Large effective population size masks population genetic structure in *Hirondellea* amphipods within the deepest marine ecosystem, the Mariana Trench.

Stuart B Piertney[1], Marius Wenzel[1] & Alan J Jamieson[2]

[1]School of Biological Sciences, University of Aberdeen, Aberdeen, UK

[2]Minderoo-UWA Deep-Sea Research Centre, School of Biological Sciences and Oceans Institute, The University of Western Australia, Perth, Australia.

Running title: Genetic structure in hadal amphipods

Corresponding author: Stuart Piertney (s.piertney@abdn.ac.uk)

Abstract

The examination of genetic structure in the deep-ocean hadal zone has focussed on divergence between the tectonic trenches to understand how environment and geography may drive species divergence and promote endemism. There has been little attempt to examine localised genetic structure within trenches, partly because of logistical challenges associated with sampling at an appropriate scale, and the large effective population sizes of species that can be sampled adequately may mask underlying genetic structure. Here we examine genetic structure in the superabundant amphipod *Hirondellea gigas* in the Mariana Trench at depths of 8126-10545 m. RADSeq was used to identify 3,182 loci containing 43,408 SNPs across individuals after stringent pruning of loci to prevent paralogous multicopy genomic regions being erroneously merged. PCA of SNP genotypes resolved no genetic structure between sampling locations, consistent with a signature of panmixia. However, DAPC identified divergence between all sites driven by 301 outlier SNPs in 169 loci and significantly associated with latitude and depth. Functional annotation of loci identified differences between singleton loci used in analysis and paralogous loci pruned from the dataset and also between outlier and non-outlier loci, all consistent with hypotheses explaining the role of transposable elements driving genome dynamics. This study challenges the traditional perspective that highly abundant amphipods within a trench form a single panmictic population. We discuss findings in relation to eco-evolutionary and ontogenetic processes operating in the deep sea, and highlight key challenges associated with population genetic analysis in non-model systems with inherent large effective population sizes and genomes.

Keywords: Deep-sea, amphipod, hadal, Mariana, RADSeq, genetic structure

Introduction

The deep-ocean hadal zone remains one of the most poorly understood and least explored ecosystems on Earth. It extends below 6000 m to full ocean depth at ca. 11,000 m in the Mariana trench, and whilst it accounts for the bottom 45% of the oceans bathymetric range it represents only 2% of the total seafloor area (Jamieson, 2015). Most of the hadal zone is made up of the deep-ocean trenches that form at subduction zones between converging tectonic plates, and most of these trenches are situated around the Pacific rim where they form a discontinuous chain of discrete topographic seafloor features separated by intervening abyssal plains (Stewart & Jamieson, 2018).

An enduring question has been to what extent the individual trenches are demographically and evolutionarily independent, each shaping the patterns of local intra- and inter-specific genetic structure through the combined effects of extreme environmental conditions and their relative geographic isolation (Ritchie et al., 2017b; Weston & Jamieson, 2022). The traditional narrative has been that the trenches promote a high level of species endemism for a specialised hadal fauna that is distinct from shallower deep-sea fauna on the abyssal plains by virtue of adaptation to extreme hydrostatic pressures and reduced food availability (Beliaev, 1989; Wolff, 1959, 1970). Many taxa show a pattern of phylogeographic structure that is consistent with this view, such as the Lysianassoid amphipod genus *Hirondellea* where several species can have restricted geographic ranges associated with individual trenches (France, 1993; Lowry & Stoddart, 2010) while others do not (Weston & Jamieson, 2022). Even in those taxa with a cosmopolitan distribution across ocean basins, such as the amphipods of the *Eurythenes gryllus* complex, there are monophyletic morphospecies found within individual trenches that indicate localised evolutionary divergence and potentially incipient speciation (Eustace et al., 2016; Havermans, 2016; Havermans & Smetacek, 2018; Havermans et al., 2013; Ritchie et al.,

2015; Weston et al., 2021). A similar picture is seen in the genus *Paralicella*, where high levels of genetic divergence exist between neighbouring Kermadec and New Hebrides trenches based upon microsatellite DNA polymorphisms. However an apparent lack of structure for a different *Paralicella* species across the Pacific Ocean questions a total lack of gene flow among individual trenches (Ritchie et al., 2017b).

Most research on how trenches influence the spatial distribution of genetic diversity has focussed on a broad geographic scale between trenches (Weston & Jamieson, 2022). However, there is also evidence for localised population and ontogenetic structure within trenches (Blankenship et al., 2006; Eustace et al., 2016; Lacey et al., 2018; Lacey et al., 2016; Weston et al., 2021). Multiple different deep-sea amphipod species across different trenches show the same patterns of ontogenetic stratification, with juveniles found in in the shallower depths within their individual species bathymetric distribution and only adults present in the deeper regions. It has been considered that at shallower depths the juveniles will experience reduced predation and intra-specific competition, gain a nutritional benefit, and the reduced hydrostatic pressures will confer a physiological and metabolic advantage during growth (Blankenship et al., 2006; Hessler et al., 1978; Lacey et al., 2018). How such ontogenetic structure affects population genetic structure through the effects of differential selection and cohort formation across depth distributions remains unknown.

A major impediment to collating the body of data and evidence necessary to generalise about how the hadal trenches shape patterns of intra-specific genetic structure is the sheer logistical challenge of sampling the deep sea at an appropriate geographic scale and for an adequate number of individuals (Jamieson et al., 2010). Indeed, many of the high profile studies on deep-sea genetic structure work within a phylogeographic rather than a population genetic

framework (Baco et al., 2016; Havermans et al., 2013; Taylor & Roterman, 2017); focus on topographic features such as hydrothermal vents with a proclivity for marked genetic structure driven by natural selection (Vrijenhoek, 2010; Xiao et al., 2020); or confirm hypotheses of wide ranging dispersal in cosmopolitan species from relatively sparse sampling (France & Kocher, 1996). Advances in autonomous underwater vehicle technology have increased the capability for sampling the deep sea in a more systematic and less opportunistic manner. This can provide much better insight of how stochastic and deterministic microevolutionary forces shape the spatial distribution of genetic diversity over different geographic scales (Jamieson, 2015).

Here we exploit a sampling campaign to the Mariana Trench that yielded multiple samples of the amphipod *Hirondellea gigas* to provide the first reports of within-species population genetic structure over a small spatial geographic scale within a single hadal trench. *H. gigas* (Birstein & Vinogradov, 1955) is the dominant scavenging amphipod species found deeper than 6000 m in the hadal trenches of the NW Pacific Ocean (Eustace et al., 2013; France, 1993). It is readily caught in large numbers using baited traps deployed on deep-sea autonomous lander vehicles, which simulates the natural occurrence of surface-derived carrion falls (Blankenship & Levin, 2007). Multiple individuals were sampled along a transect that encompasses both the overriding and subducting tectonic plates and the trench axis of the Mariana Trench. We utilise double digest restriction site-associated DNA sequencing (ddRAD) (Peterson et al., 2012) to genotype individuals at multiple thousand loci across the *Hirondellea* genome. The propensity for sampling *H. gigas* from within the NE Pacific Ocean hadal zones makes it an obvious target species to examine within-species population genetic structure in trenches. However, this also presents two major challenges.

112 The first challenge is that the ease of attracting *H. gigas* to baited traps suggests an extremely
113 large population size. Indeed, the number of individuals attracted to bait are typically hundreds
114 to thousands within hours (Gallo et al., 2015; Hessler et al., 1978). Such hyperabundancy can
115 influence the detection and interpretation of population genetic divergence (Waples &
116 Gaggiotti, 2006). Population structure in neutral markers is primarily driven by the effects of
117 genetic drift, with effects inversely proportional to population size (Hedrick et al., 1976;
118 Waples, 1998). As such there will be very slow accumulation of genetic divergence between
119 large subpopulations, even in the absence of gene flow. Such so-called "panmictic inertia"
120 causes a disconnect between genetic differentiation as estimated from neutral polymorphisms
121 and the actual demographic connectivity between locations. The use of genome-wide markers
122 generated using RADSeq in analysis means that focus can shift away from just examining
123 purely neutral markers influenced by demographic processes, and instead allow examination
124 of patterns of population structure at outlier loci that are highly differentiated relative to neutral
125 expectations reflecting the influence of selection. The spatial distribution of such adaptive
126 genetic variation can be used to infer local adaptation and be particularly useful for identifying
127 the footprint of genetic structure in marine systems where high levels of gene flow and large
128 population sizes are common (Deagle et al., 2015).
129
130 The second challenge is that deep-sea amphipods have extremely large genome sizes (Ritchie
131 et al., 2017a). These can range up to ca. 34 Gbp in *Alicellea gigantea*, with *Hirondellea dubia*
132 having a ca. 5 Gbp genome with a large component of repetitive and paralogous DNA (Ritchie
133 et al., 2017a). This makes it especially difficult to ensure that the relatively short DNA
134 sequences generated using approaches such as RADseq are assembled into true orthologous
135 loci from which single nucleotide polymorphism (SNP) genotypes and allele frequencies can
136 be drawn to infer patterns of population genetic structure. The potential for confounding

paralogous gene sequence variants has been highlighted in species that have undergone historical genome duplication events, such as the salmonids (Davidson et al., 2010; Waples et al., 2016) and also in species with large genomes (Deagle et al., 2015) especially where the focal organism lacks a reference genome and so RADseq relies on *de novo* locus assembly (Nadukkalam Ravindran et al., 2018). In the case of deep-sea amphipods, there is no evidence of ancestral genome duplication, but some indications that large genomes are a consequence of transposable element proliferation which will duplicate genomic regions as a by-product of transposition (Brown & Thatje, 2014; Ritchie et al., 2017a). Both genome duplication and the actions of repetitive elements increase the potential that sequence contigs from polymorphic multicopy genomic regions are erroneously collapsed into a single locus, artificially inflating heterozygosity and biasing allele frequency. Several bioinformatic pipelines are available for RADseq analysis to enrich for true analogues within and between samples by identifying erroneous paralogues based upon optimising sequence similarity metrics, thus preventing over-splitting of true alleles into separate loci or conversely under-splitting and incorrectly merging paralogous sequences into a single locus (Harvey et al., 2015; McKinney et al., 2017; Nadukkalam Ravindran et al., 2018; O'Leary et al., 2018; Willis et al., 2017). Here we combine several of these approaches at high stringency to identify true independent loci from which proper biological inference can be made. This study therefore presents both new insight into the patterns of population genetic structure within hadal amphipods from the Mariana Trench, but also provides a cautionary tale for population genomics studies on non-model organisms with inherently large genomes.

Materials & Methods

Sample collection

Amphipods were collected using the full ocean depth rated autonomous baited camera and trap lander; Hadal-Lander C (Linley et al., 2016) during a research cruise on the RV Falkor (Schmidt Ocean Institute cruise FK141109). The lander was equipped with a conductivity, temperature and depth (CTD) sensor (SBE-19 plus V2; SeaBird Electronics Inc, USA) to record depth every 10 s throughout deployment, and an array of three 30 cm x 6 cm funnel traps baited with 200 g of locally sourced fish. Samples were taken from five sites across the Mariana Trench from north to south, encompassing both the overriding and subducting tectonic plates and the trench axis (Figure 1).

Mature male samples were identified according to Barnard & Ingram (1990) and Barnard & Karaman, (1991). DNA was extracted according to Ritchie et al., (2017b). All samples were confirmed as *H. gigas* from mitochondrial COI DNA barcodes obtained using LCO1490 and HCO12198 PCR primers (Folmer et al., 1994) according to Ritchie et al., (2015).

ddRAD sequencing, locus assembly and paralog filtering

Individuals were genotyped using double-digest RADSeq (Peterson et al., 2012) with the two 6-bp restriction enzymes *Pst*I and *Eco*RI. Samples were multiplexed across four libraries and sequenced across two lanes on an Illumina HiSeq 2000/2500 instrument generating 125 bp paired-end reads. Read pairs with evidence of at least 4 bp of 3' read-through into the *Eco*RI restriction site plus Illumina adapter (forward reads) or the *Pst*I restriction site plus 8bp barcode and Illumina adapter (reverse reads) were removed using CUTADAPT 2.3 (Martin, 2011). The remaining reads were demultiplexed and quality-filtered (-c -r -q options) using the *process_radtags* utility of STACKS 2.58 (Rochette et al., 2019). Barcodes were rescued

allowing one nucleotide mismatch and read pairs containing residual Illumina adapters with at most two mismatches were discarded.

The sequence data were assembled using the *de-novo* STACKS pipeline with additional filtering at various points to enrich for unique non-paralogous loci (hereafter, singletons) (Supplemental Figure S1) (Nadukkalam Ravindran et al., 2018). Initial loci were assembled from the forward reads within each individual in *ustacks,* implementing a strict haplotyping approach to identifying paralogs, where no more than two haplotypes may be observed in an individual and allele merging must be optimised to maximise biallelic loci and minimise monoallelic loci per individual (Ilut et al., 2014; Willis et al., 2017). Alleles were built from primary stacks of at least three identical reads (m=3) and merged into loci via ungapped alignments allowing up to seven mismatches (M=7, equivalent to 94% sequence identity) and at most two primary stacks per locus. The deleveraging algorithm was disabled to ensure that overmerged multi-allelic loci originating from diverged paralogous genomic regions were blacklisted. We selected M=7 based on a parameter sweep from M=0 to M=15, where M=7 minimised the average number of monoallelic loci per individual (though biallelic loci had already peaked at M=4; Supplemental Document S1), yielding sets of high-quality mono- and biallelic loci that are relatively distinct in sequence space within each individual. Initial SNP and haplotype calls were made by aligning secondary stacks to loci with at most nine mismatches (N=9), calling SNPs using the bounded model with an upper error bound of 0.05, and retaining haplotypes from primary stacks only.

Loci from all individuals were then merged into a catalogue using *cstacks* in ungapped alignment mode. To prevent paralogs that were diverged between individuals from being merged into the same catalogue locus we required a 100% match (n=0) between loci from

different individuals. This strict approach ensures that paralog variants are recorded as separate loci but does not preclude *bona fide* heterozygous catalogue loci to be formed from individuals with biallelic loci. Correspondence between catalogue loci and sample alleles was obtained using *sstacks* in ungapped alignment mode, and catalogue loci that linked to alleles from multiple loci within an individual were blacklisted since these originate from within-individual paralogs that were missed by the initial locus assembly. The catalogue locus sequences were then clustered by sequence similarity to identify diverged between-individual paralogs (Nadukkalam Ravindran et al., 2018). Sequence clustering was carried out at 40% identity in ungapped end-to-end alignments using VSEARCH 2.15.1 (Rognes et al., 2016), and all loci that clustered with other loci were blacklisted as putative paralogs. A parameter sweep indicated that the numbers of clusters reached a minimum at 40% (Supplemental Document S1), suggesting that this degree of identity produces the most conservative set of singleton loci in this highly diverse dataset.

Genotyping was completed by incorporating the reverse reads using *tsv2bam* and calling final genotypes from information across all individuals using *gstacks* with the Maruki-low model (alpha = 0.05) (Maruki & Lynch, 2017). The final assembled paired-end catalogue loci consensus sequences were examined for microbial and human contamination by querying against all bacterial, viral, fungal and protozoan genomes available in the NCBI REFSEQ database using BLASTN 2.9.0 (Camacho et al., 2009) with an e-value cut-off of 0.001, and classifying against the PlusPF-16GB database using KRAKEN2 2.0.8 (Wood et al., 2019). Loci identified with at least one of these methods were blacklisted.

A final paralogue check was done based on heterozygosity and allele read ratios in the final genotypes of the paired-end catalogue contigs (McKinney et al., 2017). SNPs with minimum

allele frequency of 0.01, maximum heterozygosity of 0.6 and found in loci common among all five populations were extracted using STACKS *populations*, and paralogous loci were identified using HDplot (https://github.com/gjmckinney/HDplot), enforcing a maximum deviation from even read ratio (z-score) of +- 2 to classify a locus as a singleton. The final blacklist of paralogs consisted of the intersection set between loci blacklisted by the VSEARCH and HDplot methods (between-individuals methods), plus those blacklisted by haplotyping (within-individuals method). Final genotypes at singleton loci that were common among all five populations and identified in at least 50% of individuals per population were extracted using *populations*, requiring a minimum minor-allele frequency of 0.01 and a maximum observed heterozygosity of 0.6.

Inference of genetic structure

Genetic structure among individuals and sampling sites was examined using principal components analysis (PCA) and discriminant analysis of principal components (DAPC) in *adegenet* 2.1.3 (Jombart, 2008). Missing data were replaced with the dataset-wide mean allele frequency at the corresponding SNP. PCA eigenvalues were used to determine the number of principal components retained for DAPC such that 80% of the variance was preserved. Genetic structure among sampling sites was further examined using pairwise F_{ST} (Nei & Chesser, 1983) and population-specific F_{ST} (local F_{ST}) (Weir & Goudet, 2017), both calculated using *FinePop2* (Kitada et al., 2017, 2021). Relationships between individual-based scores on each of the first three linear discriminant functions (DAPC) and latitude or sampling depth were identified using Spearman's rank correlation. The pairwise F_{ST} matrix was summarised using PCA and relationships between the first two principal components and latitude or depth were identified likewise. Isolation by latitude or depth based on the pairwise F_{ST} matrix was tested using Mantel tests with 9999 permutations (*vegan* package) (Oksanen et al., 2013). Population-

specific F_{ST} was regressed onto latitude or depth variables using generalised least-squares models, accounting for autocorrelation in the response (Kitada et al., 2021).

SNPs driving the observed divergence between sites were identified by hierarchically clustering DAPC variable contributions to the linear discriminant functions into "outlier" and "non-outlier" SNPs using Ward's method (*snpzip* function in *adegenet*). F_{ST}-based outliers were identified using OutFLANK (Lotterhos & Whitlock, 2015) and outliers associated with individual-based structure were identified using multivariate Mahalanobis distances based on the first six principal components using PCADAPT (Luu et al., 2017). The union of the three sets was retained as a final set of outliers.

Finally, genetic structure among individuals based on outliers and non-outliers was inferred using two Bayesian methods. SNP-based inference was carried out by thinning SNPs by linkage disequilibrium using PLINK 1.9 (Chang et al., 2015) with an r^2 threshold of 0.5 and then using fastSTRUCTURE with the logistic prior model and inferring between two (K=2) and five (K=5) genetic clusters (Raj et al., 2014). Haplotype-based inference was carried out using RADpainter and fineRADstructure (Malinsky et al., 2018). The co-ancestry matrix inferred by RADpainter was inspected using PCA and used to assign individuals to an optimal number of genetic clusters using fineRADstructure with 10^6 iterations burn-in, 10^6 iterations MCMC and 10^4 thinning interval. The clustering dendrogram was obtained from the best posterior state identified from at most 10^5 search attempts.

Functional annotation of loci

The assembled ddRAD loci consensus sequences were functionally annotated to identify functional enrichment among subsets of loci. Loci containing repetitive elements were

identified using REPEATMASKER 4.0.9 with the arthropoda database. Loci overlapping coding sequences were identified via alignment against the *Hirondellea gigas* transcriptome assembly (NCBI TSA accession GEZX01) using BLASTN 2.9.0 (Camacho et al., 2009) with an e-value cut-off of 0.001. Protein and GeneOntology (GO) annotations (biological process ontology) for each locus were obtained via alignment against all arthropod proteins available on UniProt (taxonomy ID 6656) using BLASTX 2.9.0 with an e-value cut-off of 0.001. Enrichment of GO terms among outlier loci against a background of non-outlier loci were carried out using hypergeometric tests in CLUSTERPROFILER 3.14.2 (Yu et al., 2012) with an FDR-corrected significance threshold of 0.05.

Results

ddRAD data assembly

An average of 3,306,212 read pairs per sample passed quality filtering and were assembled to 19,905-34,486 loci per individual with a mean coverage of 12-35x per locus (Supplemental document S1). Sequencing depth and numbers of loci were similar among sampling sites (Kruskal-Wallis test: $P > 0.05$; Supplemental document S1), though one library produced significantly fewer loci than other libraries ($P < 0.05$; Supplemental document S1). The locus catalogue comprised 457,130 loci, of which 11,407 (2%) were removed as microbial/human contaminants and 3,182 non-paralogous loci containing 43,408 SNPs were present in all five sampling sites and at least 50% of individuals at each site.

Inspection of locus-sharing patterns among individuals revealed systematic data missingness associated with ddRAD library, which was partially confounded by sampling site. PCA on Jaccard distances and DAPC between libraries separated individuals from sites C or D in one particular library from individuals from the same site in other libraries (Supplemental Figure

2). This same separation was also apparent in SNP genotypes, even in loci shared across all individuals. The library in question had fewer loci overall (Supplemental document S1) and more overlap in read pairs (45% versus 13-25% pairs merged using FLASH 1.2.11, (Magoč & Salzberg, 2011), suggesting a difference in size-selection regime. To account for this batch effect, loci or SNPs that were among the top 10% features driving the DAPC separation in locus sharing or SNP genotypes between individuals within sites C or D were removed (Deagle et al., 2015; O'Leary et al., 2018). This removed 70% of SNPs with heterozygote deficiency in sites C and D (Supplemental Figure 3) and left 824 loci containing 5,147 SNPs with minimal residual discrimination between libraries and good discrimination between sites C and D (Supplemental Figure 2). Notwithstanding, all further analyses were also confirmed with a reduced dataset, comprising only individuals from sites C, D and E from the outlier library.

Genetic structure

PCA on SNP genotypes indicated substantial genetic differentiation among individuals but not sampling sites, whereby the two axes accounted for 4.98% and 3.92% of the variance (Figure 2a). In contrast, DAPC on individuals grouped by sampling site revealed three major genetic groups on the first two discriminant functions, comprising sites A+B, C+D and E (Figure 2b). The third discriminant function separated sites A from B and C from D (Figure 2b). The genetic structure described by the first two discriminant functions was significantly associated with latitude (axis 1; ρ = - 0.870; $P < 0.001$) and trench depth (axis 2; ρ = - 0.757; $P < 0.001$) respectively (Figure 3). These associations were stable across a broad range of PCA variance preserved in the DAPC (Supplemental Table 1). Based on DAPC variable loadings, this genetic structure was driven by 301 outlier SNPs. No significant F_{ST}-outliers were identified, and PCA-based Mahalanobis distances yielded 23 outliers, of which 6 (26%) were not identified by DAPC, yielding a combined set of 307 outlier SNPs and 4,840 non-outlier SNPs.

335

336 The pairwise F_{ST} matrix based on the outlier SNPs supported the same three genetic clusters
337 as identified via DAPC (Figure 4a) and the first two principal components tended to be
338 associated with latitude and depth respectively (ρ = -0.9; P = 0.08). Mantel tests supported a
339 weak relationship with latitude (r = 0.66; P = 0.06) but not with depth (r = -0.07; P = 0.55).
340 Pairwise F_{ST} based on non-outlier SNPs described a much weaker genetic structure, primarily
341 driven by divergence of site B, but showing no associations with latitude or depth (Figure 4a).
342 Population-specific (local) F_{ST} corroborated the contrast in population structure between
343 outlier and non-outlier loci (Figure 4b), whereby it was weakly associated with depth in outliers
344 (β = 3.87e-05; SE = 2.13e-05; Z = 1.815; P = 0.070) and strongly associated with latitude in
345 non-outliers (β = -0.418; SE = 0.130; Z = -3.216; P = 0.001).
346

347 Individual-based Bayesian inference of genetic structure failed to resolve any structure using
348 non-outlier SNPs and yielded genetic clusters inconsistent with geographic or bathymetric
349 structure using outlier SNPs (Supplemental Figure 4). Similarly, the haplotype-based
350 coancestry matrix resolved minimal structure using non-outlier loci and failed to identify
351 structure congruent with sampling site, geography or bathymetry using outlier loci
352 (Supplemental Figure 5). Analysis of a reduced dataset comprising only samples from a single
353 library covering sites C, D and E corroborated genetic structure between these three sites
354 (Supplemental Figure 6), and local F_{ST} corroborated genetic similarity between the deep sites
355 C and D compared to the shallow site E (Supplemental Figure 7).
356

357 ## Functional annotation of loci

358 Out of the total raw 457,130 loci, 33,174 (7%) contained repetitive sequence elements, 218,874
359 (48%) matched a *Hirondellea gigas* transcript and 65,921 (14%) matched an arthropod UniProt

protein. The various filters employed to enrich for singleton loci identified 25,402 loci as paralogs common between at least 50% of individuals from each sampling site - these contained significantly more repetitive elements (5.9% vs. 3.6%; $\chi^2_1 = 6.821$; $P = 0.009$), significantly fewer transcripts (53.0% vs. 57.7%; $\chi^2_1 = 6.765$; $P = 0.009$) and similar numbers of proteins (17.0% vs. 18.8%; $\chi^2_1 = 1.650$; $P = 0.199$) compared to the singleton loci. Among these singletons, proportions of annotated loci did not differ significantly between 170 outlier and 654 non-outlier loci, ranging from 2.4%-4% for repetitive elements, 52.3%-59.2% for transcripts and 18.2%-21.2% for proteins (Figure 5).

A total of 66, 116 and 345 GeneOntology (GO) biological process annotations were assigned to the outlier, non-outlier and paralog loci, respectively. None of these GO terms were significantly enriched (FDR > 0.2) against background locus sets comprising all 3182 singleton loci or (for paralogs) the whole raw locus catalogue. However, the set of GO terms assigned to paralog loci was markedly differed from those in singletons with only 54 out of 345 terms (15.7%) being shared. The two most frequently observed GO terms among outliers and non-outliers are related to DNA transposon activity (*DNA integration* [GO:0015074], *transposition, DNA-mediated* [GO:0006313]), followed by terms related to epigenetics and protein ubiquitination (Figure 5, Supplemental Document S2). The transposon terms represent several reverse transcriptase/integrase enzymes and retrotransposon polyproteins (Tf2, TC1, Gag-pol, TY3B-I/G). The most frequently observed GO term among paralogs was *viral genome replication* [GO:0019079], followed by various fundamental cellular processes (Supplemental Document S2). The GO terms unique to outlier loci comprised processes involved in protein modification, photoperception (*sina* and *SIAH1* genes) and neural function among various fundamental RNA processing terms (Supplemental Document S2).

Discussion

This study has resolved fine-scale population genetic structure in *Hirondellea gigas* deep-sea amphipods in the Mariana Trench, highlighting a signature of isolation by geography and bathymetry across a small set of strongly diverged loci against a highly diverse genomic background impacted by large population size, large genome size and extensive locus paralogy caused by transposable elements. Our analyses thus provide new insight into eco-evolutionary processes operating in the deep ocean hadal zone, emphasising how both stochastic and deterministic microevolutionary forces are shaping the spatial distribution of genetic variation and so could contribute across range of spatial scales to population and phylogeographic structure and potentially the speciation process.

The large genome sizes characteristic of many deep-sea amphipods, which range from c. 4-34 Gbp (Ritchie et al., 2017a), cause considerable technical challenges when attempting to resolve population genetic structure using approaches such as *de novo* RADseq. Proliferation of transposable elements increases levels of paralogy across the genome and thus increases the risk that sequence reads from paralogous multicopy regions are incorrectly merged into a single locus, leading to a bias in estimates of allele frequency and any downstream population genetic analysis inference. This problem is not unique to deep-sea amphipods but applies to all organisms with large and dynamic genomes, particularly those having undergone genome duplication or subject to extensive TE activity (Davidson et al., 2010; Deagle et al., 2015; Waples et al., 2016;).

We made several modifications to the standard ddRAD STACKS pipeline to ensure that the final dataset yielded robust inference of genetic structure. First, we enriched for non-paralogous loci within individuals during *ustacks* assembly by increasing the number of allowed mismatches between alleles, disabling the deleveraging algorithm and allowing at most two

alleles per locus. These settings are differemt to standard recommendations (Paris et al., 2017), but are effective in blacklisting biologically implausible loci comprising multiple alleles originating from diverged paralogs. We empirically identified the ideal threshold in our dataset to strike a balance between minimising the numbers of monoallelic loci and maximising the numbers of biallelic loci (Ilut et al., 2014; Willis et al., 2017). Second, we observed high levels of between-individual loci similarity (Supplemental Document S1), leading to a decrease in catalogue loci by >40% when allowing seven mismatches in *cstacks* as suggested by default recommendations (Paris et al., 2017). An effective way of avoiding overmerging of paralogs diverged between individuals into confounded catalogue loci is to undermerge loci in *cstacks* and blacklist catalogue loci that cluster at relatively low sequence similarity (Nadukkalam Ravindran et al., 2018). In our dataset, even a similarity threshold of 90% reduced the number of unique loci across all individuals by more than 50%, and the number of locus clusters did not stabilise until 40% similarity, consistent with extensive locus paralogy (Supplemental Document S1). Finally, we removed loci with uneven per-allele read depths, which may originate from identical paralogous genomic regions that our previous measures failed to identify (McKinney et al., 2017). Our strict modified RAD-Seq pipeline may be overly conservative, possibly falsely blacklisting diverged singleton loci (Nadukkalam Ravindran et al., 2018), but is clearly essential when dealing with data with extensive paralogy.

Although our focus was not on identifying and characterising TE content, for which more specialised pipelines exist (Chak & Rubenstein, 2019), comparing the functional annotations of the singleton versus paralogous locus sets allowed us to gain some insight into the role of TEs in affecting eco-evolutionary processes operating in *H. gigas* in the deep sea. The paralogous loci contained significantly more repetitive elements and returned fewer hits with the *H. gigas* transcriptome compared to the singleton loci. The singleton and paralogous loci also had near non-overlapping gene ontology assignments, with the singleton loci being

associated with GO terms that included regulation of integration, transposition and methylation that would also be involved in policing the activity of transposable elements. Combined, these observations reinforce the significance of transposable elements in driving genome size in deep-sea amphipods (Naville et al., 2019; Ritchie et al., 2017a), and are consistent with hypotheses implicating TE as drivers of evolutionary innovation necessary to allow the colonisation of the deep ocean from shallower bathyal depths (Brown & Thatje, 2014). There is growing recognition and focus around how TE proliferation can generate genomic novelty through a variety of mechanisms that can lead to rapid adaptation across a broad taxonomic range and under varying ecological contexts (Bourque et al., 2018; Ricci et al., 2018; Wells & Feschotte, 2020). For deep-sea amphipods what is now required is the mapping of different TE families across host genomes to examine location relative to genes that could influence the adaptive potential to extreme hydrostatic pressure, especially for those situations where there is clear gene family expansion or signatures of positive selection (Kobayashi et al., 2018, 2019; Lan et al., 2017).

A further inherent challenge for identifying patterns of population genetic structure in *H. gigas* is that the high local abundance of these amphipods could show, at most, only subtle levels of genetic divergence (Hedrick et al., 1976). Given that the effect of genetic drift on allele frequencies is inversely proportional to population size, even in the complete absence of gene flow among locations there will only be very slow accumulation of genetic differences at neutral loci that makes inferring dispersal and population connectivity problematic (Waples & Gaggiotti, 2006). This genomic inertia can cause a disconnect between patterns of demographic and population genetic structure (Lowe & Allendorf, 2010) and can be common occurrence in marine systems where species can have large effective population sizes and fewer physical barriers to gene flow (Baco et al., 2016; Deagle et al., 2015; Etter et al., 1999). Indeed, our initial data exploration using PCA suggested no genetic structure between sampling locations

459 within the Mariana Trench, consistent with a signature of panmixia. However, a parallel DAPC
460 analysis, that acts to maximise variance between sampling locations whilst minimising
461 variance within them, identified clear genetic structure across all sampling sites. This
462 differentiation was driven by a relatively small number of SNPs and loci, which display higher
463 levels of divergence among populations than the neutral genomic average and thus can be
464 considered to reflect the effects of adaptive divergence. Markers under divergent selection will
465 change in frequency more readily than neutral polymorphisms even in large populations,
466 though the capacity to identify genomic regions under selection is difficult with low levels of
467 linkage disequilibrium (Pritchard & Przeworski, 2001).

468 An intriguing finding was that the multivariate structure described by the first two discriminant
469 functions was significantly correlated with both latitude and trench depth. Relationships with
470 latitude may reflect a signature of isolation by distance, though given the structure was driven
471 primarily by a limited set of outlier loci this would also suggest some gradual signature of
472 divergence based upon some environmental gradient. These outlier loci represented GO terms
473 associated with DNA integration, transposition and methylation, in line with the broader suite
474 of singleton loci, again reflecting the potential important of transposable element activity and
475 policing in amphipod adaptation and evolution (Brown & Thatje, 2014; Ritchie et al., 2017a).
476 Critically, the association with trench depth challenges any oversimplistic perspective about
477 patterns of dispersal, gene flow and genetic structure in the deep sea and the hadal zone. The
478 traditional paradigm was centred around the depth-differentiation hypothesis (Rex & Etter,
479 2010) which postulated that there will be decreases in both environmental heterogeneity and
480 barriers to gene flow with increasing depth and hence less structure. Whilst data from some
481 deep-sea species do support the depth differentiation hypothesis (Cowart et al., 2014; Quattrini
482 et al., 2015; Ritchie et al., 2013) it is apparent that as more studies examine a broader taxonomic
483 spread and range of habitats, the more exceptions there are to the general rule. Disjunct

topographical features such as seamounts, fracture zones and ocean ridges do disrupt gene flow, and hydrothermal vents offer localised selection pressures that drive adaptive divergence (Baco et al., 2016). The hadal trenches, as a disconnected network of ultra-deep island-like habitats, are also recognised as drivers of genetic structure in the deep sea. This is evidenced both by the levels of endemism found in the hadal zone, and also from some population genetic studies that have compared across trenches in cosmopolitan amphipod species (Chan et al., 2020; Havermans et al., 2013; Ritchie et al., 2017b; Weston et al., 2022). Patterns of genetic structure have been framed in the context of understanding how the intervening abyssal plains operate as barriers to gene flow and how differences in selection between abyssal and hadal zones may drive divergence. The current study provides whole new insight into the microgeographic scales over which population genetic structure can accrue, rejecting any concept that highly abundant amphipods within a trench form a single panmictic population. Moreover, it highlights that environmental differences operating within a trench and over microgeographic scales do influence the spatial distribution of genetic variation.

Within-trench population structure has been hinted at before from patterns of ontogenetic variation in amphipods, with a consistent trend across species and trenches of juveniles residing in the shallower depths of the bathymetric range (Blankenship et al., 2006; Eustace et al., 2016; Lacey et al., 2018; Weston et al., 2021). Spatially this would mean that juveniles at shallower sites move in a downslope direction as they grow and mature, which could explain the patterns of isolation-by-depth observed in the current data given the environmental gradients associated with trench depth and the effects of genetic drift operating in a similar way to a founder effect as juveniles moved to deeper reaches of the trench. This would be especially pronounced if juveniles formed periodic cohorts. Chronobiological triggers that could entrain reproduction do occur even at full ocean depth (Taira et al., 2004), and temporal and seasonal shifts in

surfaced derived food supply with associated responses of deep-sea benthic fauna are well-documented (Durden et al., 2020; Gooday & Lambshead, 1989; Kalogeropoulou et al., 2010). It is also known that *Hirondellea* amphipods do not move long distances through the water column, with video footage from the autonomous landers showing they so not move more than a metre of so from the surface of the trench, and amphipod traps that are suspended above the trench floor rarely catching any individuals (Jamieson et al., 2009). This, coupled with a life cycle in *Hirondellea* that involves fertilisation and juvenile development in a marsupium, would predict that longer distance dispersal is limited. Moreover, the trench topology could provide multiple localised barriers to dispersal, certainly from the trench axis onto the shallower aspects of the plates, and in the case of the Mariana Trench, there are five physically partitioned areas at hadal depths within this one trench (Jamieson & Stewart, 2021).

Overall, this study exemplifies why deep-sea population genetics needs to embrace a seascape genomics approach (Liggins et al., 2020; Nielsen et al., 2020; Riginos et al., 2016) to understand better how the deep-sea environment shapes genomic diversity both in terms of how geographical features influence the connectivity of populations and how environmental gradients affect demographic structure and selective regimes. The drivers of genetic structure encompass both stochastic and deterministic microevolutionary forces spanning a range of spatial and temporal scales often underpinned by subtle if not enigmatic environmental effects. This requires both the capacity to examine genome-wide structure in non-model systems with potentially low levels of underlying genomic resources, plus also sampling at an appropriate scale that can encompass both micro- and macrogeographic patterns. The latter is being achieved through advances in deep-ocean lander and submersible technology, and we have illustrated how the former is achievable with scrutiny of the data generated and recourse to potential sources of bias and error such as large genome size and effective population size.

Acknowledgements

This work was supported by NERC (NBAF884 to AJJ and NE/N01149X/1 to SBP), the Leverhulme Trust (to SBP) and the Schmidt Ocean Institute, USA (SBP and AJJ). We are grateful to the captain and crew of the RV Falkor, Professor Jeff Drazen (University of Hawaii) as Principal Scientist for cruise FK141109 and Dr Thomas Linley (Newcastle University) for assistance in lander operations. We thank Edinburgh Genomics for RAD genotyping services, and Heather Ritchie for useful discussions.

References

Baco, A. R., Etter, R. J., Ribeiro, P. A., von der Heyden, S., Beerli, P., & Kinlan, B. P. (2016). A synthesis of genetic connectivity in deep-sea fauna and implications for marine reserve design. *Molecular Ecology*, *25*(14), 3276–3298. https://doi.org/10.1111/mec.13689

Barnard, J. L., & Ingram, C. L. (1990). Lysianassoid Amphipoda (Crustacea) from deep-sea thermal vents. *Smithsonian Contributions to Zoology*, *499*, 1–80. https://doi.org/10.5479/si.00810282.499

Barnard, J. L., & Karaman, G. S. (1991). The families and genera of marine gammaridean Amphipoda (except marine gammaroids). Part 2. *Records of the Australian Museum, Supplement*, *13*(2), 419–866.

Beliaev, G. (1989). *Deep sea ocean trenches and their fauna.* Nauka.

Birstein, J., & Vinogradov, M. (1955). Pelagicheskie gammaridy (Amphipoda, Gammaridea) Kurilo-Kamchatskoi Vpadiny. *Akademiya Nauk SSSR, Trudy Instituta Okeanologii*, *12*, 219–257.

Blankenship, L. E., & Levin, L. A. (2007). Extreme food webs: Foraging strategies and diets of scavenging amphipods from the ocean's deepest 5 kilometers. *Limnology and Oceanography*, *52*(4), 1685–1697.

Blankenship, L. E., Yayanos, A. A., Cadien, D. B., & Levin, L. A. (2006). Vertical zonation patterns of scavenging amphipods from the hadal zone of the Tonga and Kermadec Trenches. *Deep Sea Research Part I: Oceanographic Research Papers*, *53*(1), 48–61. https://doi.org/10.1016/j.dsr.2005.09.006

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., & others. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*, 1–12.

Brown, A., & Thatje, S. (2014). Explaining bathymetric diversity patterns in marine benthic invertebrates and demersal fishes: Physiological contributions to adaptation of life at depth. *Biological Reviews*, *89*(2), 406–426. https://doi.org/10.1111/brv.12061

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 1–9.

Chak, S. T., & Rubenstein, D. R. (2019). TERAD: Extraction of transposable element composition from RADseq data. *Molecular Ecology Resources*, *19*(6), 1681–1688.

Chan, J., Pan, B., Geng, D., Zhang, Q., Zhang, S., Guo, J., & Xu, Q. (2020). Genetic diversity and population structure analysis of three deep-sea amphipod species from geographically isolated hadal trenches in the Pacific Ocean. *Biochemical Genetics*, *58*, 157–170.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015.

583 Cowart, D. A., Halanych, K. M., Schaeffer, S. W., & Fisher, C. R. (2014). Depth-dependent
584 gene flow in Gulf of Mexico cold seep Lamellibrachia tubeworms (Annelida,
585 Siboglinidae). *Hydrobiologia*, *736*, 139–154.
586 Davidson, W. S., Koop, B. F., Jones, S. J., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien,
587 S., & Omholt, S. W. (2010). Sequencing the genome of the Atlantic salmon (*Salmo*
588 *salar*). *Genome Biology*, *11*(9), 403. https://doi.org/10.1186/gb-2010-11-9-403
589 Deagle, B. E., Faux, C., Kawaguchi, S., Meyer, B., & Jarman, S. N. (2015). Antarctic krill
590 population genomics: Apparent panmixia, but genome complexity and large
591 population size muddy the water. *Molecular Ecology*, *24*(19), 4943–4959.
592 Durden, J. M., Bett, B. J., Huffard, C. L., Pebody, C., Ruhl, H. A., & Smith Jr, K. L. (2020).
593 Response of deep-sea deposit-feeders to detrital inputs: A comparison of two abyssal
594 time-series sites. *Deep Sea Research Part II: Topical Studies in Oceanography*, *173*,
595 104677.
596 Etter, R. J., Rex, M. A., Chase, M. C., & Quattro, J. M. (1999). A genetic dimension to deep-
597 sea biodiversity. *Deep-Sea Research Part I: Oceanographic Research Papers*, *46*(6),
598 1095–1099. https://doi.org/10.1016/S0967-0637(98)00100-9
599 Eustace, R. M., Kilgallen, N. M., Lacey, N. C., & Jamieson, A. J. (2013). Population
600 structure of the hadal amphipod *Hirondellea gigas* (Amphipoda: Lysianassoidea)
601 from the Izu-Bonin Trench. *Journal of Crustacean Biology*, *33*(6), 793–801.
602 Eustace, R. M., Ritchie, H., Kilgallen, N. M., Piertney, S. B., & Jamieson, A. J. (2016).
603 Morphological and ontogenetic stratification of abyssal and hadal *Eurythenes gryllus*
604 sensu lato (Amphipoda: Lysianassoidea) from the Peru–Chile Trench. *Deep Sea*
605 *Research Part I: Oceanographic Research Papers*, *109*, 91–98.
606 https://doi.org/10.1016/j.dsr.2015.11.005

607 Folmer, O., Black, M, Hoeh, W, Lutz, R, & Vrijenhoek, R. (1994). DNA primers for
608 amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan
609 invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299.
610 France, S. C. (1993). Geographic variation among three isolated populations of the hadal
611 amphipod *Hirondellea gigas* (Crustacea: Amphipoda: Lysianassoidea). *Marine*
612 *Ecology Progress Series*, *92*(3), 277–287. https://doi.org/10.3354/meps092277
613 France, S. C., & Kocher, T. D. (1996). Geographic and bathymetric patterns of mitochondrial
614 16S rRNA sequence divergence among deep-sea amphipods, *Eurythenes gryllus.*
615 *Marine Biology*, *126*(4), Art. 4. https://doi.org/10.1007/BF00351330
616 Gallo, N. D., Cameron, J., Hardy, K., Fryer, P., Bartlett, D. H., & Levin, L. A. (2015).
617 Submersible-and lander-observed community patterns in the Mariana and New
618 Britain trenches: Influence of productivity and depth on epibenthic and scavenging
619 communities. *Deep Sea Research Part I: Oceanographic Research Papers*, *99*, 119–
620 133.
621 Gooday, A. J., & Lambshead, P. (1989). Influence of seasonally deposited phytodetritus on
622 benthic foraminiferal populations in the bathyal northeast Atlantic: The species
623 response. *Marine Ecology Progress Series*, 53–67.
624 Harvey, M. G., Judy, C. D., Seeholzer, G. F., Maley, J. M., Graves, G. R., & Brumfield, R. T.
625 (2015). Similarity thresholds used in DNA sequence assembly from short reads can
626 reduce the comparability of population histories across species. *PeerJ*, *3*, e895.
627 https://doi.org/10.7717/peerj.895
628 Havermans, C. (2016). Have we so far only seen the tip of the iceberg? Exploring species
629 diversity and distribution of the giant amphipod *Eurythenes*. *Biodiversity*, *17*(1–2),
630 12–25. https://doi.org/10.1080/14888386.2016.1172257

631 Havermans, C., & Smetacek, V. (2018). Bottom-up and top-down triggers of diversification:
632 A new look at the evolutionary ecology of scavenging amphipods in the deep sea.
633 *Progress in Oceanography*, *164*, 37–51. https://doi.org/10.1016/j.pocean.2018.04.008
634 Havermans, C., Sonet, G., d'Udekem d'Acoz, C., Nagy, Z. T., Martin, P., Brix, S., Riehl, T.,
635 Agrawal, S., & Held, C. (2013). Genetic and morphological divergences in the
636 cosmopolitan deep-sea amphipod *Eurythenes gryllus* reveal a diverse abyss and a
637 bipolar species. *PLoS ONE*, *8*(9), Art. 9. https://doi.org/10.1371/journal.pone.0074218
638 Hedrick, P. W., Ginevan, M., & Ewing, E. (1976). Genetic polymorphism in heterogeneous
639 environments. *Annual Review of Ecology and Systematics*, *7*, 1–32.
640 Hessler, R. R., Ingram, C. L., Aristides Yayanos, A., & Burnett, B. R. (1978). Scavenging
641 amphipods from the floor of the Philippine trench. *Deep-Sea Research*, *25*(11).
642 https://doi.org/10.1016/0146-6291(78)90585-4
643 Hollister, J., Shah, T., Robitaille, A. L., Beck, M. W., & Johnson, M. (2021). *elevatr: Access*
644 *Elevation Data from Various APIs*. https://doi.org/10.5281/zenodo.5809645
645 Ilut, D. C., Nydam, M. L., & Hare, M. P. (2014). Defining loci in restriction-based reduced
646 representation genomic data from nonmodel species: Sources of bias and diagnostics
647 for optimal clustering. *BioMed Research International*, *2014*.
648 Jamieson, A. (2015). *The Hadal Zone: Life in the Deepest Oceans*. Cambridge University
649 Press.
650 Jamieson, A. J., Fujii, T., Mayor, D. J., Solan, M., & Priede, I. G. (2010). Hadal trenches:
651 The ecology of the deepest places on Earth. *Trends in Ecology and Evolution*, *25*(3),
652 190–197. https://doi.org/10.1016/j.tree.2009.09.009
653 Jamieson, A. J., Fujii, T., Solan, M., & Priede, I. G. (2009). HADEEP: Free-Falling Landers
654 to the Deepest Places on Earth. *Marine Technology Society Journal*, *43*(5), 151–160.
655 https://doi.org/10.4031/MTSJ.43.5.17

656 Jamieson, A. J., & Stewart, H. A. (2021). Hadal zones of the northwest Pacific Ocean.
657 *Progress in Oceanography*, *190*, 102477.
658 Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers.
659 *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129
660 Kahle, D. J., & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *R J.*, *5*(1),
661 144.
662 Kalogeropoulou, V., Bett, B., Gooday, A., Lampadariou, N., Arbizu, P. M., & Vanreusel, A.
663 (2010). Temporal changes (1989–1999) in deep-sea metazoan meiofaunal
664 assemblages on the Porcupine Abyssal Plain, NE Atlantic. *Deep Sea Research Part*
665 *II: Topical Studies in Oceanography*, *57*(15), 1383–1395.
666 Kitada, S., Nakamichi, R., & Kishino, H. (2017). The empirical Bayes estimators of fine-
667 scale population structure in high gene flow species. *Molecular Ecology Resources*,
668 *17*(6), 1210–1222.
669 Kitada, S., Nakamichi, R., & Kishino, H. (2021). Understanding population structure in an
670 evolutionary context: Population-specific F ST and pairwise F ST. *G3*, *11*(11),
671 jkab316.
672 Kobayashi, H., Nagahama, T., Arai, W., Sasagawa, Y., Umeda, M., Hayashi, T., Nikaido, I.,
673 Watanabe, H., Oguri, K., Kitazato, H., & others. (2018). Polysaccharide hydrolase of
674 the hadal zone amphipods *Hirondellea gigas*. *Bioscience, Biotechnology, and*
675 *Biochemistry*, *82*(7), 1123–1133.
676 Kobayashi, H., Shimoshige, H., Nakajima, Y., Arai, W., & Takami, H. (2019). An aluminum
677 shield enables the amphipod Hirondellea gigas to inhabit deep-sea environments. *PloS*
678 *One*, *14*(4), e0206710.
679 Lacey, N. C., Mayor, D. J., Linley, T. D., & Jamieson, A. J. (2018). Population structure of
680 the hadal amphipod *Bathycallisoma (Scopelocheirus) schellenbergi* in the Kermadec

Trench and New Hebrides Trench, SW Pacific. *Deep Sea Research Part II: Topical Studies in Oceanography*, *155*, 50–60.

Lacey, N. C., Rowden, A. A., Clark, M. R., Kilgallen, N. M., Linley, T., Mayor, D. J., & Jamieson, A. J. (2016). Community structure and diversity of scavenging amphipods from bathyal to hadal depths in three South Pacific Trenches. *Deep-Sea Research Part I: Oceanographic Research Papers*, *111*, 121–137. https://doi.org/10.1016/j.dsr.2016.02.014

Lan, Y., Sun, J., Tian, R., Bartlett, D. H., Li, R., He, T. X. L., Him, Y., Weipeng, W., Qiu, Z. J., & Qian, H. G. T. P. (2017). *Molecular adaptation in the world ' s deepest-living animal: Insights from transcriptome sequencing of the hadal amphipod Hirondellea gigas*. *March*, 3732–3743. https://doi.org/10.1111/mec.14149

Liggins, L., Treml, E. A., & Riginos, C. (2020). Seascape genomics: Contextualizing adaptive and neutral genomic variation in the ocean environment. *Population Genomics: Marine Organisms*, 171–218.

Linley, T. D., Gerringer, M. E., Yancey, P. H., Drazen, J. C., Weinstock, C. L., & Jamieson, A. J. (2016). Fishes of the hadal zone including new species, in situ observations and depth records of Liparidae. *Deep-Sea Research Part I: Oceanographic Research Papers*, *114*, 99–110. https://doi.org/10.1016/j.dsr.2016.05.003

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*.

Lowe, W. H., & Allendorf, F. W. (2010). What can genetics tell us about population connectivity? *Molecular Ecology*, *19*(15), 3038–3051.

Lowry, J., & Stoddart, H. (2010). The deep-sea scavenging genus *Hirondellea* (Crustacea: Amphipoda: Lysianassoidea: Hirondelleidae fam. Nov.) in Australian waters. *Zootaxa*, *2329*(1), 37–55.

Luu, K., Bazin, E., & Blum, M. G. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77.

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963.

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure: Population inference from RADseq data. *Molecular Biology and Evolution*, *35*(5), 1284–1290.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, *17*(1), 10–12.

Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, *7*(5), 1393–1404.

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. https://doi.org/10.1111/1755-0998.12613

Nadukkalam Ravindran, P., Bentzen, P., Bradbury, I. R., & Beiko, R. G. (2018). PMERGE: Computational filtering of paralogous sequences from RAD-seq data. *Ecology and Evolution*, *8*(14), 7002–7013. https://doi.org/10.1002/ece3.4219

Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N., & Chourrout, D. (2019). Massive changes of genome size driven by expansions of non-autonomous transposable elements. *Current Biology*, *29*(7), 1161–1168.

729 Nei, M., & Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Annals*
730 *of Human Genetics*, *47*(3), 253–259.
731 Nielsen, E. S., Henriques, R., Beger, M., Toonen, R. J., & Von der Heyden, S. (2020). Multi-
732 model seascape genomics identifies distinct environmental drivers of selection among
733 sympatric marine species. *BMC Evolutionary Biology*, *20*(1), 1–17.
734 Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P., O'hara, R., Simpson, G.,
735 Solymos, P., Stevens, M. H. H., Wagner, H., & others. (2013). Community ecology
736 package. *R Package Version*, *2*(0).
737 O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These
738 aren't the loci you'e looking for: Principles of effective SNP filtering for molecular
739 ecologists. *Molecular Ecology*, *27*(16), 3193–3206.
740 https://doi.org/10.1111/mec.14792
741 Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: A road map for
742 stacks. *Methods in Ecology and Evolution*, *8*(10), 1360–1373.
743 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double
744 digest RADseq: An inexpensive method for de novo SNP discovery and genotyping
745 in model and non-model species. *PLoS ONE*, *7*(5), Art. 5.
746 https://doi.org/10.1371/journal.pone.0037135
747 Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and
748 data. *The American Journal of Human Genetics*, *69*(1), 1–14.
749 Quattrini, A. M., Baums, I. B., Shank, T. M., Morrison, C. L., & Cordes, E. E. (2015).
750 Testing the depth-differentiation hypothesis in a deepwater octocoral. *Proceedings of*
751 *the Royal Society B: Biological Sciences*, *282*(1807), 20150008.
752 Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: variational inference of
753 population structure in large SNP data sets. *Genetics*, *197*(2), 573–589.

754 Rex, M. A., & Etter, R. J. (2010). *Deep-sea biodiversity: Pattern and scale*. Harvard
755 University Press.
756 Ricci, M., Peona, V., Guichard, E., Taccioli, C., & Boattini, A. (2018). Transposable
757 elements activity is positively related to rate of speciation in mammals. *Journal of*
758 *Molecular Evolution*, *86*(5), 303–310.
759 Riginos, C., Crandall, E. D., Liggins, L., Bongaerts, P., & Treml, E. A. (2016). Navigating
760 the currents of seascape genomics: How spatial analyses can augment population
761 genomic studies. *Current Zoology*, *62*(6), 581–601.
762 Ritchie, H., Cousins, N. J., Cregeen, S. J., & Piertney, S. B. (2013). Population genetic
763 structure of the abyssal grenadier (*Coryphaenoides armatus*) around the mid-Atlantic
764 ridge. *Deep Sea Research Part II: Topical Studies in Oceanography*, *98*, 431–437.
765 https://doi.org/10.1016/j.dsr2.2013.06.014
766 Ritchie, H., Jamieson, A. J., & Piertney, S. B. (2015). Phylogenetic relationships among
767 hadal amphipods of the Superfamily Lysianassoidea: Implications for taxonomy and
768 biogeography. *Deep Sea Research Part I: Oceanographic Research Papers*, *105*,
769 119–131. https://doi.org/10.1016/j.dsr.2015.08.014
770 Ritchie, H., Jamieson, A. J., & Piertney, S. B. (2017a). Genome size variation in deep-sea
771 amphipods. *Royal Society Open Science*, *4*(9), Art. 9.
772 https://doi.org/10.1098/rsos.170862
773 Ritchie, H., Jamieson, A. J., & Piertney, S. B. (2017b). Population genetic structure of two
774 congeneric deep-sea amphipod species from geographically isolated hadal trenches in
775 the Pacific Ocean. *Deep-Sea Research Part I: Oceanographic Research Papers*, *119*.
776 https://doi.org/10.1016/j.dsr.2016.11.006

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584.

Stewart, H. A., & Jamieson, A. J. (2018). Habitat heterogeneity of hadal trenches: Considerations and implications for future studies. *Progress in Oceanography*, *161*, 47–65. https://doi.org/10.1016/j.pocean.2018.01.007

Taira, K., Kitagawa, S., Yamashiro, T., & Yanagimoto, D. (2004). Deep and bottom currents in the Challenger Deep, Mariana Trench, measured with super-deep current meters. *Journal of Oceanography*, *60*(6), 919–926.

Taylor, M. L., & Roterman, C. N. (2017). Invertebrate population genetics across Earth's largest habitat: The deep-sea floor. *Molecular Ecology*, *26*(19), 4872–4896. https://doi.org/10.1111/mec.14237

Vrijenhoek, R. C. (2010). Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations: *Molecular Ecology*, *19*(20), 4391–4411. https://doi.org/10.1111/j.1365-294X.2010.04789.x

Waples, R. K., Seeb, L. W., & Seeb, J. E. (2016). Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Molecular Ecology Resources*, *16*(1), 17–28. https://doi.org/10.1111/1755-0998.12394

Waples, R. S. (1998). Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, *89*(5), 438–450.

Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, *15*(6), 1419–1439.

Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, *206*(4), 2085–2103.

Wells, J. N., & Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annual Review of Genetics*, *54*, 539–561.

Weston, J. N., Espinosa-Leal, L., Wainwright, J. A., Stewart, E. C., González, C. E., Linley, T. D., Reid, W. D., Hidalgo, P., Oliva, M. E., Ulloa, O., & others. (2021). *Eurythenes atacamensi*s sp. Nov.(Crustacea: Amphipoda) exhibits ontogenetic vertical stratification across abyssal and hadal depths in the Atacama Trench, eastern South Pacific Ocean. *Marine Biodiversity*, *51*(3), 1–20.

Weston, J. N. J., Peart, R. A., Stewart, H. A., Ritchie, H., Piertney, S. B., Linley, T. D., & Jamieson, A. J. (2021). Scavenging amphipods from the Wallaby-Zenith Fracture Zone: Extending the hadal paradigm beyond subduction trenches. *Marine Biology*, *168*(1), 1. https://doi.org/10.1007/s00227-020-03798-4

Weston, J. N., & Jamieson, A. J. (2022). The multi-ocean distribution of the hadal amphipod, *Hirondellea dubia* (Crustacea, Amphipoda). *Frontiers in Marine Science*, *9*, 824640.

Weston, J. N., Jensen, E. L., Hasoon, M. S., Kitson, J. J., Stewart, H. A., & Jamieson, A. J. (2022). Barriers to gene flow in the deepest ocean ecosystems: Evidence from global population genomics of a cosmopolitan amphipod. *Science Advances*, *8*(43), eabo6672.

Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD loci: An efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*, *17*(5), 955–965.

Wolff, T. (1959). The hadal community, an introduction. *Deep Sea Research (1953)*, *6*(C), 95–124. https://doi.org/10.1016/0146-6313(59)90063-2

Wolff, T. (1970). The concept of the hadal or ultra-abyssal fauna. *Deep Sea Research Oceanogrpahic Abstracts*, *17*(6), 983–1003.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 1–13.

Xiao, Y., Xu, T., Sun, J., Wang, Y., Wong, W. C., Kwan, Y. H., Chen, C., Qiu, J.-W., & Qian, P.-Y. (2020). Population Genetic Structure and Gene Expression Plasticity of the Deep-Sea Vent and Seep Squat Lobster Shinkaia crosnieri. *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.587686

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, *16*(5), 284–287.

Data Accessibility Statement

Raw DNA sequence reads are deposited in NCBI SRA bioproject PRJNA834786. Sample genotypes and analysis scripts are available on GitHub repository https://github.com/wenzelm/hgigas_ddRAD.

Benefit-Sharing Statement

Benefits Generated: A research collaboration was developed with scientists from the countries providing genetic samples, the results of research have been shared with the provider communities and the broader scientific community. More broadly, our group is committed to international scientific partnerships, as well as institutional capacity building.

Author Contributions

SP and AJJ designed the research and collected samples. SP performed labwork prior to RADSeq. MW analysed the data. All authors wrote and corrected the manuscript.

Table and Figure legends

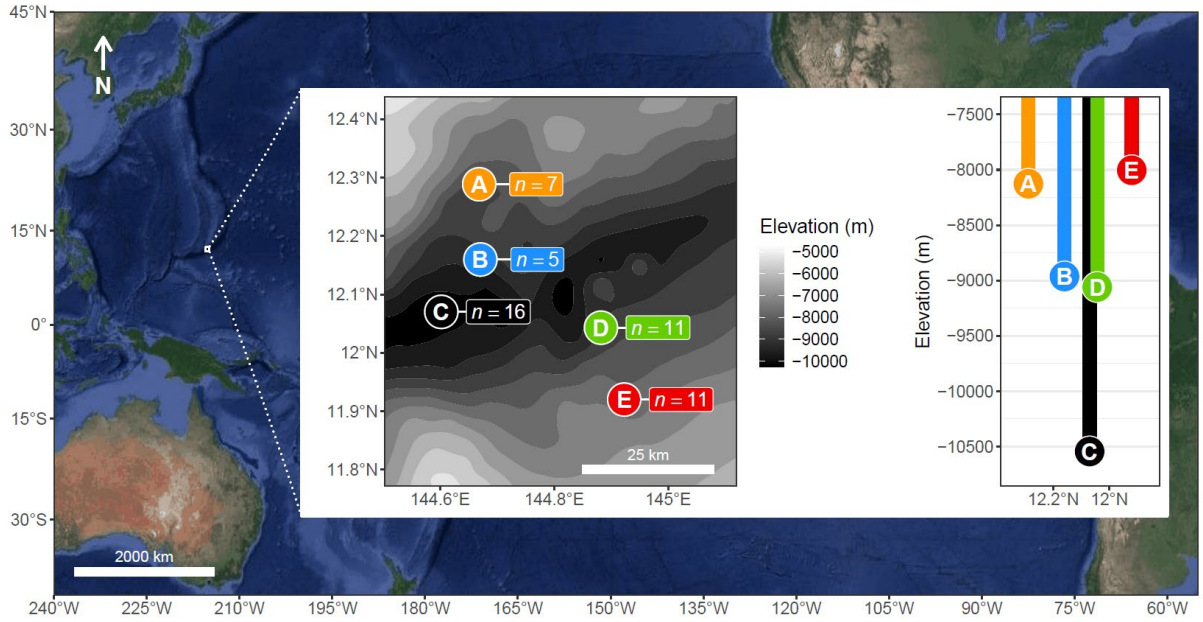


Figure 1: Sampling locations of *Hirondellea gigas* across the Mariana Trench in the Pacific Ocean. Inset left: bathymetric contour map indicating sample coordinates, depth and sample size (*n*). Inset right: sample depths organised by latitude (north-south transect). Maps obtained via *ggmap* (Kahle & Wickham, 2013)and *elevatR* (Hollister et al., 2021)R packages.

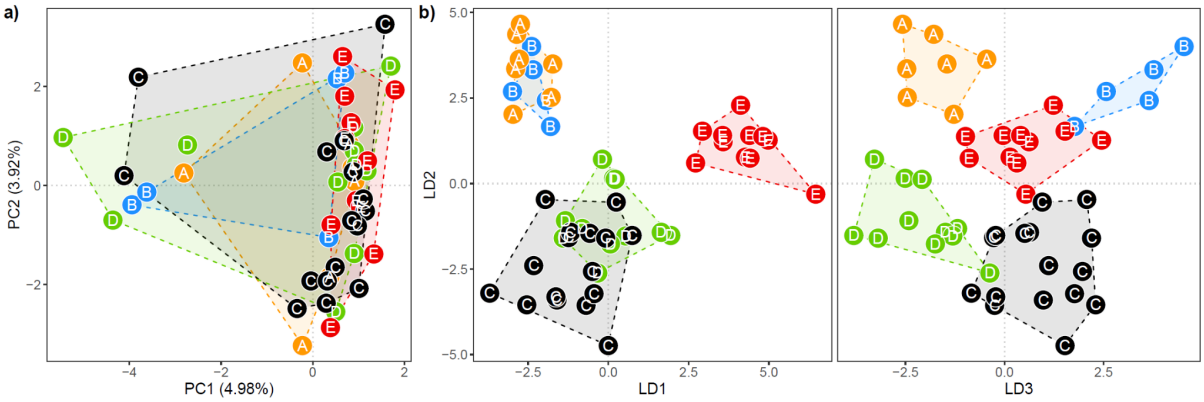


Figure 2: Genetic structure among individuals from five sampling sites (A-E). a) Scatter plot of the first two eigenvectors (PC1 and PC2) following principal components analysis of genotypes. b) Scatter plots of the first three discriminant functions (LD2 vs LD1 and LD2 vs LD3) from discriminant analysis of principal components (DAPC) among genotypes.

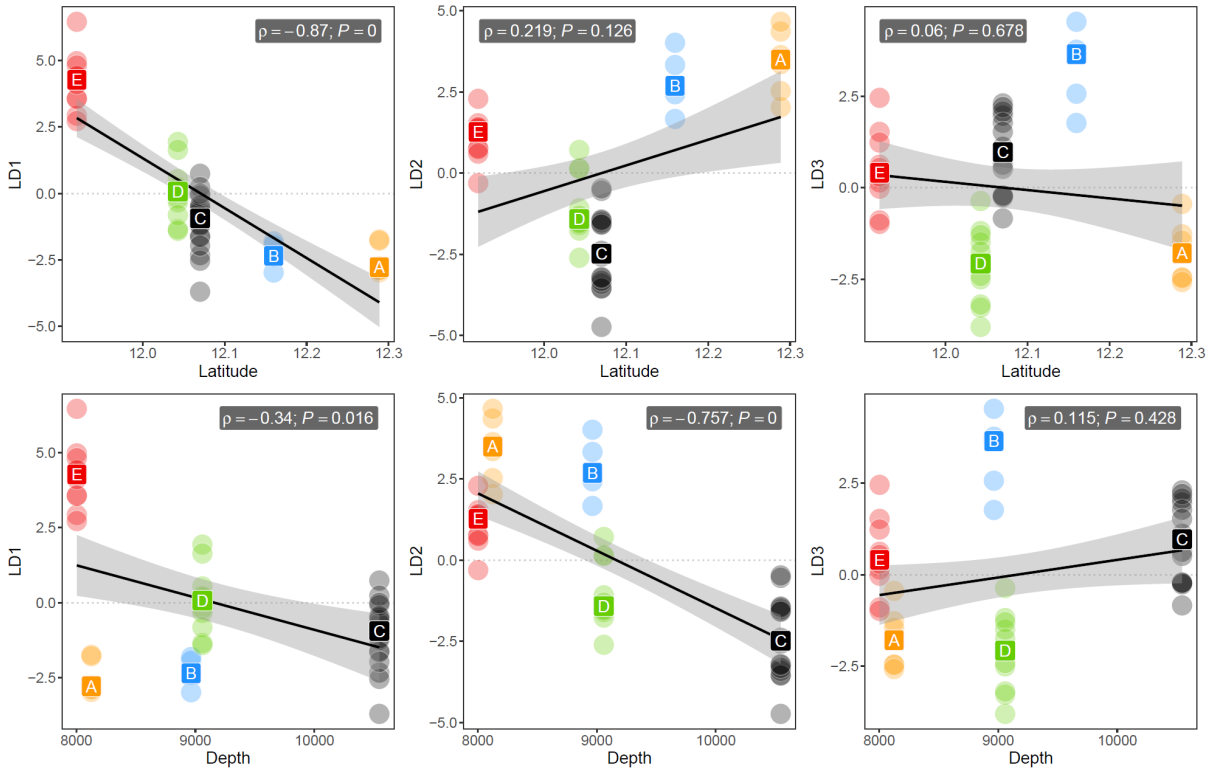
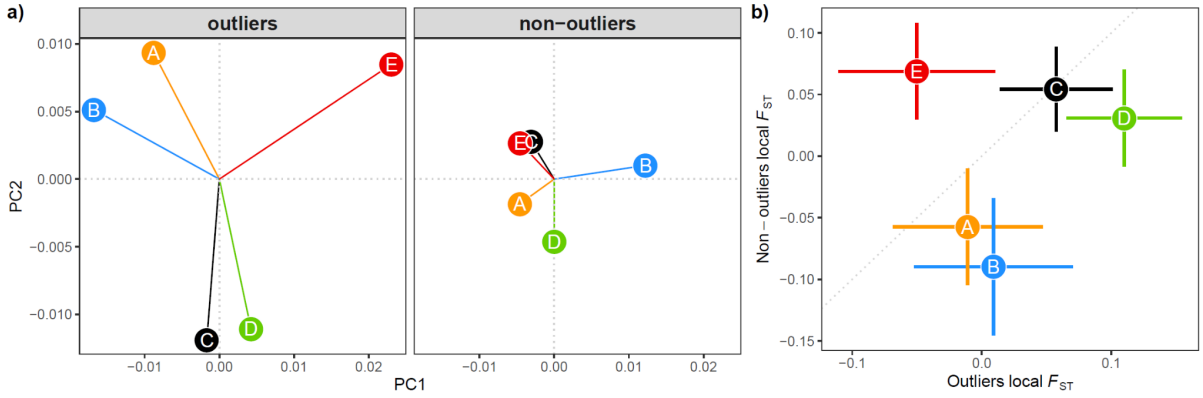


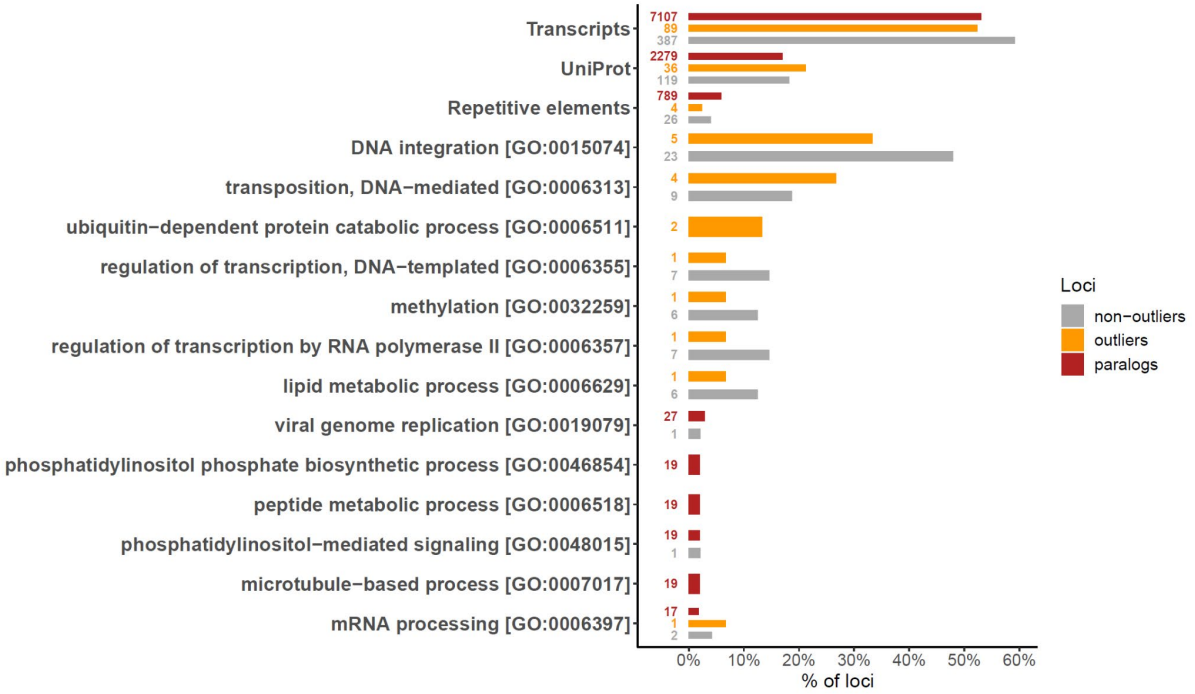
Figure 3. Relationships between genetic structure, latitude and trench depth. Individual scores on the first three linear discriminant functions (LD1, LD2 and LD3) from DAPC are plotted against latitude (top panels) or depth (bottom panels) and are overlaid with a linear regression model. Spearman's rho and *P* value are indicated in panel corners.

877



878

879 **Figure 4.** Relationships between F_{ST}-based genetic structure in outlier versus non-outlier SNPs
880 among sampling sites. a) Scatter plots of the first two principal components (PC1 and PC2) of
881 pairwise F_{ST}. b) Scatter plot of population-specific (local) F_{ST} ± 2*SE with a hypothetical 1:1
882 relationship indicated by a dotted grey line.
883



884

885 **Figure 5**. Functional annotation of ddRAD loci identified as paralogs, divergence outliers or
886 singleton non-outliers. Numbers and proportions are given for loci annotated as transcripts
887 (*Hirondellea gigas* transcriptome), UniProt proteins (arthropoda), repetitive elements
888 (RepeatMasker) and the six most frequently observed GeneOntology (*biological process*)
889 terms in each locus category. Percentages for GO terms are ratios based on total loci with GO
890 annotations.

891

892