genome Institute of Singapore, 60 Biopolis Street, #02-01 Genome, Singapore 138672, Singapore.

[4] National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077, Singapore.

[5] Sorbonne Université, CNRS, UMR 7009 LBDV, Laboratoire 'Evolution des génomes et des protéines animales', Observatoire océanologique, Quai de la Darse, 06234, Villefranche/mer, cedex, France.

[6] CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680, France.

[7] Sorbonne Université, CNRS UMR 8222, Laboratoire d'Ecogéochimie des Environnements Benthiques (LECOB), Observatoire Océanologique de Banyuls, Avenue du Fontaulé, 66650 Banyuls-sur-mer.

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

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

Keywords

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

Abstract

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

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

(96%) exhibiting high divergence. This congealing effect however seems to be linked to habitat specialization and still allows positive selection on genes involved in gamete recognition, as a possible long-duration process of species reinforcement.

Conclusion: 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 ligth on the evolution of gene divergence during the process of speciation and species specialization over a very long period of time.

Background

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

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

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

106 concept used (*e.g.*, complete reproductive isolation, hybrid counter-selection, phylogenetic
107 isolation with complete sorting of allelic lineages between populations). In the initial part of the
108 speciation process, individuals are likely to still exchange genes across semi-permeable barriers
109 resulting in heterogeneous migration rates along the genome but could also endure local
110 adaptation resulting in a loss of genetic diversity over some portions of their interacting genomes
111 [32].

112 Recent studies on the genetic architecture of reproductive isolation across the entire
113 genome often found a genome-wide heterogeneity of genetic differentiation between
114 populations, ecotypes or even well-separated morphological species, with highly divergent
115 regions of the genome contrasting with regions of lower divergence rather than a homogeneous
116 rate of differentiation across the genome [30, 33-34]. To this extent, genome scans of
117 differentiation and genome-wide linkage maps based on polymorphic sites at multiple loci across
118 the genome, represent powerful tools to address the genomic landscape of population/species
119 isolation. Regions displaying high levels of differentiation between lineages, so-called “genomic
120 islands of speciation” were logically assumed to exhibit loci underlying reproductive isolation
121 when the time elapsed since isolation was long enough to create genetic incompatibilities [35].
122 Most of the time, “islands of speciation” are believed to result from local barriers to gene flow at
123 some specific genes that rapidly led to allele fixation, gene hitchhiking and the subsequent
124 accumulation of divergence [36]. These islands differ from incidental islands of divergence
125 resulting from accelerated rates of lineage sorting within populations due to recurrent events of
126 either selective sweeps or background selection not necessarily related to reproductive isolation
127 [37]. The genomic island metaphor has proven popular and a wide array of studies searching for
128 “islands of speciation” in multiple taxa has been published over the last decade [30,34,35,38-42].

129 These studies identified outlier loci or regions that stand out from the distribution of the multi-
130 loci genetic differentiation (most often through F_{ST} scans) expected under models of migration
131 with or without isolation and heterogeneous rates of migration across genomes (*e.g.*, assortative
132 mating, hybrid counter-selection, divergent selection). Such outlier genes usually fall into three
133 categories: (1) outliers resulting from a differential effect of purifying selection in non-crossover
134 regions of the separating genomes, (2) outliers involved in Dobzhansky-Muller incompatibilities
135 that affect the fitness of hybrids/heterozygotes, and (3) outliers resulting from differential local
136 adaptation between the splitting lineages. In the last category, a limited number of alleles
137 involved in local adaptation have been found to carry a selective value strong enough to initiate a
138 barrier to gene flow primarily by divergent selection. When considering speciation from both
139 ends of the road, how do these patterns of divergence differ between the early and the late stages
140 of isolation?

141 To better understand the role of genomic architecture on speciation, the number, width,
142 and level of clustering of divergent regions within the genome of early incipient species are
143 usually compared with more divergent congeneric pairs. It is expected that early in speciation,
144 regions of divergence should be small and randomly scattered in the genome and possibly
145 restricted to a few non-recombining regions. These regions may correspond to the accumulation
146 of genetic incompatibilities or to loci under diversifying selection strong enough to overcome
147 gene flow which would be expected to homogenize the rest of the genome. Conversely, later in
148 speciation, when gene flow is greatly reduced between the two genomes by the cumulated effect
149 of multiple genetic barriers, divergent selection, physical linkage between targets of selection
150 and chromosomal rearrangements may create longer genomic regions of divergence. Such
151 processes acting over larger regions of the genome, followed by gene specialization within each

152 genome in response to local adaptation contribute to a genome-wide "congealing" (GWC) effect
153 that finalizes speciation with reproductive isolation [43-46]. GWC refers to the period along the
154 speciation continuum when the genomes of the splitting populations essentially become sealed
155 and when reproductive isolation becomes a property of the whole genome not just of a few loci
156 [43]. Therefore, incomplete lineage sorting and semi-permeable barriers to gene flow (associated
157 with the early stages of speciation) and the wide synergetic effects of genome congealing due to
158 chromosomal linkage between diverging genes (associated with the latest stages) are likely to
159 result in very distinct patterns of divergence at a given stage of the speciation process. Moreover,
160 positive selection over the genome due to gene specialization under local conditions should
161 affect KEGG pathways differently over the course of the splitting process. Although
162 fundamental to our understanding of speciation, such comparisons between the early and late
163 stages of speciation, and the time of transition between these two stages have rarely been done
164 when looking at the relationship between the number of diverging genes in association
165 (congealing effect) and their accumulated divergence (time since species separation).

166

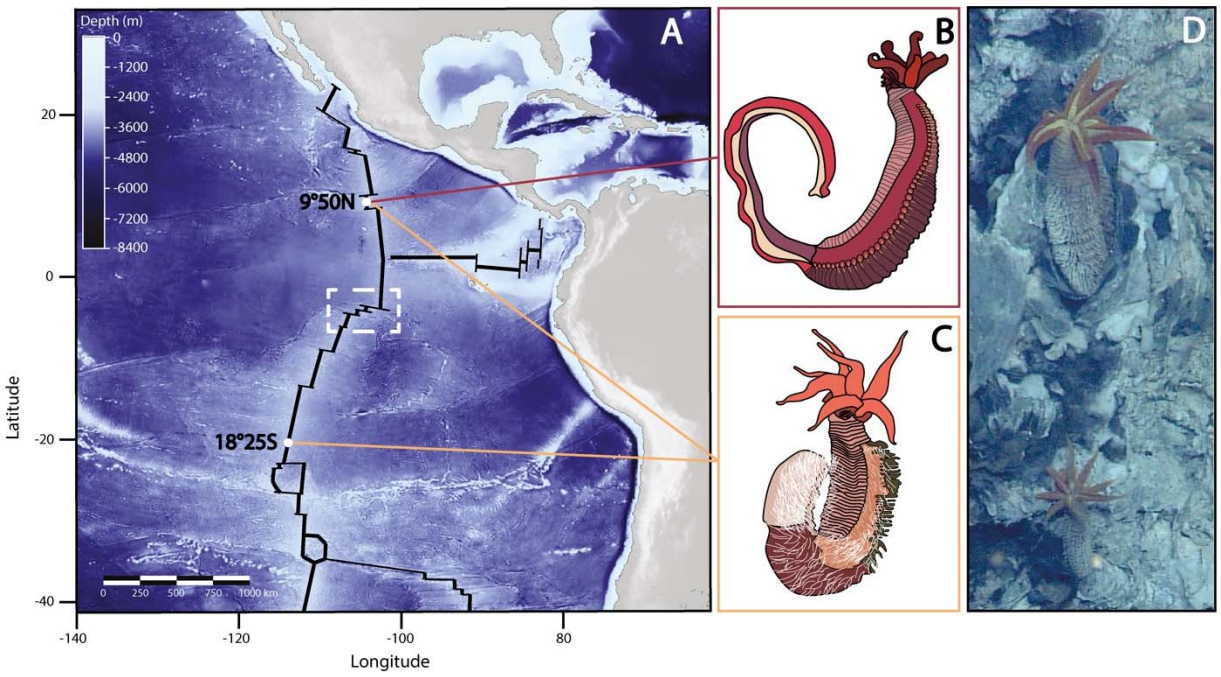


Figure 1: A) Bathymetric map of the East Pacific Ridge (in black), sampled populations sites along the ridge (white dots) and the geographical barrier between Southern and Northern populations (white dashed square); B) Drawing of a specimen of *Alvinella caudata;* C) Drawing of a specimen of *Alvinella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

hydrothermal chimneys along the East Pacific Rise (EPR) (Figure 1) [47-48]. Using transcriptome assemblies, we carefully examined the relative role of geographic isolation and habitat preference in shaping gene divergence between newly isolated populations (in allopatry) of *Alvinella pompejana* and between this latter species and its syntopic species *Alvinella caudata* (niche specialization & reinforcement in sympatry). The tube-dwelling worms *Alvinella pompejana* Desbruyères & Laubier 1980 [49] (Figure 1C and 1D) called the Pompeii worm and its sister species *A. caudata* Desbruyères & Laubier 1986 [50] (Figure 1B) live on the wall of still-hot chimneys of the EPR from 23°N (Guaymas basin) to 38°S (Pacific-Antarctic Ridge). The first description of these two worms considered them as two morphologically distinct ontogenic forms of a single species, one juvenile form with a long tail bearing filamentous campylobacteria on modified notopodia (corresponding to *A. caudata*) and the adult reproductive

form without tail but bearing the same epibionts on the whole dorsal part of its body (corresponding to *A. pompejana*). But enzyme polymorphism analyses led to the subdivision of this taxon into two distinct species without any shared electromorphs between them [51-52]. For both species, a strong physical barrier to gene flow was depicted at the Equator near 7°S/EPR and separates Northern and Southern vent populations in two allopatric groups of putative cryptic species with differentiation particularly marked in the Pompeii worm *Alvinella pompejana* [3,15]. Differences in the vent-site turn-over from both sides of the EPR and its subsequent effect on local hydrothermal vent conditions (*i.e.* changes in thermal conditions due to a highest proportion of still-hot chimneys and fluid chemistry heterogeneity at both local and regional scales) could represent favorable conditions for local adaptation and subsequent ecological filtering when migration is low or episodic across the barrier. In addition, depth gradient along the ridge or between ridges, extreme variations in the hydrothermal fluid composition and subsequent effects on symbiotic associations with divergent bacterial strains are likely to induce divergent selection between vent populations [53].

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

Results

Assembly and annotation of the reference genome and associated transcriptomes

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

Identification of orthologous genes

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

Distribution of genetic divergence and d_N/d_S values

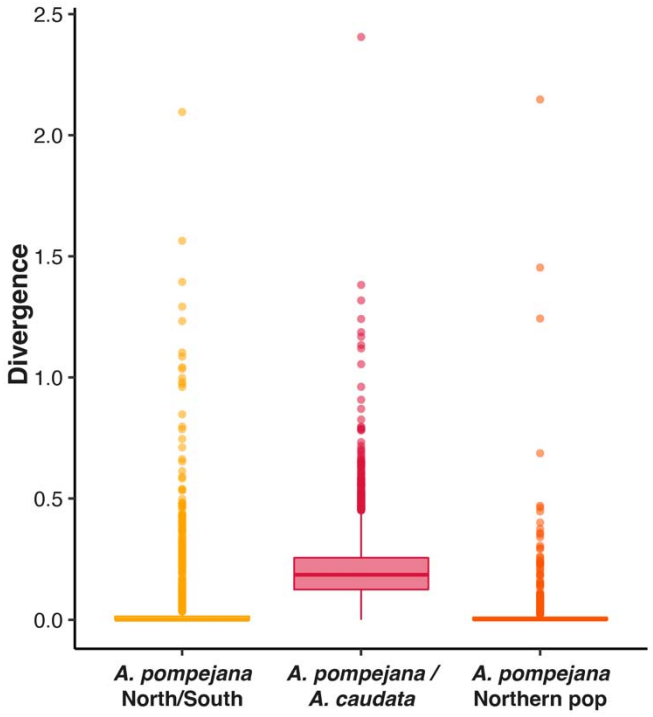


Figure 2: Boxplots of the distribution of the divergence values found in different stages of speciation in *Alvinella spp.* worms. The yellow color corresponds to the comparison *A. pompejana south/north*. The red color corresponds to the comparison *A. caudata/A. pompejana*. The orange color corresponds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

The evolutionary dynamics of genes during speciation was investigated by calculating pairwise synonymous to non-synonymous substitution rates and the associated divergence between orthologous transcripts of divergent individuals within and between closely-related species (see the "Methods" section). Among the 6,916 orthologous genes studied between the two Northern and Southern individuals of *A. pompejana*, 58% of the coding sequences are strictly identical with no nucleotide differences, 26% of the genes display divergences between 0.5% and 1.75% (0.01>t>0.05), and only 4% of the genes represent outliers that diverge by more than 2% (t>0.05) along their sequence (Figure 2). As expected in early speciation steps, the

overall mean divergence of genes is low (t=0.016) with values following a negative exponential distribution with a long tail of outliers. However, the distribution has a slight bimodal shape, the second unexpected mode being centered around a value ten times greater (i.e., about 0.020 substitution per codon) (Figure 2). This bimodal shape was not encountered when assessing the intra-specific allelic diversity between two Northern individuals (average diversity = 0.0096 ± 0,0051 with fewer outliers (t>0.05 =2%), see Figure 2). Analysis of both transition and transversion rates between individuals from North and South were 4 times greater than those found by comparing two distinct individuals of the same population in the northern part of the EPR (Figure 3).

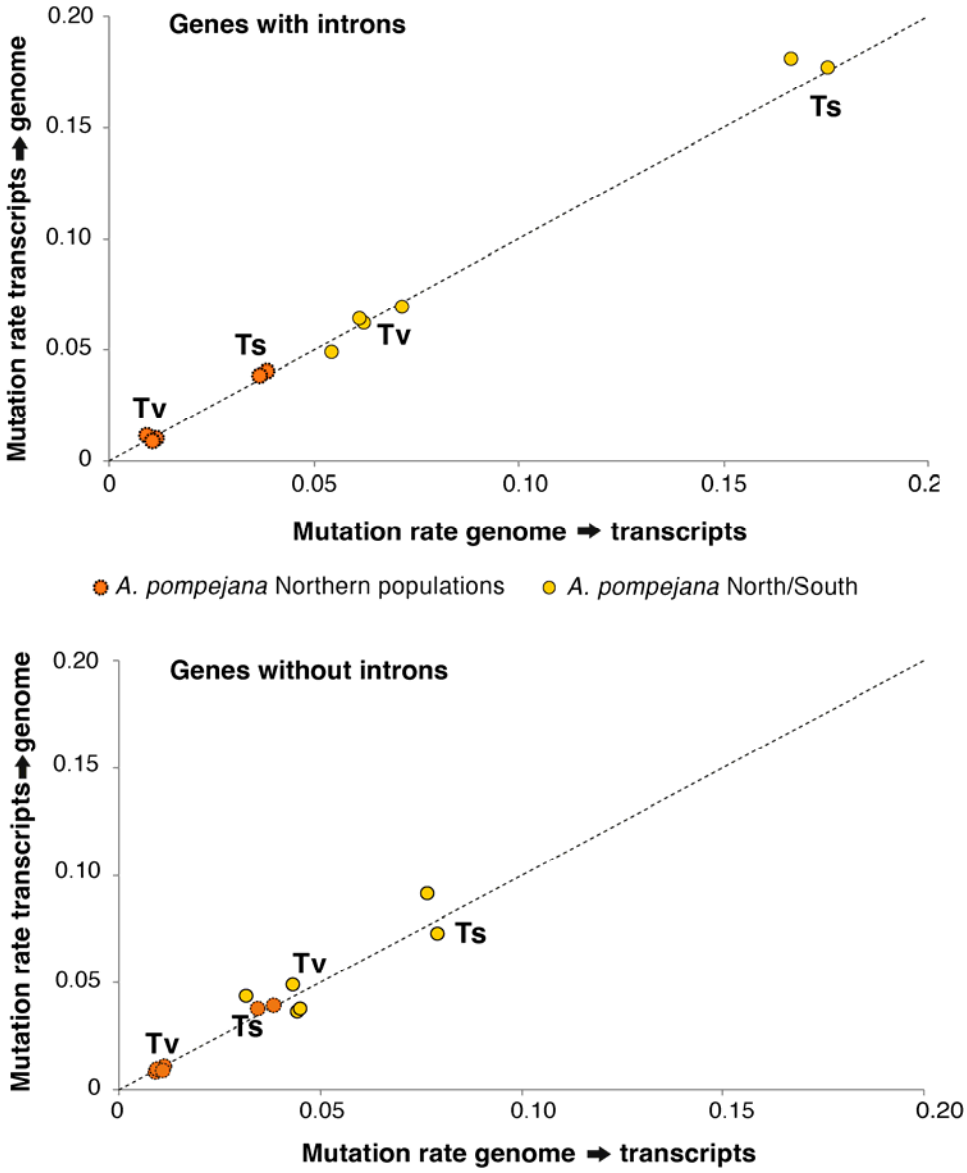
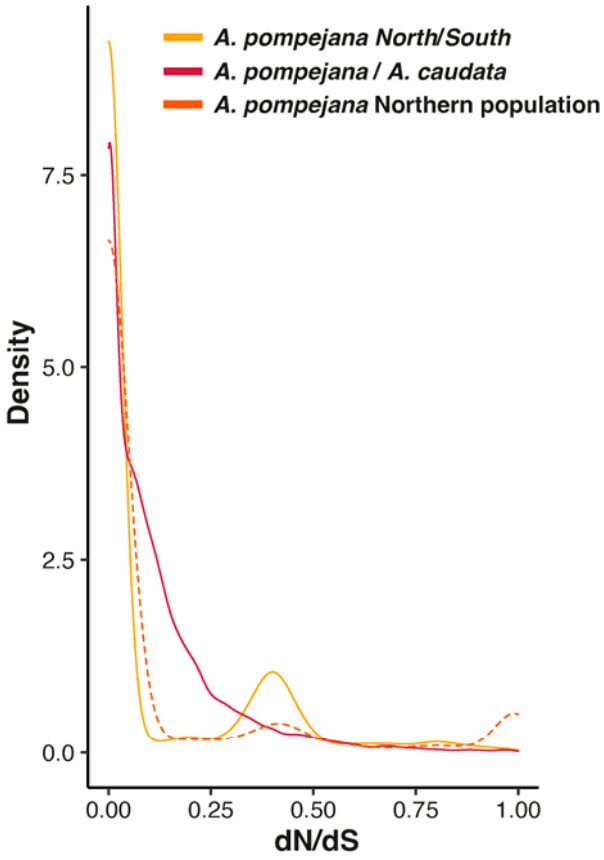


Figure 3: Transition and translation rates obtained between orthologous pairs of genes of *Alvinella pompejana* by comparing single-individual transcriptomes from the north (9°50N) and south (18°25S) EPR onto the genome of the worm (northern individual). Rates were obtained from coding sequences obtained 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).

255 On the opposite, the large majority of the 6,687 genes are highly divergent between *A.*
256 *pompejana* and *A. caudata*. Divergence follows a Gaussian distribution centered around 0.2
257 substitution per codon with an asymmetric tail of outliers (Figure 2). Only 0.4% of the sequences
258 are strictly identical with no nucleotide differences corresponding to only 33 pairwise
259 alignments, 3% of the genes display divergence levels between 10% and 25% (0.25>t>0.75), and
260 96% of the genes diverge by more than 25% (t>0.75) along their sequence. Thus, the distribution
261 of the divergence between *A.caudata* and the 2 populations of *A.pompejana* is significantly
262 different (Figure 2; Wilcoxon test, p < 2.2e-16), but this distribution is not significantly different
263 within *A. pompejana* populations (Wilcoxon test, p = 0.8266).



264

Figure 4: Density distributions of d_N/d_S values estimated for each pairwise alignment of orthologous
265 genes in both the early and late states of speciation in *Alvinella spp*. worms using the method of Nielsen &
266 Yang (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.

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

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

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

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

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

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

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

Patterns of nucleotide substitutions and highly divergent genes

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

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

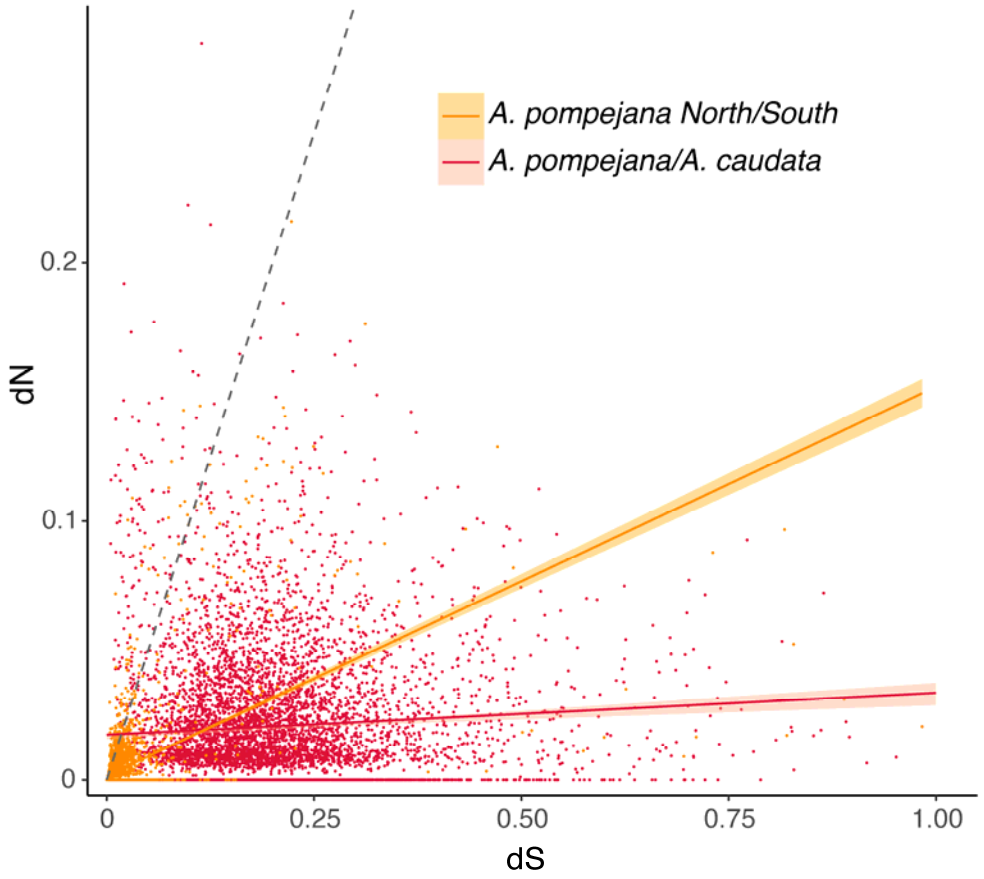


Figure 5: Rates of synonymous and non-synonymous divergences between *A. pompejana* North and South populations (yellow) and *A. pompejana* and *A. caudata* (red). The grey dashed line represents the expected linear relationship between d_S and d_N under neutral evolution.

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

(peroxidase, peroxiredoxin 4, glutathione peroxidase, Hsp70, Hsp83), DNA repairs and methylation (ATP-dependent RNA helicase, DNA ligase1, RNA polymerase II transcription subunit, adenosylhomocysteinase B, BRCA1-associated RING domain protein1, histone deacetylase3), mRNA A-to-I editing (Double-stranded RNA-specific adenosine deaminase, xanthine dehydrogenase), protein glycosylation (Protein O-linked-mannose beta-1.2-N-acetylglucosaminyltransferase1), adaptation to hypoxia (Hypoxia-inducible factor1, carbonic anhydrase) and proteins involved in protein homo- heterodimerization (complement factor H-related protein2, BTB/POZ domain-containing protein7).

Distribution of divergence and associated d_N/d_S along the genome

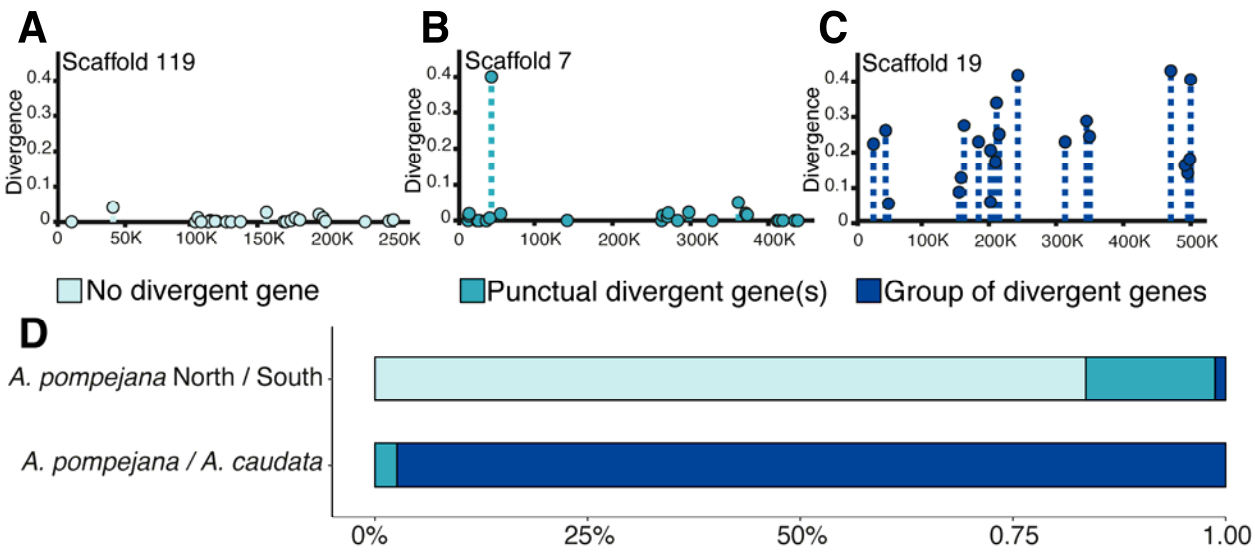


Figure 6: Genomic distribution of the divergence along the longest scaffolds (>400,000bp) *A. pompejana* draft genome (subset of genes representing about 1/10 of the genome). According to the index of distribution and our criterion of selection by eye, three distribution patterns of the divergence are 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).

387 The distribution of gene divergence was examined over the 79 scaffolds exceeding
388 300,000 bp with the *A. pompejana* North/South comparison (Figure 6D). Because the global
389 level of divergence is low and close to zero, values of divergence are homogeneously distributed
390 with a few punctual observable outliers. As a consequence, about 83% of the examined scaffolds
391 have no divergent genes (Figure 6A and 6D). Scaffolds containing only one divergent gene
392 (Figure 6B and 6D) account for 15% of the examined portion of the genome and only one of the
393 largest scaffolds displays an island of divergence pattern, and may be linked to sexual
394 chromosome (Figure 6C and 6D). On the contrary, for the *A.caudata/A.pompejana* comparison,
395 divergence is clearly structured along the genome in large islands of highly divergent genes for
396 97% of the scaffolds (Figure 6D). Each of the 76 scaffolds only carries very few genes with no
397 divergence and we were not able to identify any region or island of no divergence among the
398 examined scaffolds. Here, divergence is clearly structured in large blocks where all the genes and
399 their putative exons are divergent. The average values of d_N/d_S per scaffold are globally
400 homogeneous between scaffolds but the dispersion index of the d_N/d_S values around this mean
401 fluctuates substantially. Gene density among scaffold however greatly varies and therefore
402 impacts the number of genes involved in the genomes congealing.
403

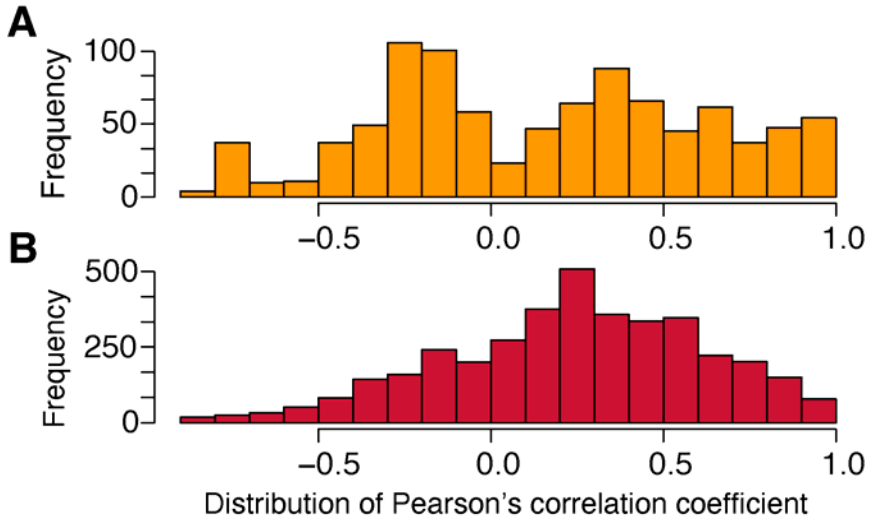


Figure 7: Distribution of the Pearson's correlation coefficient between divergence and associated dN/dS values calculated for each scaffold containing more than 5 genes: (A) *A. pompejana* North/South in yellow and (B) *A. pompejana / A. caudata* in red.

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

421 are highly divergent, this overall increase of d_N/d_S with divergence therefore suggests that most
422 of the genes have evolved to adapt to a new ecological niche after the separation of the two sister
423 species.
424
425 ## Discussion
426
427 Considering the speciation continuum from one end to the other is essential to shed light
428 on the mechanisms by which genomes separate, and more specifically the timing of reproductive
429 isolation and species specialization. In this study, we first examined an early stage of allopatric
430 speciation between populations of the Pompeii worm, *Alvinella pompejana* geographically
431 separated by a physical barrier to gene flow at the Equatorial triple junction of the East Pacific
432 Rise and the Galapagos rift [3,17]. This geographic separation of the Pompeii worm populations
433 is also associated with different venting dynamics related to different spreading rates of the ridge
434 segments that may play a crucial role in the proportion of still-hot chimneys and colder vent
435 habitats and thus the distribution of the microbiota/biota [8,56]. Then we also examined the
436 molecular evolution signature associated with a much more advanced stage of the speciation
437 process that separated *A. pompejana* and its syntopic sister species *A. caudata*, which exhibits a
438 slightly different niche. This event is relatively old and possibly predates the separation of the
439 EPR and the North East Pacific ridge (JdF) communities by vicariance that led for example to
440 the separation of sibling species like *Paralvinella pandorae pandorae* and *P. p. irlandei*, about
441 28 Mya [1]. Mitochondrial and nuclear divergences estimated between *A. pompejana* and *A.*
442 *caudata* are indeed greater than those estimated for these latter species (supplementary data in
443 [57]).

444

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

460

461 During the early steps of speciation, the large majority of genes (about 70%) are stacked
462 around null values of d_N and d_S but the relationship between the two rates for the remaining
463 genes is on average 5:1, one non synonymous substitution for five synonymous ones, probably
464 due to unfixed mutations in the worm's polymorphism. Allele divergence and associated d_N/d_S
465 distributions between (North vs South individuals) and within (North individuals) the Pompeii
466 worm's populations are indeed quite similar although a greater number of higher values has been

observed in the first comparison. This can be partially explained by the fact that the Equatorial barrier to gene flow is not completely sealed [15] allowing rare episodes of allele introgression. On the opposite, the slope of the regression line between d_N and d_S is close to zero during the late steps of speciation and clearly highlights a large accumulation of synonymous substitutions close to saturation for a large number of genes. This strongly suggests that even if a great number of non-synonymous substitutions are positively-selected, most of this signal is likely to have been erased in fast-evolving genes. Outlier genes associated with high evolutionary rate (d_S>0.5) should thus represent fast-evolving genes that may have endured duplications, selective losses or gains of duplicates according to habitat/geographic isolation during the early steps of speciation, but also genes that may have already lost their 'adaptive' signal in the face of saturation at synonymous sites during the late steps of speciation.

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

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

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

513 substitutions (and hence associated genetic incompatibilities) between the two isolated entities
514 represents 15560 substitutions in coding sequences (as the product of 2,075 300-codons genes
515 with a divergence of 0.025 substitutions per codon (30% of the investigated genes)) for
516 approximately 1/10 of the whole genome. Compared to other eukaryotic models (*e.g.*, about
517 70,000 changes accumulated between sub-species of the *Drosophila simulans* complex over 0.25
518 My [65] for about the same number of coding sequences), the number of accumulated changes
519 (i.e. ~15,560) is quite low if the two populations separated 1.5 Mya [15], even if the generation
520 time may be about 10 times greater in *Drosophila*. Such discrepancy might be explained by a
521 strong stabilizing selection associated with these extreme but constant environmental conditions
522 over space and time. Gene networks models indeed predict that the hybrid incompatibility
523 dynamics may be greatly reduced by stabilizing selection while producing a “basin of attraction”
524 for the optimal genotype on both sides of the barrier to gene flow [66].
525
526 In this early stage of speciation, very few genes showed evidence of positive selection
527 and these genes do not relate to genes currently depicted as involved in reproductive isolation
528 (*e.g.*, genes involved in gamete recognition, pheromones, mito-nuclear incompatibilities, hybrid
529 sterility or immune incompatibility [67-70]. Neither sperm-egg binding or egg-fusion proteins
530 nor nuclear proteins transferred to the mitochondria displayed signatures of positive selection or
531 strong divergence. However, almost all genes showing high evolution rates in our data encode
532 for tubulins, which represents highly diversified multigene families and can have a great
533 influence on cell division (mitosis/meiosis) and the spermatozoon flagellum architecture (which
534 is extremely reduced as a byproduct of the internal fertilization of oocytes in *Alvinella* [71]. It
535 should also be noted that a meaningful proportion of positively-selected genes could indirectly

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

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

559

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

582 time of genome congealing around 26.5 Mya by using the substitution rate previously estimated
583 by Chevaldonné *et al.* [83]. The first mapping of divergence and associated d_N/d_S showed
584 however a strong heterogeneity of these variables along the genome with some scaffolds having
585 relatively higher divergence values. Although we cannot rule out that these scaffolds could
586 represent either regions of lower recombination rates or a portion of sex-linked chromosome,
587 their high divergence could also trace back to the early steps of the speciation process (i.e.
588 primary islands of speciation), in this case suggesting that the separation of the two species may
589 have been initiated much more earlier.
590
591 Gene divergence is also slightly positively correlated with d_N/d_S within each scaffold.
592 This therefore suggests that the increase in gene divergence have been accompanied by the
593 accumulation of non-synonymous mutations. Assuming that d_N/d_S only increase when
594 synonymous sites become saturated, this trend could either indicate an adaptive divergence
595 associated with a long period of gene specialization for the two sister species regarding their
596 habitat after their separation, or that the time since species separation is so high that most genes
597 reached saturation but this later interpretation does not fit well with the presence of positively-
598 selected genes. The genes under positive selection ($d_N/d_S > 1$) are more specifically involved in
599 the repair/replication/biosynthesis of nucleic acids but also in several metabolic pathways usually
600 involved in the stress response, carbohydrate catabolism and lipid/steroid metabolisms associated
601 with membrane organization and tegument formation. Targets of positive selection seem to be
602 consistent with gene specialization in response to different thermo-chemical regimes encountered
603 by the worms, and more specifically high temperatures, natural doses of radioactivity and high
604 concentrations of sulfides and heavy metals producing reactive oxygen species (ROS) and

reactive sulfur species (RSS), and suggest that the two species do not share exactly the same niche. Annotation of fast evolving genes also revealed that most of them belonged to exactly the same metabolic pathways impacted by positive selection including membrane/tegument organization. Regarding fast evolving genes, we noted a net enrichment in proteins involved in oxidative stress response (peroxidase, Hsps), adaptation to hypoxia, DNA repairs and methylation, mRNA A-to-I editing, and protein glycolysation. The two latter molecular processes represent a powerful way to adapt to differing thermal regimes [84-85]. These results therefore indicate that the adaptive evolution of the two species has been a sufficiently 'old' process to erase most signatures of positive selection that has likely played a crucial role in their ecological isolation. As opposed to *A. caudata*, *A. pompejana* is one of the first colonizer of newly-formed and still hot chimneys [86] when these edifices are made of porous anhydrite (barium/calcium sulfates) with temperatures above 100°C. Consistently, slight differences in terms of protein composition with more charged residues in *A. pompejana* and more hydrophobic/aromatic ones in *A. caudata* have been described [57]. Interestingly, when looking at the proportion of each species between chimneys at a local spatial scale (*i.e.*, hundreds of meters), it can be observed that newly-formed anhydrite edifices and 'white smokers' are mostly inhabited by *A. pompejana*, whereas older edifices and 'black smokers' are mainly characterized by a greater abundance of *A. caudata* [7]. Accelerated differentiation of genes involved in the adaptive response of the worms to environmental variations are likely to explain their different ecologies and behaviors to cope with 'high' temperatures. For instance, tubes of the two species are closely entangled in colonies but with differences in their opening [72] (D. Jollivet, *pers. obs.*). While the aperture of the tube of *A. pompejana* is widely opened like a funnel with septa, that of *A. caudata* is ended by a small hole. The shape of the tubes is likely related to the

628 thermoregulatory behavior of the worms that helps individuals to refresh their microenvironment
629 through water pumping [77,87]. While *A. pompejana* is widely observed moving inside and
630 outside its tube to refresh it with the cold surrounding waters, *A. caudata* usually stays in its tube
631 but exhibits a tail without organs used to cultivate its epibiotic flora that may be viewed as a
632 thermal sensor to position the worm at exactly the right temperature.
633
634 Despite this very long history of speciation and subsequent species specialization, one of
635 the most striking observations we made is that many genes involved in gamete recognition (*e.g.*,
636 zonadhesin, innexin) are under positive selection, suggesting that species sympatry may still play
637 a role in species reinforcement. These positively-selected proteins are accompanied by several
638 fast-evolving proteins involved in spermatogenesis, and egg-sperm fusion (*e.g.*, dynein, integrin,
639 disintegrin, protocadherin), and embryonic development (*e.g.*, methenyltetrahydrofolate
640 synthase). Moreover, specific pathways such as the Acetyl-CoA cycle (3 targets) together with a
641 lipid desaturase seems to become highly differentiated. In *Drosophila*, desaturase genes and the
642 Acetyl-CoA cycle are both involved in the production of cuticular hydrocarbons (CHCs) which
643 have a crucial role in adaptation to desiccation and also act as pheromones involved in mating
644 behavior [70]. It is therefore possible that these two worm species preliminary evolved in a
645 different geographical context for a long period of time and then met secondarily during the
646 recolonization of the EPR after the episodes of tectonic isolation that structured the hydrothermal
647 fauna between 11-18 Mya as recently suggested to explain the phylogeographic patterns of the
648 *Lepetodrilus elevatus* complex of limpet species [4]. This rather late evolution of prezygotic
649 mechanisms of isolation contrasts with the allopatric situation encountered between the northern

650 and southern populations of *A. pompejana* and supposes that these mechanisms could last for a
651 very long time even after a genetic barrier has been erected.
652
653

654 Conclusion

655 Our analyses pointed out the non-negligible role of natural selection on both the early and
656 late stages of speciation in the emblematic thermophilic worms living on the walls of deep-sea
657 hydrothermal chimneys. They shed ligth on the evolution of gene divergence during the process
658 of speciation and species specialization over time. Due to habitat fragmentation, populations
659 separate and start to accumulate putative adaptive mutations (and possibly genetic
660 incompatibilities) in allopatry where they find differing local conditions but are likely to
661 consolidate reproductive isolation in sympatry following secondary contacts during the
662 specialization process, even after genomes diverged almost completely (congealing effet). These
663 first analyses raise many further questions about the evolutionary mechanisms that led to the
664 speciation of *Alvinella* spp. and their subsequent distribution within the spatial micro-mosaic of
665 habitats typifying hydrothermal vents. Additional studies combining polymorphism and
666 divergence are needed to better understand the respective roles of geographic and ecological
667 histories of the worm's populations in speciation in a such fragmented and instable environment.
668

669 Material and methods

670

671 1) Animal sampling and genome/transcriptome sequencing

672 The genome of the Pompeii worm was assembled from a single individual collected at
673 9°50N/EPR. Several DNA purification methods using alcohol precipitation and/or column
674 purification resulted in DNA samples that could not be efffectively digested by restriction
675 enzymes. To reduce mucopolysaccharides or similar co-precipitating contaminants, we purified
676 the DNA from worm tissue with phenol extraction and isopropanol precipitation followed by a
677 clean-up step with cesium chloride gradient ultracentrifugation [88]. The DNA was used to
678 prepare paired-end shotgun and 5 kb mate-pair libraries for Illumina genome sequencing [89].
679
680 We used at least two additional individuals of *A. pompejana* and *A. caudata* coming from the
681 same locality to perform RNA sequencing and the subsequent assembly of transcriptomes for the
682 northern specimens of the two species [57]. In parallel, a Sanger sequencing project of a
683 reference and fully annotated transcriptome was performed using several individuals coming
684 from a single chimney of the hydrothermal vent field 18°25S [90], for additional information
685 about this annotated transcriptome) and, represented the third transcriptome used to characterize
686 the specimens of *A. pompejana* located further south to the Equatorial barrier to gene flow
687 previously described [3,15].
688
689 **2) Sequencing** (see Metrics in Supplementary Table 2)
690
691 • Genome and transcriptomes
692 Genome of the Pompeii worm (about 370 Mb) was sequenced from a single individual using
693 both a shotgun and mate-pair 5 kb sequencing using the Illumina HiSeq2500 technology at the
694 A*Star Molecular and Cell Biology Institute (*Alvinella* genome project: coord. A. Claridge-

Chang). Paired-end reads libraries (insert size 300bp) were assembled using SOAPdenovo2 (v2.0, -K 41 -d -R), contigs were then scaffolded with OPERA (v1.3.1) using both paired-end (insert size 300bp) and mate pair libraries (insert size 5 kb). Finally, the assembly was gapfilled with Gapcloser (v1.12) to subsequently produce about 3000 scaffolds with a size greater than 15 kb.

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

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

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

- Megablast

Because coding sequences are compared between divergent individuals within a given species or between closely-related species, we performed a sequence-similarity search using each collection

718 of transcripts obtained for *A. caudata* and the two geographic forms of *A. pompejana* against the
719 entire *Alvinella pompejana's* genome with the software MegaBLAST implemented in NCBI
720 BLAST+ 2.2.30 nucleotide-nucleotide search, optimized for very similar sequences (set
721 expectation value cutoff = 10^{-15}, maximum hits to show = 1).
722
723 • Paralog
724 The 24-columns MegaBLAST output was then filtered using a custom-made Perl script (called
725 "Paralog") to discriminate genes according to their position and occurrences in the genome. This
726 allowed us to rule out putative duplicated genes in our subsequent analyses. Based on the
727 positions of the hits on both scaffolds and transcripts, genes were separated into distinct
728 categories, namely scaffold paralog, tandemly repeated paralogs, allelic forms of a single gene in
729 either genes with or without introns. Scaffold paralogs correspond to coding sequences found on
730 distinct scaffolds, tandemly-repeated genes to sequences found at non-overlapping positions onto
731 the same scaffold, allelic forms of a gene corresponded to several transcripts matching exactly
732 the same positions onto a scaffold, and the two others categories were orthologous transcripts
733 only found once into the genome at a given scaffold position with a discrimination exon/uniq
734 when the Megablast hits corresponded or not to a succession of fragments with different
735 positions both onto the scaffold and the transcript (i.e. exonic regions). Paired orthologous
736 sequences obtained in the two last categories (exon and uniq) were then used to estimate
737 divergence and associated d_N/d_S ratios in the pipeline.
738
739 • Search for coding DNA sequences (CDS) in pairwise alignments

Nucleic acid sequences were translated in the six reading frames to select for the longest CDS between two subsequent stop codons. Coding sequences containing gaps, undetermined nucleic acids (N) or alternative frames without stop were removed. To avoid false positives and reduce the risk of calculating erroneous d_N/d_S values from non-coding or incorrectly framed sequences, we performed a sequence-similarity search against the UniProt database using BlastP from NCBI BLAST+ 2.2.30 (set expectation value cutoff = 10^{-10}, maximum hits to show = 1). Sequences with no annotation from UniProt with a CDS shorter than 300 nucleotides were removed from the analysis.

- Divergence and d_N/d_S calculation

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

762 To avoid any potential effect of greater divergence at non-synonymous sites possibly due to
763 sequencing errors when values of neutral divergence (d_S) departed from zero, the significance of
764 all d_N/d_S values greater than zero was tested using a custom-made pipeline. For that purpose, we
765 produced 1,000 random alignments of each pair of coding sequences using a bootstrap
766 resampling of codons (from the bootstrap option of codeML in PamlX 1.3.1). We then estimated
767 the associated d_N and d_S for each resampled alignment, and calculated the difference (D) between
768 each d_N and d_S resampled values in order to test whether the observed value of this difference
769 was significantly greater than zero (*i.e.*, fall outside the distribution of the pseudo-replicates). To
770 this end, we performed a unidirectional one-sample *z*-test for each resampled set of paired
771 sequences and adjusted p-values over the whole set of genes using the false discovery rate
772 correction representing the expected proportion of false positives [94].
773
774 **4) Identification of islands of divergence and/or adaptation**
775
776 To investigate the genomic architecture of molecular divergence of putative genes under
777 strong purifying (d_N/d_S close to zero) or positive selection (d_N/d_S >1) along chromosomes, both
778 the divergence and their d_N/d_S associated values were positioned along the longest scaffolds
779 (>300,000 bp with more than 5 genes per scaffold) of the draft genome of *A. pompejana* for each
780 pairwise comparisons (*i.e.*, AP north vs AP south, and AP vs AC). According to the gene
781 distribution along each scaffold, three different types of patterns were characterized depending
782 on the number and level of clustering of divergent genes (divergence > 0.1) among scaffolds: (1)
783 scaffolds without divergent genes, (2) scaffolds with one or two isolated divergent genes and, (3)
784 scaffolds carrying a genes cluster of divergence corresponding to a putative island of speciation.

The proportion of each type was then calculated for the two sets of analyses. In addition, we calculated the mean scaffold distance (d) between two divergent genes and its standard deviation ($s^2(d)$) within each scaffold. We then tested for contiguous distribution patterns of divergence along each scaffold thanks to the Fisher's index of dispersion ($s^2(d)/mean(d)$). Divergent genes are randomly distributed when the index is equal to one, over-dispersed for values lower than one and aggregated for values greater than one.

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

Abbreviations

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

Acknowlegments

We would like to thank the captain and crews of the research vessel L'Atalante and the pilots and associated teams of the manned submersible Nautile without which no specimen would have been sampled. We are also very grateful to the chief-scientists of the oceanographic cruises Biospeedo 2004 and Mescal 2010 for collecting *Alvinella* specimens and to S.C. Cary who kindly provided in 2006 the AP specimen used to produce the genome. We thank Romain P.

Boisseau for considerable feedback on the manuscript and Ulysse Guyet for useful discussions on the statistical analyses.

Authors' contributions

CTB and DJ performed analyses and wrote the manuscript. EB and DJ conceptualized, designed and co-supervised the study. SH and DJ were in charge of the funded research project. ACC supervised the AP genome project. DB, NN, RC and EC were involved in the AP genome assembly, transcriptome assembly and mapping, and gene annotation. DJ, SH, and ACC edited and contributed to the proofreading of the manuscript. All authors read and approved the final manuscript.

Funding

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

Availability of data and materials

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All the authors declare no competing interests.

Authors details

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

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

References

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

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

[3] Plouviez S, Shank TM, Faure B, Daguin□Thiebaut C, Viard F, Lallier FH, Jollivet D. Comparative phylogeography among hydrothermal vent species along the East Pacific Rise

854 reveals vicariant processes and population expansion in the South. Mol. Ecol.
855 2009;18(18):3903–17.
856 [4] Matabos M, Jollivet D. Revisiting the *Lepetodrilus elevatus* species complex
857 (Vetigastropoda: Lepetodrilidae), using samples from the Galápagos and Guaymas hydrothermal
858 vent systems. J Moll Stud. 2019;85(1):154–65.
859 [5] Lutz RA. Dispersal of organisms at deep-sea hydrothermal vents: a review. Oceanol Acta.
860 Special Issue. 1988;8:23-30.
861 [6] Tyler PA, Young CM. Reproduction and dispersal at vents and cold seeps. J mar Biol Ass
862 UK. 1999;79(2):193-208.
863 [7] Jollivet D. Specific and genetic diversity at deep-sea hydrothermal vents: an overview.
864 Biodiv Conserv. 1996;5(12):1619–53.
865 [8] Vrijenhoek RC. Gene flow and genetic diversity in naturally fragmented metapopulations of
866 deep-sea hydrothermal vent animals. J Hered. 1997;88(4):285–93.
867 [9] Jollivet D, Chevaldonné P, Planque B. Hydrothermal-vent alvinellid polychaete dispersal in
868 the Eastern Pacific. 2. a metapopulation model based on habitat shifts. Evolution.
869 1999;53(4):1128–42.
870 [10] Johnson KS, Childress JJ, Beehler CL. Short-term temperature variability in the Rose
871 Garden hydrothermal vent field: an unstable deep-sea environment. Deep Sea Res Part
872 A. 1988;35(10-11):1711-21.
873 [11] Le Bris N, Sarradin PM, Caprais JC. Contrasted sulphide chemistries in the environment of
874 13°N EPR vent fauna. Deep Sea Res Part I. 2003;50(6):737-47.

875 [12] Matabos M, Le Bris N, Pendlebury S, Thiébaut E. Role of physico-chemical environment
876 on gastropod assemblages at hydrothermal vents on the East Pacific Rise (13°N/EPR). J mar Biol
877 Ass UK. 2008;88(5):995–1008.
878 [13] O'Mullan GD, Maas PAY, Lutz RA, Vrijenhoek RC. A hybrid zone between hydrothermal
879 vent mussels (Bivalvia: Mytilidae) from the Mid-Atlantic Ridge. Mol Ecol. 2001;10(12):2819-
880 31.
881 [14] Johnson SB, Won Y-J, Harvey JBJ, Vrijenhoek RC. A hybrid zone between *Bathymodiolus*
882 mussel lineages from Eastern Pacific hydrothermal vents. BMC Evol Biol. 2013;13(1):21.
883 [15] Plouviez S, Le Guen D, Lecompte O, Lallier FH, Jollivet D. Determining gene flow and the
884 influence of selection across the equatorial barrier of the East Pacific Rise in the tube-dwelling
885 polychaete *Alvinella pompejana*. BMC Evol Biol 2010;10(1):220.
886 [16] Plouviez, S, Faure B, Le Guen D, Lallier FH, Bierne N, Jollivet D. A new barrier to
887 dispersal trapped old genetic clines that escaped the Easter Microplate tension zone of the Pacific
888 vent mussels. PLoS One. 2013;8(12):e81555.
889 [17] Matabos M, Plouviez S, Hourdez S, Desbruyères D, Legendre P, Warén A, Jollivet D,
890 Thiébaut E. Faunal changes and geographic crypticism indicate the occurrence of a
891 biogeographic transition zone along the Southern East Pacific Rise. J Biogeogr. 2011;38(3):575–
892 94.
893 [18] Won Y-J, Hallam SJ, O'Mullan GD, Vrijenhoek RC. Cytonuclear disequilibrium in a hybrid
894 zone involving deep-sea hydrothermal vent mussels of the genus *Bathymodiolus*. Mol Ecol.
895 2003;12(11):3185–90.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

940 [39] Hofer T, Foll M, Excoffier L. Evolutionary forces shaping genomic islands of population
941 differentiation in humans. BMC Genomics. 2012;13(1):107.
942 [40] Larson WA, Limborg MT, McKinney GJ, Schindler DE, Seeb JE, Seeb LW. Genomic
943 islands of divergence linked to ecotypic variation in Sockeye salmon. Mol Ecol.
944 2017;26(2):554–70.
945 [41] Duranton M, Allal F, Fraïsse C, Bierne N, Bonhomme F, Gagnaire P-A. The origin and
946 remolding of genomic islands of differentiation in the European sea bass. Nature. Comm.
947 2018;9(1):1-11.
948 [42] Tavares H, Whibley A, Field DL, Bradley D, Couchman M, Copsey L, Elleouet J, et al.
949 Selection and gene flow shape genomic islands that control floral guides. Proc Nat Acad Sci
950 USA. 2018;*115*(43):11006-11011.
951 [43] Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends Genet.
952 2012;28(7):342–50.
953 [44] Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. Genome-Wide
954 congealing and rapid transitions across the speciation continuum during speciation with gene
955 flow. J Hered. 2014;105(S1):810–20.
956 [45] Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the influence of
957 genomic architecture on the dynamics of speciation. Mol Ecol. 2014;23(16):4074–88.
958 [46] Tittes S, Kane NC. The genomics of adaptation, divergence and speciation: a congealing
959 theory. Mol Ecol. 2014;23(16):3938–40.
960 [47] Chevaldonné P, Desbruyères D, Childress JJ. ... and some even hotter. Nature.
961 1992;359(6396):593-594.

962 [48] Cary SC, Shank T, Stein J. Worms bask in extreme
963 temperatures. Nature. 1998;391(6667):545-546.
964 [49] Desbruyères, D, Laubier L. *Alvinella pompejana* gen. sp. nov., ampharetidae aberrant des
965 sources hydrothermales de la ride Est-Pacifique. Oceanol Acta. 1980;3(3):267-274.
966 [50] Desbruyères, D, Laubier L. Les Alvinellidae, une famille nouvelle d'annélides polychètes
967 inféodées aux sources hydrothermales sous-marines: systématique, biologie et écologie. Can J
968 Zool 1986;64(10):2227–45.
969 [51] Autem M, Salvidio S, Pasteur N, Desbruyères D, Laubier L. Mise en évidence de
970 l'isolement génétique de deux formes sympatriques d'*Alvinella pompejana* (Polychaeta:
971 Ampharetidae), annélides inféodées aux sites hydrothermaux actifs de la dorsale du Pacifique
972 Oriental. C R Acad Sci Paris. 1985;301:131–35.
973 [52] Jollivet D, Desbruyères D, Ladrat C, Laubier L. Evidence for differences in the allozyme
974 thermostability of deep-sea hydrothermal vent polychaetes (Alvinellidae): a possible selection by
975 habitat. Mar Ecol Prog Ser. 1995;123(1–3):125–136.
976 [53] Faure B, Jollivet D, Tanguy A, Bonhomme F, Bierne N. Speciation in the deep sea: multi-
977 locus analysis of divergence and gene flow between two hybridizing species of hydrothermal
978 vent mussels. PLoS One. 2009;4(8):e6485.
979 [54] Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of
980 molecular evolution. Nature. 1977; 267(5608):275–76.
981 [55] Stoletzki N, Eyre-Walker A. Synonymous codon usage in *Escherichia coli*: selection for
982 translational accuracy. Mol Biol Evol. 2007;24(2):374–81.

983 [56] Carbotte S, Macdonald K. East Pacific Rise 8°–10°30′N: Evolution of ridge segments and
984 discontinuities from SeaMARC II and three-dimensional magnetic studies. J Geophys Res: Solid
985 Earth. 1992;97(B5):6959–82.
986 [57] Fontanillas E, Galzitskaya OV, Lecompte O, Lobanov MY, Tanguy A, Mary J, Girguis PR,
987 Hourdez S, Jollivet D. Proteome evolution of deep-sea hydrothermal vent alvinellid polychaetes
988 supports the ancestry of thermophily and subsequent adaptation to cold in some lineages.
989 Genome Biol Evol 2017;9(2):279-296.
990 [58] Mikkelsen PM. Speciation in modern marine bivalves (Mollusca: Bivalvia): insights from
991 the published record. Am Malacol Bull. 2011;29(1–2):217–45.
992 [59] Rat Genome Sequencing Project Consortium. Genome sequence of the brown Norway rat
993 yields insights into mammalian evolution. Nature. 2004;428(6982):493–521.
994 [60] Gaut B, Yang L, Takuno S, Eguiarte LE. The patterns and causes of variation in plant
995 nucleotide substitution rates. Ann Rev Ecol Evol Syst. 2011; 42(1):245–66.
996 [61] Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. Quantifying the slightly deleterious
997 mutation model of molecular evolution. Mol Biol Evol. 2002; 19(12):2142–49.
998 [62] Hurtado LA, Lutz RA, Vrijenhoek RC. Distinct patterns of genetic differentiation among
999 annelids of Eastern Pacific hydrothermal vents. Mol Ecol. 2004;13(9):2603–15.
1000 [63] Jang S-J, Park E, Lee W-K, Johnson SB, Vrijenhoek RC, Won Y-J. Population subdivision
1001 of hydrothermal vent polychaete *Alvinella pompejana* across Equatorial and Easter Microplate
1002 boundaries. BMC Evol Biol. 2016;16(1):235.
1003 [64] Orr HA, Turelli M. The evolution of postzygotic isolation: accumulating Dobzhansky-
1004 Muller incompatibilities. Evolution. 2001;55(6):1085–94.

1005 [65] Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC.
1006 Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. Genome Res.
1007 2012;22(8):1499–1511.
1008 [66] Palmer ME, Feldman MW. Dynamics of hybrid incompatibility in gene networks in a
1009 constant environment. Evolution. 2009;63(2):418–31.
1010 [67] Palumbi SR. Genetic divergence, reproductive isolation, and marine speciation. Ann Rev
1011 Ecol Syst. 1994;25(1):547-572.
1012 [68] Orr HA. The genetic basis of reproductive isolation: insights from *Drosophila*. Proc Nat
1013 Acad Sci USA. 2005;102(1):6522–26.
1014 [69] Presgraves DC. The molecular evolutionary basis of species formation. Nature Rev Genet.
1015 2010;11(3):175–80.
1016 [70] Castillo, DM, Barbash DA. Moving speciation genetics forward: modern techniques build
1017 on foundational studies in *Drosophila*. Genetics. 2017;207(3):825–42.
1018 [71] Jouin-Toulmond C, Mozzo M, Hourdez S. Ultrastructure of spermatozoa in four species of
1019 Alvinellidae (Annelida: Polychaeta). Cah. Biol. Mar. 2002;*43*(3/4):391-394.
1020 [72] Desbruyères D, Chevaldonné P, Alayse A-M, Jollivet D, Lallier FH, Jouin-Toulmond C, Zal
1021 F, et al. Biology and Ecology of the 'Pompeii Worm' (*Alvinella pompejana* Desbruyères and
1022 Laubier), a normal dweller of an extreme deep-sea environment: a synthesis of current
1023 knowledge and recent developments. Deep Sea Res Part II: Topical Studies in Oceanography.
1024 1998;45(1):383–422.
1025 [73] Breusing C, Castel J, Yang Y, Broquet T, Sun J, Jollivet D, Qian P-Y, Beinart RA. Global
1026 16S rRNA diversity of provannid snail endosymbionts from Indo-Pacific deep-sea hydrothermal
1027 vents**.** Environ Microbiol Reports, 2021;in press.

1028 [74] Papot C, Massol F, Jollivet D, Tasiemski A. Antagonistic evolution of an antibiotic and its
1029 molecular chaperone: how to maintain a vital ectosymbiosis in a highly fluctuating habitat.
1030 Scientific Reports. 2017;7(1):1–14.
1031 [75] Swanson, WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nature Rev
1032 Genet. 2002;3(2):137–44.
1033 [76] Chevaldonné, P, Desbruyères D, Le Haître M. Time-series of temperature from three deep-
1034 sea hydrothermal vent sites. Deep Sea Res Part A. 1991;38(11):1417–30.
1035 [77] Le Bris N, Gaill F. How does the annelid *Alvinella pompejana* deal with an extreme
1036 hydrothermal environment? Rev Environ Sci Biotechnol. 2006;6(1):197.
1037 [78] Gaill F, Wiedemann H, Mann K, Kühn K, Timpl R, Engel J. Molecular characterization of
1038 cuticle and interstitial collagens from worms collected at deep sea hydrothermal vents. J Mol
1039 Biol. 1991;221(1):209–23.
1040 [79] Cary, SC, Cottrell MT, Stein JL, Camacho F, Desbruyères D. Molecular identification and
1041 localization of filamentous symbiotic bacteria associated with the hydrothermal vent annelid
1042 *Alvinella pompejana*. Appl Environ Microbiol. 1997;63(3):1124–30.
1043 [80] Phleger, CF, Nelson MM, Groce AK, Cary SC, Coyne K, Gibson JAE, Nichols PD. Lipid
1044 biomarkers of deep-sea hydrothermal vent polychaetes - *Alvinella pompejana*, *A. caudata*,
1045 *Paralvinella grasslei* and *Hesiolyra bergi*. Deep Sea Res Part I. 2005;52(12):2333–52.
1046 [81] Toulmond A, Slitine FEI, De Frescheville J, Jouin C. Extracellular hemoglobins of
1047 hydrothermal vent annelids: structural and functional characteristics in three alvinellid species.
1048 Biol Bull. 1990;179(3):366–73.
1049 [82] Jouin-Toulmond C, Augustin D, Desbruyères D, Toulmond A. The gas transfer system in
1050 alvinellids (Annelida Polychaeta, Terebellida). anatomy and ultrastructure of the anterior

1051 circulatory system and characterization of a coelomic, intracellular haemoglobin, Cah Biol
1052 Mar. 1996;37(2):135-152.
1053 [83] Chevaldonné P, Jollivet D, Desbruyères D, Lutz RA, Vrijenhoek RC. Sister-species of
1054 Eastern Pacific hydrothermal vent worms (Ampharetidae, Alvinellidae, Vestimentifera) provide
1055 new mitochondrial COI clock calibration. Cah Biol Mar. 2002;43(3–4):367–70.
1056 [84] Sicot F-X, Mesnage M, Masselot M, Exposito J-Y, Garrone R, Deutsch J, Gaill F.
1057 Molecular adaptation to an extreme environment: origin of the thermal stability of the Pompeii
1058 worm collagen. J Mol Biol. 2000;302(4):811–20.
1059 [85] Garrett SC, Rosenthal JJC. A role for A-to-I RNA editing in temperature adaptation.
1060 Physiology. 2012;27(6):362–69.
1061 [86] Pradillon F, Zbinden M, Mullineaux LS, Gaill F. Colonisation of newly-opened habitat by a
1062 pioneer species, *Alvinella pompejana* (Polychaeta: Alvinellidae), at East Pacific Rise vent
1063 sites. Mar Ecol Prog Ser. 2005;302:147-157.
1064 [87] Chevaldonné P, Jollivet D. Videoscopic study of deep-sea hydrothermal vent alvinellid
1065 polychaete populations: estimation of biomass and behaviour. Mar Ecol Prog Ser.
1066 1993;95(3):251-262.
1067 [88] Lee PN, McFall-Ngai MJ, Callaerts P, de Couet HG. Preparation of genomic DNA from
1068 Hawaiian bobtail squid (Euprymna scolopes) tissue by cesium chloride gradient
1069 centrifugation. Cold Spring Harbor Protocols. 2009;11,pdb-prot5319.
1070 [89] Yao B, Zhao Y, Zhang H, Zhang M, Liu M, Liu H, Li J. Sequencing and *de novo* analysis of
1071 the Chinese Sika deer antler-tip transcriptome during the ossification stage using Illumina RNA-
1072 Seq technology. Biotechnol lett. 2012;34(5):813-822.

1073 [90] Gagnière N, Jollivet D, Boutet I, Brélivet Y, Busso D, Da Silva C, Gaill F, et al. Insights
1074 into metazoan evolution from *Alvinella pompejana* cDNAs. BMC Genomics. 2010;11(1):634.
1075 [91] Bankar KG, Todur VN, Shukla RN, Vasudevan M. Ameliorated *de novo* transcriptome
1076 assembly using Illumina paired end sequence data with Trinity Assembler. Genomics
1077 data. 2015;5:352-359.
1078 [92] Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol.
1079 2007;24(8):1586–91.
1080 [93] Xu B, Yang Z. PamlX: a graphical user interface for PAML. Mol Biol Evol. 2013;30
1081 (12):2723–24.
1082 [94] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
1083 approach to multiple testing. J Roy Stat Soc, ser B. 1995;57(1):289–300.