# Distinct genomic routes underlie transitions to specialised symbiotic lifestyles in deep-sea annelid worms

Giacomo Moggioli[1], Balig Panossian [1], Yanan Sun[2,3,4], Daniel Thiel[5], Francisco M. Martín-Zamora [1], Martin Tran[1], Alexander M. Clifford[6], Shana K. Goffredi [7], Nadezhda Rimskaya-Korsakova[8], Gáspár Jékely [5], Martin Tresguerres [6], Pei-Yuan Qian [2,4], Jian-Wen Qiu [3,4], Greg W. Rouse [6], Lee M. Henry [1] ✉ & José M. Martín-Durán [1] ✉

Bacterial symbioses allow annelids to colonise extreme ecological niches, such as hydrothermal vents and whale falls. Yet, the genetic principles sustaining these symbioses remain unclear. Here, we show that different genomic adaptations underpin the symbioses of phylogenetically related annelids with distinct nutritional strategies. Genome compaction and extensive gene losses distinguish the heterotrophic symbiosis of the bone-eating worm *Osedax frankpressi* from the chemoautotrophic symbiosis of deep-sea Vestimentifera. *Osedax*'s endosymbionts complement many of the host's metabolic deficiencies, including the loss of pathways to recycle nitrogen and synthesise some amino acids. *Osedax*'s endosymbionts possess the glyoxylate cycle, which could allow more efficient catabolism of bone-derived nutrients and the production of carbohydrates from fatty acids. Unlike in most Vestimentifera, innate immunity genes are reduced in *O. frankpressi*, which, however, has an expansion of matrix metalloproteases to digest collagen. Our study supports that distinct nutritional interactions influence host genome evolution differently in highly specialised symbioses.

Symbioses have shaped life on Earth, from the origin of the eukaryotic cell to the formation of biodiversity hotspots such as coral reefs[1,2]. Animal chemosynthetic symbioses, where bacteria convert inorganic compounds to organic matter, are ubiquitous in marine habitats[3] and fuel some of the most productive communities, such as those around hydrothermal vents[4]. Siboglinid worms (Annelida) often dominate deep-sea chemosynthetic environments through symbioses with environmentally acquired bacteria[5,6] that adults harbour within a specialised organ called a trophosome[7]. Despite their ecological importance, the host's genetic traits sustaining these symbioses have only been studied in Vestimentifera[8–10], one of the four main lineages in Siboglinidae (Fig. 1a). The genomes of *Lamellibrachia luymesi*[8], *Paraescarpia echinospica*[9], *Riftia pachyptila*[10] and *Ridgeia piscesae*[11] have revealed a complex molecular interplay between Vestimentifera and their endosymbionts to fulfil their nutritional demands[12]. For example, the hosts have lost genes involved in amino acid biosynthesis[8,10,11] and carbohydrate catabolism[9] but expanded gene families involved in nutrient transport[8], gas exchange[8–10,13,14], innate immunity[9,11] and

[1]School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, E1 4NS London, UK. [2]Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong, China. [3]Department of Biology, Hong Kong Baptist University, Hong Kong, China. [4]Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China. [5]Living Systems Institute, University of Exeter, Exeter, UK. [6]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093, USA. [7]Department of Biology, Occidental College, Los Angeles, LA, USA. [8]Friedrich Schiller University Jena, Faculty of Biological Sciences, Institute of Zoology and Evolutionary Research, Erbertstr. 1, 07743 Jena, Germany. ✉e-mail: l.henry@qmul.ac.uk; chema.martin@qmul.ac.uk
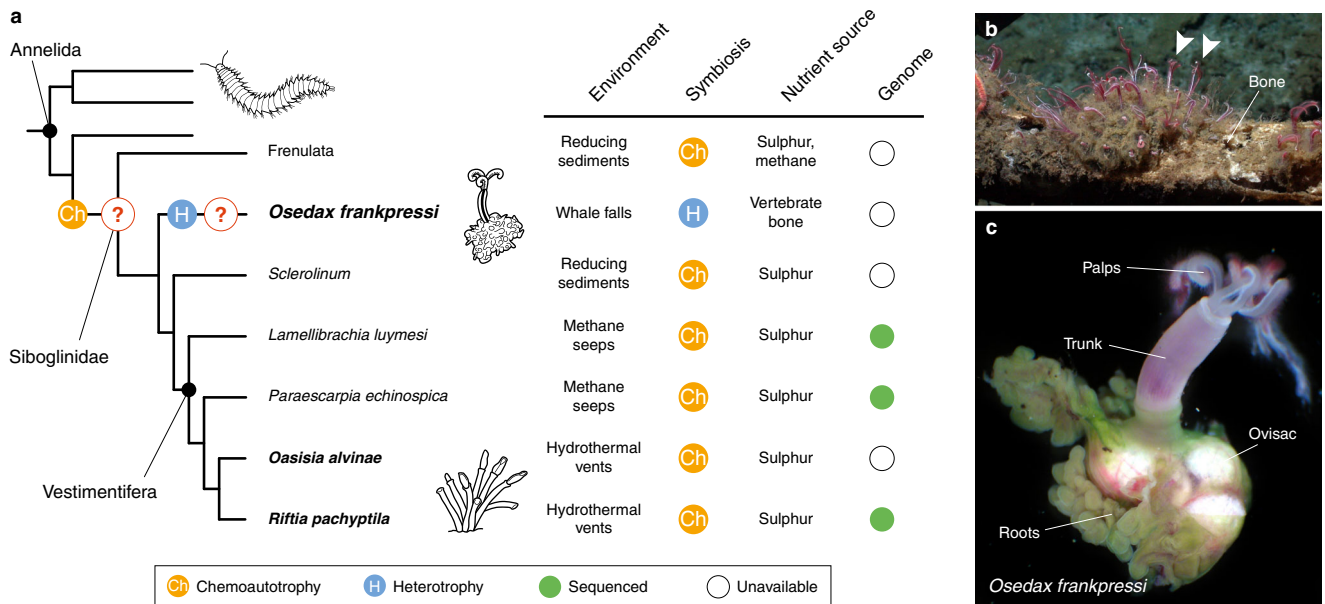
**Fig. 1 | Siboglinidae is a symbiotic annelid group. a** Siboglinidae is a diverse clade of annelid worms that evolved chemosynthetic symbioses (left side). There are four main lineages within Siboglinidae, namely Frenulata, *Osedax*, *Sclerolinum* and Vestimentifera. Chemolithoautotrophy occurs in Frenulata, *Sclerolinum* and Vestimentifera, which associate with gammaproteobacteria that employ sulphur or methane to produce organic compounds in an array of marine ecosystems, from reducing sediments to methane seeps and hydrothermal vents (right side of the panel). Differently, *Osedax* worms (e.g., *O. frankpressi*; **b**, **c**) have secondarily evolved a heterotrophic association with *Oceanospirillales* to exploit decaying vertebrate bones. The genomic basis for the evolution of these nutritional symbioses in Siboglinidae is unclear (question marks on the left) because genomic information only exists for Vestimentifera hosts (green circles on the right). The species herein studied are highlighted in boldface. **b**, **c** Photographs of *O. frankpressi* in a whale bone (**b**; arrowheads point to *O. frankpressi*) and a mature female adult (**c**). *O. frankpressi* settles and colonises decaying vertebrate bones (**b**). There, the posterior part of the body becomes stably infected with environmentally acquired *Oceanospirillales* bacteria. This body part (the so-called roots) harbours the bacteria and grows to penetrate the bone, dissolving the organic components. These nutrients are absorbed and transported towards the bacteriocytes containing the endosymbionts, which will proliferate and act as food for the worm. Anterior to the root tissue there are the reproductive ovisacs and the head bears two pairs of palps.

lysosomal digestion[8–10,15]. On the other hand, there is genomic information for the endosymbionts of most major clades of Siboglinidae, including Vestimentifera, *Osedax* and Frenulata[16–18]. The endosymbionts of Vestimentifera and Frenulata are mixotrophs[19] and show a diverse metabolic repertoire for energy production (e.g., the reductive tricarboxylic acid cycle in the endosymbionts of Vestimentifera) and nutrient biosynthesis that complements the metabolic deficiencies of, at least, the vestimentiferan host[16,17]. In addition, an increase in the genetic repertoire to infect and evade the host's immunity[16–18], transport nutrients[18] and metabolise nitrogen compounds[16,17] is common in endosymbionts of Siboglinidae. Notably, many of these genetic changes also occur in other distantly related chemosymbiotic animals, including bivalves[20], gastropods[21], and the clitellate annelid *Olavius algarvensis*[22]. Therefore, disparate animal groups have convergently evolved similar genetic mechanisms to sustain different chemosynthetic symbioses in marine ecosystems.

Within Siboglinidae, the marine *Osedax* annelids have evolved a unique endosymbiosis[23–27] with heterotrophic bacteria in the order *Oceanospirillales*[18,24,28–30] (Fig. 1a) that allows them to obtain nutrients from bones of dead animals lying on the ocean's floor (Fig. 1b). While *Osedax* shares some morphological features with other siboglinids[31], including the lack of a gut, mouth and anus, *Osedax* contains bacteriocytes concentrated in the subepidermal connective tissue of the lower trunk that grows directly into the bone[24,28] (Fig. 1c). This amorphous tissue, referred to as "roots", expresses high levels of V-type H⁺-ATPase and carbonic anhydrase[32], indicating acid is used to dissolve the bone matrix to access collagen and lipids, which are then absorbed across the root epithelium. Enzymatic[28,29] and transcriptomic data[33] support this theory by showing that the roots of *Osedax* express many proteases and solute carrier transporters that are thought to be involved in bone degradation and nutrient absorption, perhaps with the aid of the endosymbionts[18]. However, it is currently unclear whether the specialised heterotrophic symbiosis of *Osedax* is based on homologous genetic traits to those discovered in Vestimentifera and other chemoautotrophic invertebrates or if it relies on unique genomic adaptations. Untangling the molecular mechanisms behind this remarkable symbiosis is, therefore, central to understanding the evolution of *Osedax* and Siboglinidae, as well as the ecological principles and succession of bone-eating communities[34].

In this study, we sequenced the genome of *Osedax frankpressi* Rouse, Goffredi & Vrijenhoek, 2004[24], as well as that of two vent-dwelling Vestimentifera, *Oasisia alvinae* Jones, 1985 and *Riftia pachyptila* Jones, 1981, and compared them with nearly 40 eukaryote and prokaryote genomes to better understand the genomic changes leading to these distinct symbiotic lifestyles. In contrast to Vestimentifera, we found that *O. frankpressi* has a small AT-rich genome with a reduced gene repertoire. Gene families typically expanded in chemosymbiotic hosts, such as innate immunity components, are reduced in *O. frankpressi*. Instead, the *Osedax-Oceanospirillales* symbiosis has unique genomic adaptations for bone digestion, including the loss of biosynthetic pathways of amino acids that are abundant in vertebrate bones in the host, the presence of the glyoxylate cycle in the endosymbiont that could allow the production of carbohydrates from the lipids present in vertebrate bones, and the expansion of matrix metalloproteases in the host that could aid in bone digestion. Together, our findings demonstrate that different genomic principles sustain the nutritional symbioses of *Osedax* and Vestimentifera, providing critical insight into the genetic and metabolic changes that have enabled symbiotic siboglinids to colonise diverse nutrient-imbalanced feeding niches.
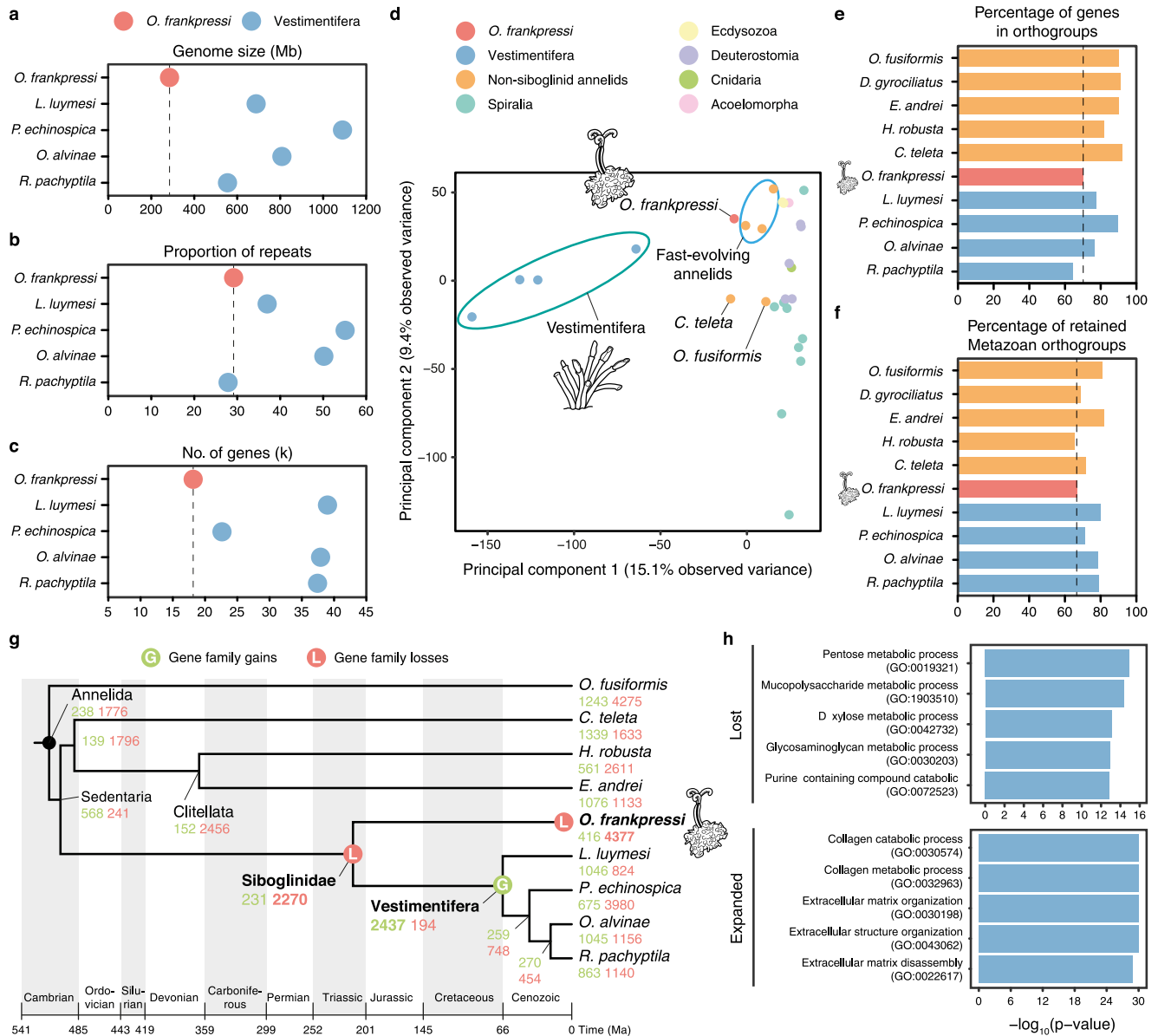
**Fig. 2 | *Osedax* and Vestimentifera exhibit distinct genome evolutionary trends. a–c** Plots comparing genome size (**a**), repeat content (**b**) and number of genes (**c**) between *O. frankpressi* and the four Vestimentifera with sequenced genomes. *Osedax frankpressi* has a smaller genome, with less genes but relatively similar repeat content. **d** Principal component analyses of the gene content of 28 metazoan genomes show that differently from symbiotic bivalves and gastropods, the gene content of Vestimentifera and *O. frankpressi* differs from slow-evolving asymbiotic species (as represented by *Owenia fusiformis* and *C. teleta*). While Vestimentifera has a unique gene content, *O. frankpressi* is like other fast-evolving annelid lineages. **e**, **f** Bar plots of the percentage of genes in gene families (i.e., orthogroups; **e**) and retained ancestral metazoan gene families (**f**) for ten annelid

lineages. *Osedax frankpressi* is amongst the annelids with less genes in gene families and less retained ancestral metazoan genes. **g** Patterns of gene family gains (in green) and loss (in red) during the evolution of Annelida under a consensus tree topology[31] and a consensus of published molecular dates[8,9]. A major event of gene loss is common to all Siboglinidae. While *O. frankpressi* continued experiencing high rates of gene loss, a major event of gene innovation is common to all Vestimentifera. **h** Top five enriched gene ontology terms (Biological Process) for gene families lost (top) and expanded (bottom) in *O. frankpressi*. While *O. frankpressi* has further lost genes involved in metabolism (e.g., carbohydrate metabolism), genes involved in collagen and extracellular matrix degradation are expanded. *P*-values were derived from upper-tail Fisher's exact tests.

## Results

### The genome of *O. frankpressi*

To identify genomic signatures that could inform the genetic and physiological basis of the heterotrophic symbiosis in *Osedax*, we used long PacBio reads and short Illumina reads to assemble the genome of *O. frankpressi*[24] (Supplementary Table 1). We also sequenced the genomes of two Vestimentifera from hydrothermal vents, *Oasisia alvinae* and *R. pachyptila* (Supplementary Fig. 1), complementing previous genome sequencing efforts[8–10]. We generated almost entirely haploid draft assemblies (Supplementary Fig. 2a–d), which included the

circularised endosymbiont genomes of *O. frankpressi* and *Oasisia alvinae* and several epibionts associated with *O. frankpressi* (Supplementary Fig. 2h–j; Supplementary Table 2). Consistent with *k*-mer-based analyses (Supplementary Fig. 2e–g), previously reported genome size estimation for *Oasisia alvinae*[35], and a recent genome assembly of *R. pachyptila*[10], the assembled genomes for *O. frankpressi*, *Oasisia alvinae* and *R. pachyptila* span 285 Mb (1,185 scaffolds with an N50 of 426 Kb), 808 Mb (642 scaffolds with an N50 of 2.975 Mb) and 554 Mb (918 scaffolds with an N50 of 1.424 Mb) after removal of bacterial contigs, respectively (Fig. 2a; Supplementary Fig. 2k).

The genome assemblies for *Oasisia alvinae* and *R. pachyptila* shows high completeness (96.9% and 95.6% BUSCO presence, respectively; Supplementary Fig. 2l; Supplementary Table 3). The assembly for *O. frankpressi* appeared to have lower completeness (80.1% BUSCO presence; Supplementary Fig. 2l). However, 95.62% and 97.77% of the de novo assembled transcripts from the body and root tissue mapped to the genome assembly of *O. frankpressi*, respectively. Accordingly, BUSCO completeness increased to a final score of 96.23% after gene annotation (Supplementary Fig. 2l) and manual curation (26 out of the 62 missing BUSCO could be manually annotated; Supplementary Data 1). Together, this suggests that the fast rates of molecular evolution in coding sequences observed in *Osedax* worms[36] are likely responsible for the relatively low initial, assembly-based BUSCO completeness in the genome of *O. frankpressi*.

Although the genome of *O. frankpressi* is ~50–75% smaller than the sequenced genomes and estimated genome sizes of Vestimentifera[8–10,35] (Fig. 2a), the fraction of simple repeats and transposable elements in *O. frankpressi* (29.16%) is comparable to that of the vestimentiferan *R. pachyptila* (27.87%) and asymbiotic annelids with similar genome sizes (Fig. 2b; Supplementary Fig. 3a). As in Vestimentifera, the repeat landscape in *O. frankpressi* shows signs of expansions (Supplementary Fig. 3b), unlike in asymbiotic annelids with slow rates of molecular evolution[37,38]. Combining transcriptomic evidence (Supplementary Table 1) with ab initio gene prediction (Supplementary Fig. 2a), we functionally annotated 37,777 and 38,179 protein-coding transcripts in *Oasisia alvinae* and *R. pachyptila*, respectively (Supplementary Fig. 2k), which have a similar number of genes to other Vestimentifera and asymbiotic annelids[8,37,38]. The number of genes annotated in our assembly for *R. pachyptila* is higher than in a previous report[10] (Supplementary Fig. 4a). Still, both annotations and assemblies are broadly equivalent (Supplementary Fig. 4b–d). Unlike Vestimentifera, *O. frankpressi* has a smaller repertoire of 18,657 transcripts (Fig. 2c), comparable to that of the miniaturised *Dimorphilus gyrociliatus*[36], another annelid species with a compact genome and a streamlined gene set (14,203 genes). Therefore, *O. frankpressi* has the smallest genome of all sequenced siboglinids. Given the number of genes in genomes of asymbiotic annelids, gene loss rather than removal of repeat content seems to account for the genome size difference between these two lineages of Siboglinidae.

## Gene gains and losses shape the evolution of Siboglinidae

To investigate gene content evolution between major lineages of Siboglinidae, we first reconstructed the gene families of 28 highly complete metazoan genomes, including seven symbiotic annelid and molluscan lineages (Supplementary Data 2). This taxonomic sampling provides sufficient resolution to infer the time of origin of each gene family while minimising potential biases in orthology inference in fast-evolving species[39]. A principal component analysis of the number of orthologs per gene family in the 28 species clustered the symbiotic molluscs *Bathymodiolus platifrons*[20] and *Gigantopelta aegis*[21] with their asymbiotic bivalve and gastropod relatives, respectively (Supplementary Fig. 5a). However, the four Vestimentifera species are markedly differentiated from the other annelid and animal genomes, and *O. frankpressi* is closer to heterotrophic annelids with fast rates of molecular evolution and divergent gene repertoires, such as the leech *Helobdella robusta* and the earthworm *Eisenia andrei*—which also harbour bacterial symbionts[40–42]—and the marine worm *D. gyrociliatus* (Fig. 2d; Supplementary Fig. 5a). Indeed, after *R. pachyptila*, *O. frankpressi* is the annelid with the second lowest percentage of genes assigned to gene families (Fig. 2e) and has only retained a fraction of ancestral metazoan gene families comparable to more rapidly evolving annelids such as *H. robusta* and *D. gyrociliatus* (Fig. 2f). Therefore, unlike symbiotic molluscs, the evolution of nutritional symbioses in

Siboglinidae correlates with divergent host gene repertoires compared to their asymbiotic annelid counterparts.

To identify and characterise the evolutionary events underpinning the divergent gene repertoires of Siboglinidae, we reconstructed the patterns of gene family evolution in those 28 metazoan genomes under a consensus tree topology (Supplementary Fig. 5b). Vestimentifera and *O. frankpressi* share a major gene loss event involving 2270 gene families of mainly ancient origins (61.23% of the lost families originated before Metazoa and the Bilateria/Nephrozoa ancestor) (Fig. 2g) and enriched in Gene Ontology (GO) terms associated with metabolism (Supplementary Fig. 5c). This loss thus coincides with the evolution of nutritional symbioses in the last common ancestor of Siboglinidae. A high rate of gene loss continued in the *O. frankpressi* lineage (Fig. 2g), which ultimately accounts for its reduced gene repertoire and primarily affected genes associated with carbohydrate and nitrogen metabolism (Fig. 2h; Supplementary Fig. 5d). Notably, Vestimentifera experienced an event of gene family expansion in its last common ancestor (2,437 gene families), mainly affecting genes related to immunity, cell communication, and response to stimuli[9] (Fig. 2g; Supplementary Fig. 5e). However, high lineage-specific rates of gene loss also occur in some Vestimentifera[10,11], as in *O. frankpressi* (Fig. 2g). Compared to Vestimentifera, *O. frankpressi* has had few gene family gains (Supplementary Fig. 5b) but has experienced a large expansion of gene families associated with extracellular matrix remodelling and degradation (e.g., collagen degrading proteases; Fig. 2h; Supplementary Fig. 5f) in agreement with previous transcriptomic observations[33]. Altogether, our findings indicate that the evolution of symbiosis in *Osedax* and Vestimentifera relies on different host gene repertoires, one sculptured predominantly through gene loss (in *O. frankpressi*) and another through ancestral gene gains followed by varying, species-specific rates of gene loss (in Vestimentifera)[8–10] (Fig. 2g).

## The different genomic traits of Siboglinidae endosymbionts

To investigate the genetic and functional contribution of the endosymbionts to the nutritional symbioses of Siboglinid worms, we used our PacBio long-read data to assemble the genomes of the primary endosymbionts of *O. frankpressi* (Rs1 ribotype; Genome Taxonomy Database accession number Rs1 sp000416275) (Fig. 3a; Supplementary Data 3), and *Oasisia alvinae* (Supplementary Fig. 6; Supplementary Data 4), as well as several epibionts associated with *Osedax*[43] (Supplementary Table 2). The circularised assembly of the endosymbionts of *O. frankpressi* improved the previously published genome[18], revealing 95 new functional genes that provide additional insights into its symbiosis (Supplementary Data 5). Compared to deep-sea free-living relatives, the *O. frankpressi* endosymbiont has a genome enriched in metabolic genes for protein secretion systems, carbohydrate metabolism, and coenzyme and amino acid biosynthesis (Supplementary Data 6b). This includes additional virulence factors, such as multiple complete copies of Type 5a, 5b, and 6i secretion system pathways (Supplementary Data 6c) that are important for modulating interactions with other bacteria and eukaryotic hosts. *Neptunomonas japonica*, a close relative of the *Oceanospirillales* endosymbionts recovered from marine sediments near a whale fall, has many of the same metabolic capabilities of the endosymbionts; however, it lacks the additional secretion systems[44]. The Type 5a and Type 5b secretion systems are also largely absent in the endosymbionts of Vestimentifera (Supplementary Data 7g). This increase in virulence factors may reflect that *Oceanospirillales* repeatedly infect the roots of *Osedax* as it grows through bone material, unlike the trophosome of Vestimentifera, which is colonised early during host development[7,45]. In addition, all Siboglinidae endosymbionts contain numerous genes encoding eukaryote-like protein domains, which, interestingly, tend to be host-lineage-specific (Supplementary Data 8). Eukaryote-like proteins
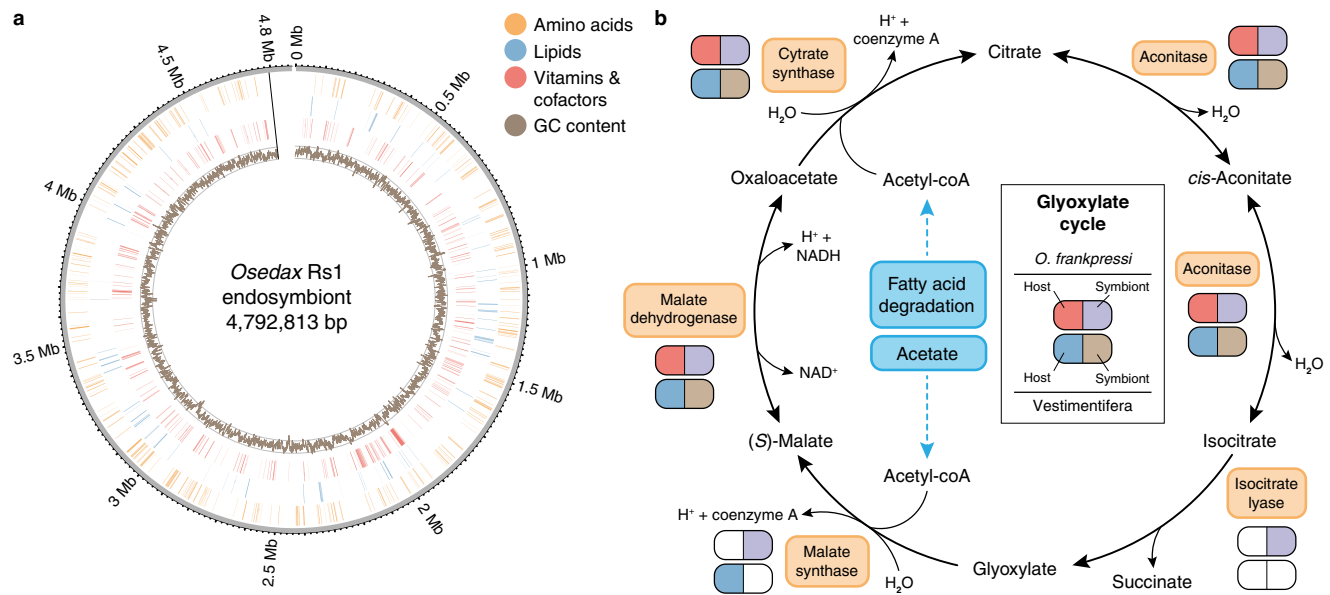
**Fig. 3 | *Osedax* and its endosymbiont reconstruct the glyoxylate cycle.**
**a** Circular schematic representation of the genome of *Osedax* endosymbiont Rs1, assembled into a single contig. The plot shows the genomic location of genes involved in amino acid, lipid and vitamin/cofactor metabolism (in orange, blue and red, respectively) and the GC content (inner circle; brown colour).
**b** *Oceanospirillales* endosymbionts possess the glyoxylate cycle, a metabolic pathway that can produce oxaloacetate, which can serve as the precursor to synthesise carbohydrates from the oxidation of fatty acids. This metabolic pathway could thus contribute to synthesising glucose in a diet (the bone) that is naturally poor in carbohydrates. Notably, this molecular and metabolic interaction does not occur between Vestimentifera and its symbionts because the host and microbes lack the enzyme isocitrate lyase.

modulate important processes in many symbioses, including extracellular secretions, cell binding and colonisation[12,46,47]. Therefore, the specificity of the endosymbionts' eukaryote-like proteins in the different lineages of Siboglinidae suggests they may be important for host and clade-specific annelid-symbiont communication, as shown in *Riftia*[12].

*Osedax frankpressi*'s endosymbionts and those of Vestimentifera and Frenulata shared a broadly similar repertoire of genes involved in core cellular processes (Supplementary Data 7a). However, as we may expect for a heterotrophic microbe, the *Oceanospirillales* endosymbionts have significantly more genes involved in the metabolism and uptake of amino acids, coenzymes, lipids, and carbohydrates (Supplementary Fig. 7a–c; Supplementary Data 7a, e, f). This includes several complete pathways to convert oxaloacetate into ribose 5-phosphate that can be used in the biosynthesis of nucleotides and histidine, the Entner-Doudoroff and De Ley-Doudoroff pathways to catabolise carbohydrates, and multiple sugar, amino acid and oligopeptide ATP-binding transporters (Supplementary Data 7). In addition, the endosymbionts of *O. frankpressi* can produce all essential amino acids (including methionine and threonine) and vitamin B6, unlike Vestimentifera and Frenulata endosymbionts, as well as vitamin B2, which Vestimentifera symbionts cannot make (Fig. 4a; Supplementary Data 7d, e). Notably, the B2 pathway was considered missing in the previous draft genome of the *Oceanospirillales* endosymbiont[18], but it is present in ours. As in some of the bacteria comprising the microbiome of degrading bones[48], *O. frankpressi*'s endosymbionts can catabolise hydroxyproline, one of the most abundant amino acids in collagen[49,50], but lacks a secreted M9 peptidase to cleave extracellular collagen (Supplementary Data 3). Finally, all endosymbionts of Vestimentifera are enriched in genes involved in chemosynthesis, most of which are absent in the heterotroph endosymbionts of *Osedax* (Supplementary Fig. 7c). Taken together, our results confirm and expand previous genomics efforts on the *Oceanospirillales* endosymbionts[18], further demonstrating that Siboglinidae has partnered with metabolically versatile microbes that are suited to sustain symbioses with eukaryotes in diverse environments.

## Metabolic adaptations for bone digestion

Vertebrate bones are nutrient-imbalanced food sources enriched in lipids and proteins and deficient in carbohydrates[50]. Given the reduced gene repertoire of the host (Fig. 2c) and the metabolic versatility of its endosymbionts[18] (Supplementary Data 7), we next explored potential molecular and metabolic interactions that could facilitate the nutritional specialisation of *O. frankpressi*. Combining highly sensitive profile hidden Markov Models sequence similarity searches with KEGG and COG functional annotations, we reconstructed all metabolic routes in *O. frankpressi* and its endosymbionts and the published genomes of Vestimentifera and their respective endosymbionts. *Osedax frankpressi* and Vestimentifera have similar metabolic capabilities to produce and process lipids (Supplementary Data 9). However, *O. frankpressi* has six incomplete pathways for carbohydrate metabolism that are intact in Vestimentifera and other asymbiotic annelids (Supplementary Data 9), consistent with the loss of gene families involved in carbohydrate metabolism (Fig. 2h). In one case (UDP-N-acetyl-D-glucosamine biosynthesis), the endosymbiont possesses the enzymes that would complement the losses in *O. frankpressi* (Supplementary Data 7c, Supplementary Data 9). Notably, the endosymbionts, unlike the host, lack the enzymes glycogen synthase and glycogen phosphorylase and, therefore, cannot produce glycogen (Supplementary Data 7) and possess the enzymes to complete the glyoxylate cycle (Fig. 3b, Supplementary Data 7), which allows the production of glucose from the catabolism of fatty acids and acetate[51–54]. This metabolic pathway does not occur in Vestimentifera because both the host and endosymbiont lack an isocitrate lyase (Fig. 3b). Therefore, the glyoxylate cycle may play a role in the metabolic interaction of *Osedax* and its endosymbionts by collectively converting bone lipids into carbohydrates, which are often nearly absent in bones[50]. Although *Osedax* appears to use wax esters to store energy[29], the fat content of bones varies widely, and *Osedax* can grow in dentin (Greg W. Rouse and Shana K. Goffredi, personal observation), where lipids are a minor component. Functional studies are thus warranted to assess the nutritional and physiological relevance of this metabolic pathway in *Osedax* and under different nutritional sources.
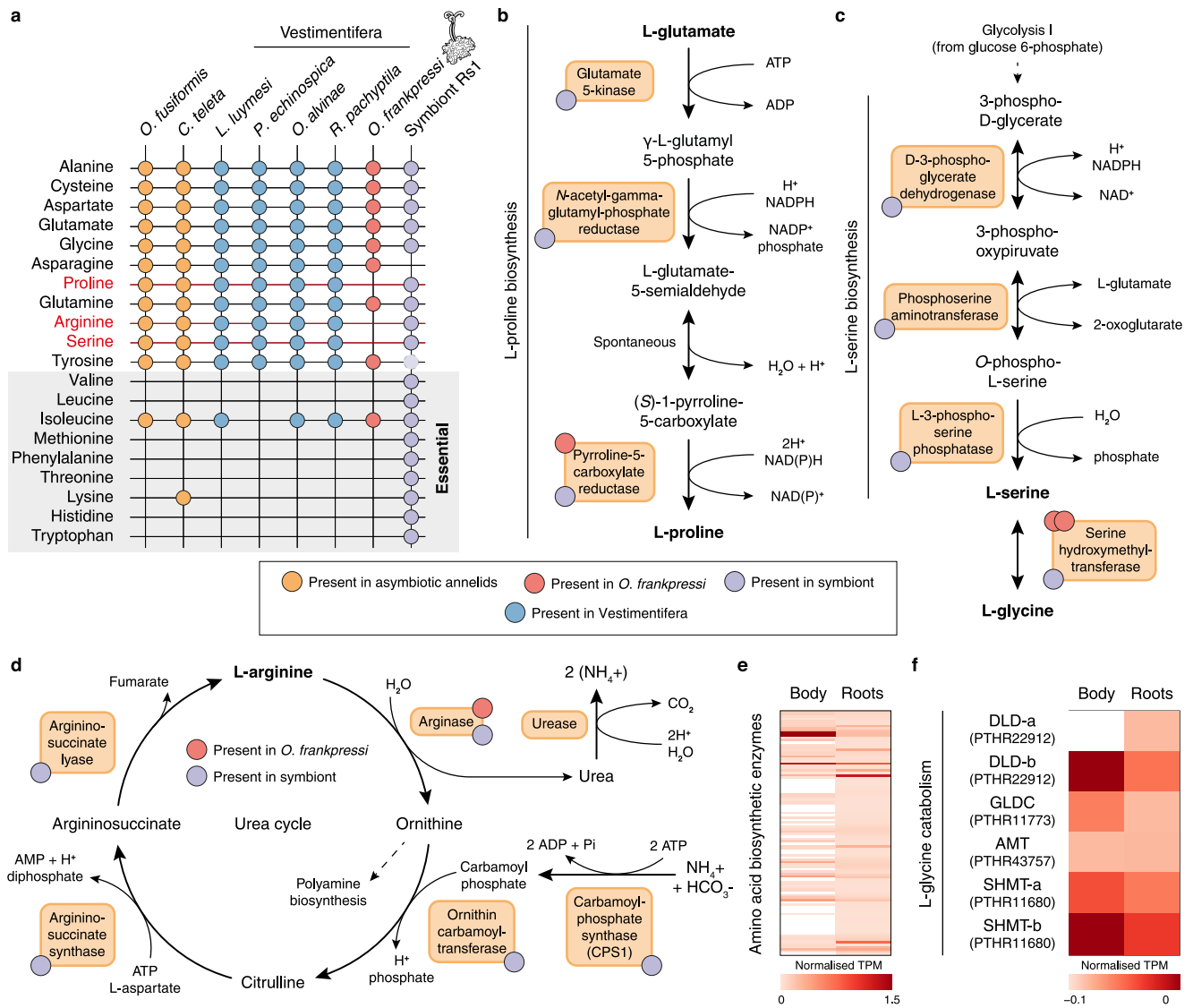
**Fig. 4 | *Osedax*'s metabolic adaptations to bone digestion. a** Summary table of the presence (filled circles) and absence (empty crosses) of amino acid biosynthetic pathways in seven annelid genomes and *O. frankpressi* endosymbiont (symbiont Rs1). While Vestimentifera and asymbiotic annelids can synthesise all amino acids that are non-essential and conditional for humans, *O. frankpressi* shows incomplete pathways to synthesise proline, arginine, and serine (in red). Some of these amino acids are abundant in the bone (e.g., proline) and all can be produced by the symbiont (tyrosine biosynthetic pathway is truncated in the symbiont; dotted and lighter circle). **b**–**d** Schematic representation (as in MetaCyc database) of the biosynthetic pathways for proline (**b**), serine (**c**) and arginine (**d**) indicating with red and violet circles the enzymes present in *O. frankpressi* and its endosymbiont, respectively. *Osedax frankpressi* cannot produce serine from glycolytic metabolites but can either produce serine from collagen-derived glycine or take it from the diet. In addition, *O. frankpressi* can only convert arginine into ornithine, producing urea as a result. **e**, **f** Heatmaps of normalised mRNA expression levels for amino acid biosynthetic enzymes (**e**) and glycine catabolising enzymes (**f**) in the body and roots of *O. frankpressi*. Biosynthetic enzymes (**e**), including the two copies of serine hydroxymethyltransferase (SHMT-a and SHMT-b) that convert glycine into serine, are more expressed in the roots than in the body of *O. frankpressi*. Source data for (**e**, **f**) are provided as a Source Data file.

Proteins, predominantly collagen[50], are the core organic component of bone. Collagen is rich in proline/hydroxyproline and glycine[49], and thus its amino acid composition is also imbalanced. Consistent with previous genomic analyses[8–10], Vestimentifera and asymbiotic annelids (*Owenia fusiformis* and *C. teleta*) can produce all non-essential and conditionally essential amino acids. However, *O. frankpressi* cannot synthesise the amino acids proline, serine, and arginine (which are non-essential or conditional for mammals), but its endosymbionts can (Fig. 4a). Indeed, only one enzyme (pyrroline-5-carboxylate reductase) of the proline biosynthetic pathway remains, which is expressed at similar levels in the roots and the rest of the body, unlike most amino acid biosynthetic enzymes that are enriched in roots (Fig. 4b, e). Similarly, the entire pathway to synthesise serine from intermediates of glycolysis is missing in *O. frankpressi* (Fig. 4c). However,

*O. frankpressi* (like other annelids) has an intact glycine cleavage system (Fig. 4f), which would favour the conversion of collagen-derived glycine into serine through serine hydroxymethyltransferase[55]. The two copies of this enzyme are highly expressed throughout *O. frankpressi* (Fig. 4f) and could provide an additional source of serine on top of those offered by the diet and endosymbionts. Therefore, *O. frankpressi* shows genomic-inferred metabolic adaptations to its unique bone-eating diet in its gene complement, which differs from the more intact metabolic repertoire of Vestimentifera and other asymbiotic annelids[12].

The catabolism of amino acids produces ammonia, a compound that can be toxic but can also serve as a substrate for amino acid biosynthesis by both animals and bacteria. Most aquatic organisms excrete excess ammonia into the water, but a few aquatic animals and

most air-breathing vertebrates shuttle ammonia into the urea cycle leading to urea production[56]. *Osedax frankpressi* lacks four urea cycle enzymes and only possesses arginase (Fig. 4d). Interestingly, the urea cycle is also incomplete in the leech *Poecilobdella granulosa*[57], another symbiotic heterotrophic annelid with a protein-rich diet that excretes ammonia as a waste product. In *O. frankpressi*, the lack of CPS1 is especially significant because this enzyme is the rate-limiting step that mediates the entry of ammonia into the urea cycle; in fact, CPS1 genetic deficiency in humans leads to episodic toxic ammonia levels in the blood ("hyperammonemia")[58]. However, *O. frankpressi* additionally lacks urease; therefore, this enzyme is not available to convert ammonia (and carbon dioxide) into urea, thus ensuring elevated internal ammonia levels. The only enzyme present in the urea cycle of *O. frankpressi* is arginase, which catalyses the interconversion of arginine—which the worm likely obtains from bone-derived collagen and the endosymbionts (Fig. 4a)—into ornithine and urea. Although the urea produced by this pathway can be expected to be negligible for ammonia homeostasis, the ornithine may generate putrescine and other polyamines essential for multiple cellular functions[59]. Therefore, the amino acid-rich diet and lack of a urea cycle almost certainly imply chronic hyperammonemia in *Osedax*. This would favour amino acid biosynthesis by both *Osedax* and their endosymbionts; however, further functional experiments are needed to test this scenario.

## Lineage-specific expansions of matrix metalloproteinases

As a core component of vertebrate bones, collagen is poised to be an essential nutrient for *Osedax*[28,29,32] and the bone-associated microbiome[48]. Accordingly, transcriptomic analyses uncovered numerous metalloproteases expressed in the root tissue of *O. japonicus*[33]. Our gene family evolutionary analyses also showed that genes involved in collagen catabolism and extracellular matrix organisation are expanded in the genome of *O. frankpressi* (Fig. 2h; Supplementary Fig. 5f). Amongst these expanded families, genes annotated as matrix metalloproteases (MMPs) are the greatest fraction (24.3%). To investigate how MMPs diversified in *O. frankpressi*, we extracted the reconstructed gene families and functional annotations of symbiotic and asymbiotic annelids to identify sequences containing a metallopeptidase domain (InterPro accession IPR006026). We then reconstructed a phylogeny of the metallopeptidase genes using maximum likelihood and Bayesian approaches (Fig. 5a; Supplementary Fig. 8, 9). Our analyses recovered all previously described classes of vertebrate MMPs with high statistical support (bootstrap node support >80%) (Fig. 5a, highlighted in green) and discovered eight new highly supported invertebrate-specific classes of MMPs, labelled A to H (Fig. 5a, highlighted in blue). In addition, we identified two *Osedax*-specific large clades of MMPs, which we referred to as MMP-Os1 and MMP-Os2 (Fig. 5a, highlighted in red). The *Osedax*-specific expansions are more closely related to invertebrate than to vertebrate collagenases, supporting previous enzymatic observations that suggested generic proteolysis rather than an actual collagenase activity in *Osedax* worms[28]. The majority of MMPs belonging to MMP-Os1 (37.5%) had a metallopeptidase domain combined with a C-terminal hemopexin-like repeats domain (IPR018487) thought to facilitate binding to other components of the extracellular matrix[60] (Fig. 5b; Supplementary Fig. 10). As observed with the 12 MMPs reported in *O. japonicus*[33], all but two of the 63 MMPs found in *O. frankpressi* are more highly expressed in root tissue than in the rest of the body (Fig. 5c). At least 43 out of 63 (68.25%) have a signal peptide. This suggests the MMPs are excreted across the root-bone interface—similar to bone-degrading osteoclast cells of vertebrate animals[61]—allowing *Osedax* to digest bone-derived collagen extracellularly and absorb the resulting nutrients through the root epithelium for direct consumption, transport to the endosymbiont for further catabolism[18,32,33], or both. Therefore, the large expansion of MMPs in an otherwise reduced genome is a unique trait of *Osedax* that may be related to their ability to exploit bones from

diverse vertebrates, hence collagens with different amino acid sequences and protease-cleavage sites.

## Divergence in innate immunity repertoire

Establishing stable and specific host-bacterial associations involves innate immunity genes, which are expanded in some Vestimentifera[8,9] (Supplementary Fig. 5e) and other symbiotic oligochaetes[22]. To identify the immune gene repertoire in *O. frankpressi*, we investigated the reconstructed gene families for innate immune pattern recognition receptors corresponding to six major classes, namely lectins, peptidoglycan recognition proteins, Toll-like receptors, scavenger receptors, bactericidal permeability-increasing proteins, and NOD-like receptors[62]. Compared to asymbiotic annelids (i.e., *Owenia fusiformis* and *C. teleta*) and Vestimentifera, *O. frankpressi* has fewer immunity genes in all considered classes (Fig. 6; Supplementary Table 7; Supplementary Data 11–17). This includes a smaller repertoire of Toll-like receptors, which are expanded in some species of Vestimentifera[8,9], and the loss of galectin and a NOD-like receptor, which is a family of cytosolic immune receptors that recognises and triggers inflammatory responses to bacterial pathogens[63] that are also largely expanded in Vestimentifera[9] (Supplementary Table 7; Supplementary Data 12–15). Notably, there is no clear association between the expression levels of the different classes of pattern recognition receptors and the body regions and tissues of Siboglinidae. Yet, a C-type lectin is highly expressed in the root tissue of *O. frankpressi* (Fig. 6). Our findings indicate that *O. frankpressi* and Vestimentifera have different innate immune complements that are simplified in the former and generally expanded in the latter. Further research in Frenulata and *Sclerolinum* will inform whether this divergence in the repertoire of innate immune genes may underpin the evolution of a novel symbiotic association with *Oceanospirillales* bacteria in *Osedax* worms.

## A conserved developmental toolkit in Siboglinidae

In addition to lacking a gut, Vestimentifera and *Osedax* also lack eyes and any other sensory structure in their most anterior region, the prostomium[64]. Yet unlike other annelids with unusual body plans, such as the leech *H. robusta*[38], the genomes of Vestimentifera contain a complete developmental toolkit[9,10]. To investigate genes involved in body patterning and organogenesis in the reduced gene set of *O. frankpressi*, we first focused on the repertoire of G protein-coupled receptors (GPCRs; Supplementary Data 18), a large family of evolutionarily related membrane receptors involved in an array of developmental, sensory, and hormonal processes[65,66]. All siboglinids show a conserved repertoire of GPCRs of class B (secretins), C (metabotropic glutamate receptors) and F (frizzled and smoothened receptors) (Supplementary Fig. 11b–d). However, Siboglinidae has a more divergent complement of rhodopsin-like receptors (class A), with five expanded clusters, one specific to *O. frankpressi* (Supplementary Fig. 11a, highlighted in pink). Notably, *O. frankpressi* and Vestimentifera have lost four GPCR families, including opsins (Supplementary Fig. 11a, highlighted in grey), suggesting an ancestral loss of light perception to these groups in parallel to the colonisation of light-deprived deep marine environments.

The bulk of the body of Siboglinidae has only two segments and the posterior end (i.e., the opisthosoma), which is often multisegmented, is lacking in *Osedax*[31,64]. Nevertheless, the complement of *Hox* genes—a conserved family of transcription factors that define a molecular code throughout the many trunk segments in Annelida[37,67]—is largely conserved in Vestimentifera, only missing the gene *Antennapedia (Antp)*[9,10]. *Osedax frankpressi* has a similar *Hox* gene repertoire, and thus the loss of *Antp* might have occurred in the last common ancestor of Siboglinidae (Supplementary Figs. 12a, 13a). Indeed, the number and complement of transcription factors involved in animal development are comparable in *O. frankpressi*, Vestimentifera and asymbiotic annelids, except for Basic Leucine Zipper Domain
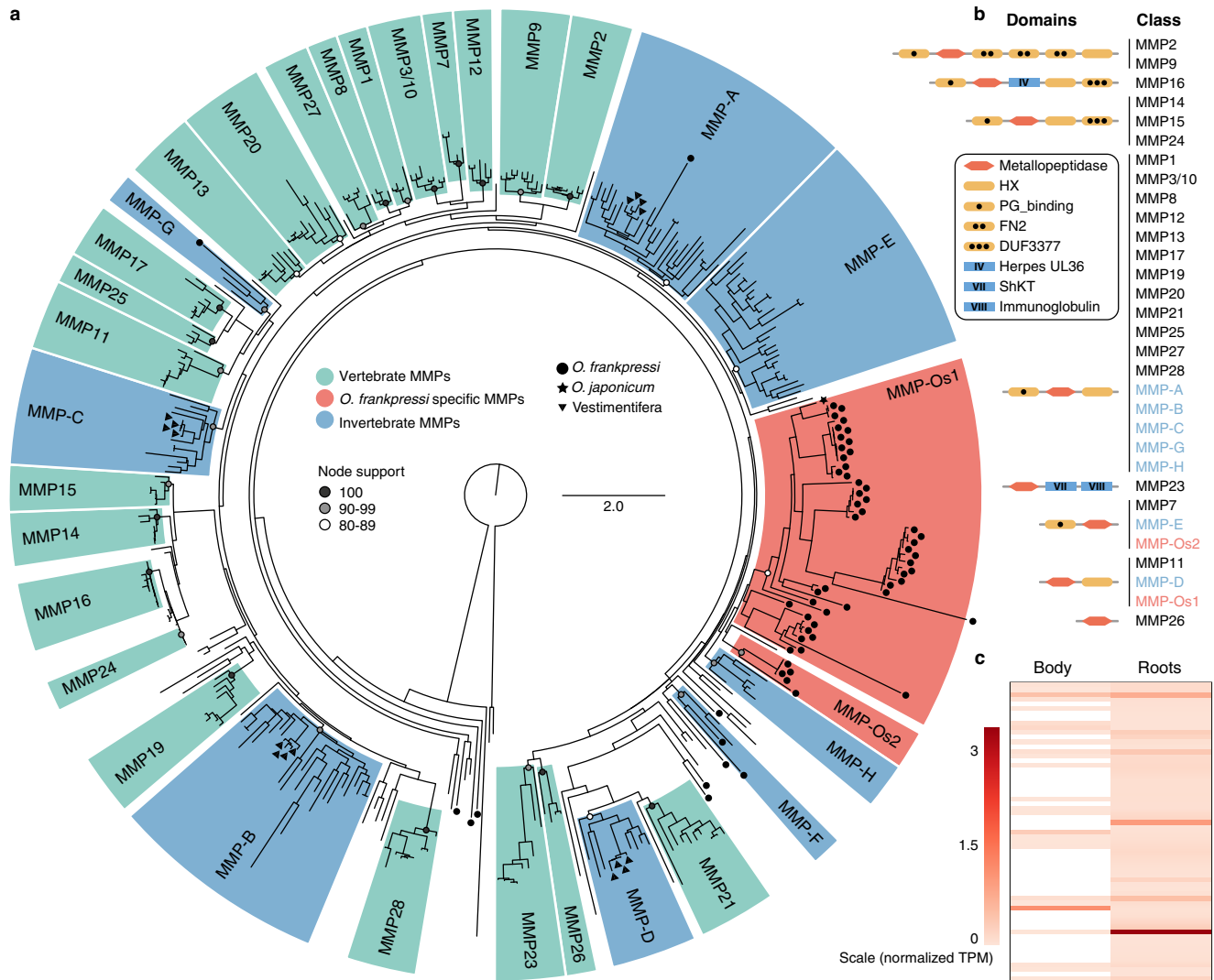
**Fig. 5 | Matrix metalloproteases experienced lineage-specific expansions in *O. frankpressi*. a** Phylogenetic reconstruction of animal matrix metalloproteases (MMPs) based on the metallopeptidase domain. Tree topology is based on maximum likelihood reconstruction and node bootstrap support for each major class is colour coded (white circles show an 80–89 bootstrap support; grey circles indicate a 90–99 bootstrap values and black dots highlight fully supported nodes). Vertebrate-specific MMP classes are highlighted in green and named according to existing literature[161]. New monophyletic clades of invertebrate MMPs are in blue and named from A to H. *Osedax frankpressi* experienced two independent expansions of MMPs, shown in red and named as MMP-Os1 and MMP-Os2. **b** Schematic drawings of the protein domain composition of the different MMP classes recovered in (**a**). For each class, only the most abundant domain architecture is shown. A complete characterisation of the domain composition of MMPs is in Supplementary Fig. 10. Drawings are not to scale. **c** Heatmap of normalised expression levels of MMPs in the body and roots of *O. frankpressi*. Most MMPs show higher expression levels in the roots than in the body. Source data for (**c**) are provided as a Source Data file.

containing proteins (bZIP; PF00170) and zinc finger transcription factors (C2H2-Zn; PF00096), which are reduced (Supplementary Fig. 13b; Supplementary Data 19), as well as certain specific classes, such as the *ParaHox* genes (Supplementary Fig. 12a). Similarly, *O. frankpressi* retains all major developmental signalling pathways, yet it has a lower number of Notch containing proteins (Supplementary Fig. 13c, d) and a simplified repertoire of signalling ligands (Supplementary Figs. 12b, 14, 15), as also observed in the miniaturised annelid *D. gyrociliatus*[36]. Therefore, *O. frankpressi* and Vestimentifera show a similar and generally conserved developmental toolkit, suggesting that changes in gene regulation rather than deviations in the gene complement underpin the development of the divergent adult morphology of Siboglinidae after symbiont acquisition.

## Species-specific repertoires of DNA damage repair mechanisms
Changes in the machinery that repair DNA damage can cause biases in the GC composition of the genome[68,69], and such changes have been

associated with genome compaction and gene loss in animals[70]. *Osedax frankpressi* has an AT-rich genome (29.08% GC content versus ~41% observed in Vestimentifera; Supplementary Fig. 2k, m) and unlike other annelids, it has three major DNA repair pathways that are largely incomplete, namely the base excision repair, the non-homologous end joining, and the Fanconi anaemia DNA repair pathway (Supplementary Figs. 12c, 16). The base excision repair pathway corrects DNA damage from base lesions caused by deamination, oxidation and methylation, and is thought to increase GC to AT base transitions when impaired[71]. The lack of the non-homologous end joining pathway—the most common mechanism to repair double-strand DNA breaks[72]—triggers the error-prone microhomology-mediated end joining pathway, which is intact in *O. frankpressi* and all other annelids but causes microdeletions[73] (Supplementary Fig. 16f; Supplementary Data 20). Therefore, the loss of genes involved in the repair of double-strand DNA breaks and chemical base modifications might underpin the reduction in genome size and GC content observed in *O. frankpressi* in
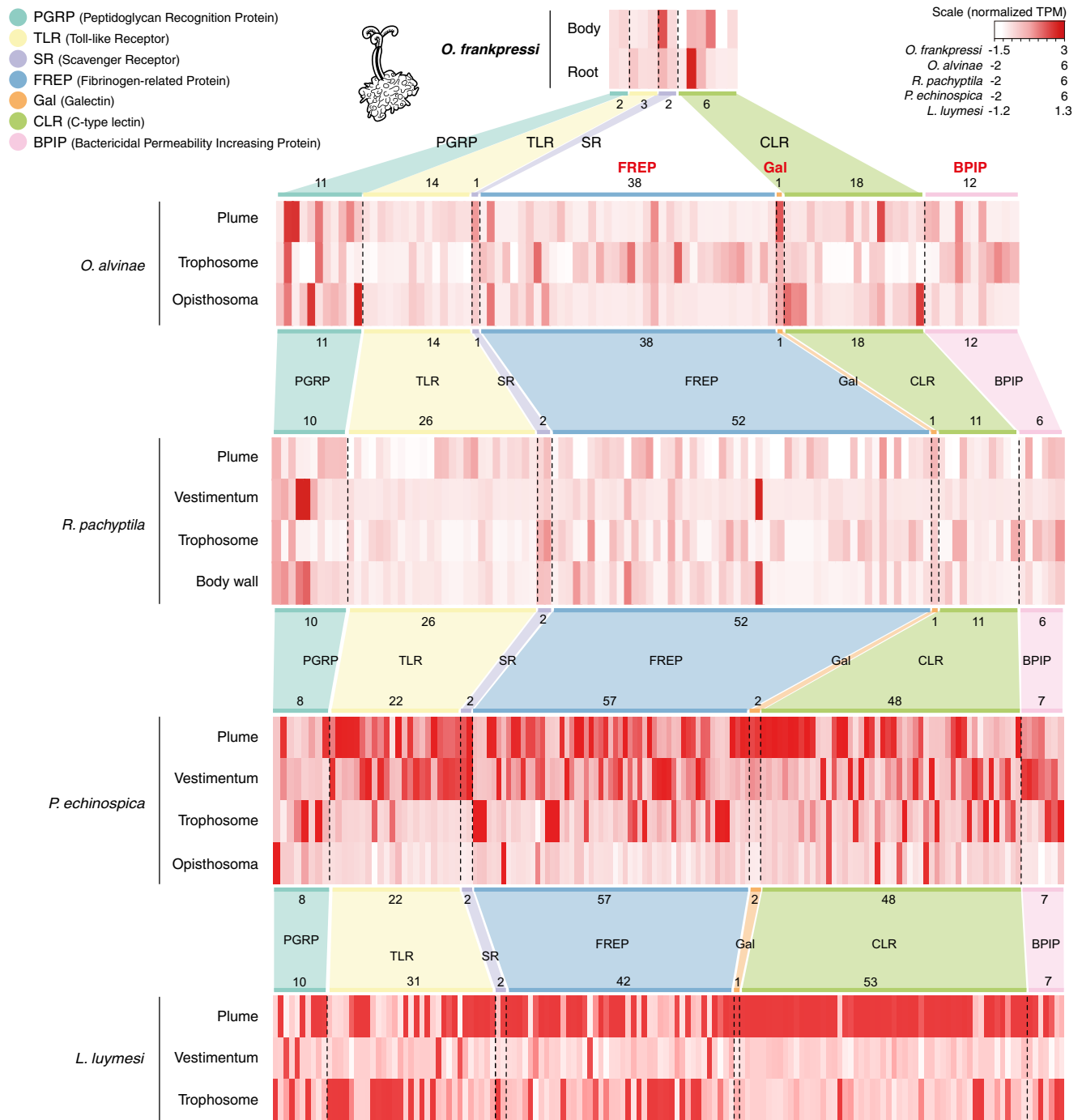
**Fig. 6 | _O. frankpressi_ has a reduced innate immune gene repertoire compared to Vestimentifera.** Heatmaps of tissue-specific normalised gene expression of innate immune genes in four species of Siboglinidae, including _O. frankpressi_ (top) and the Vestimentifera _Oasisia alvinae_, _R. pachyptila_ and _P. echinospica_. While Vestimentifera have relatively similar repertoires of innate immune genes, _O. frankpressi_ has a much-reduced complement (Supplementary Table 7). Notably,

innate immune genes do not show a clear tissue-specific expression within or among species of Siboglinidae. The immune repertoire and gene expression values for _P. echinospica_ and _L. luymesi_ is based on previously published genome resources[8, 9]. Source data for _O. frankpressi_, _Oasisia alvinae_ and _R. pachyptila_ are provided as a Source Data file.

comparison with Vestimentifera, thus differing from other annelids with reduced genomes, such as _D. gyrociliatus_, whose genome eroded without changes in DNA repair pathways[36].

## Discussion

Our data reveal additional evidence on the genetic interactions and co-dependencies of animal hosts and bacterial symbionts that have enabled distinct symbiotic lifestyles, including the exploitation of

sunken vertebrate bones as a food source (Fig. 7a). Our analyses of the genomes of _Oasisia alvinae_ and _R. pachyptila_ confirm what was previously reported for other species of Vestimentifera[8,9] and _R. pachyptila_ itself[10] and support that broadly similar genomic adaptations underpin the different symbioses of Vestimentifera, even between species occupying distinct environments, such as hydrothermal vents and methane seeps. However, compared to Vestimentifera, _O. frankpressi_ shows a fast evolving[36], divergent gene
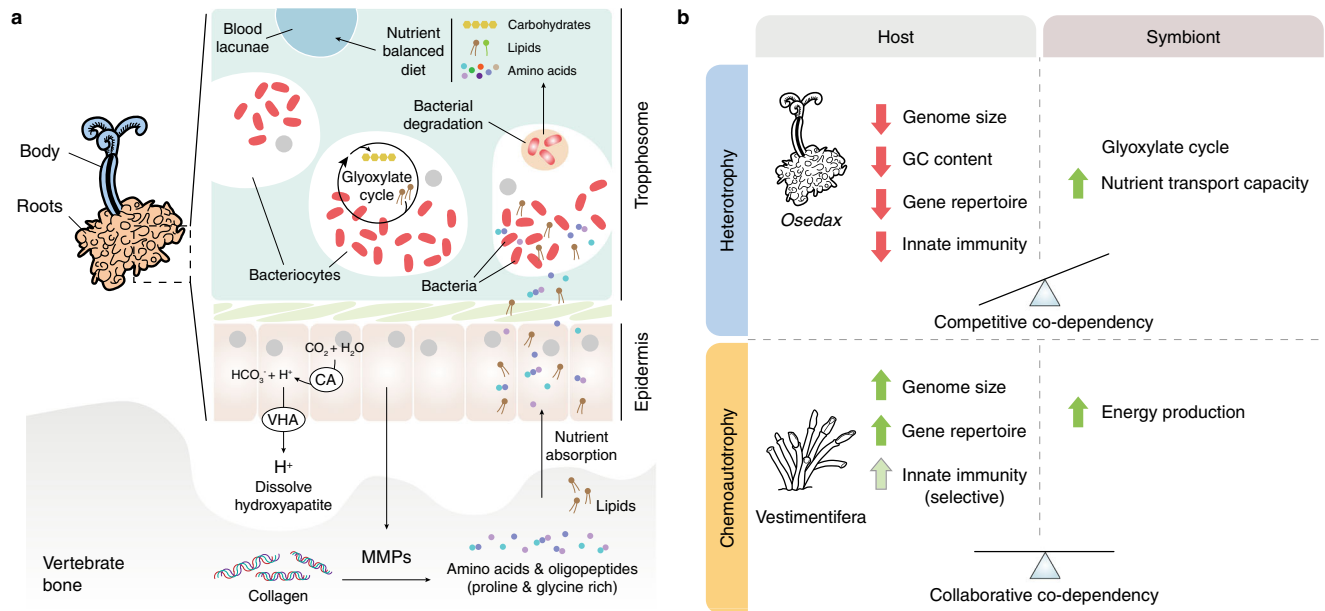
**Fig. 7 | The genomic basis and evolution of different nutritional symbioses in Siboglinidae. a** Schematic drawing of the metabolic interaction for bone digestion between *Osedax* and its endosymbiont (red kidney-shaped ovals), which are harboured in the trophosome inside bacteriocytes. The root epidermis secretes acid to dissolve the inorganic component of the bone (via carbonic anhydrase, CA, and V-type H+-ATPase, VHA) and matrix metalloproteases (MMPs) that break collagen, one of the most abundant organic components of the bone, into amino acids and oligopeptides, which are rich in proline and glycine. These amino acids and the lipidic content of the bone are absorbed by the epidermis and used either directly by *Osedax* or transported to bacteriocytes. The host and endosymbiont cooperate to generate carbohydrates (in low amounts in bone) from the oxidation of fatty acids (abundant in bone and roots) through the glyoxylate cycle, most likely inside the bacteriocytes. Ultimately, these interactions transform the originally unbalanced diet into complex and diverse macronutrients, which are then taken directly or after the digestion of the bacteria by the host. **b** *Osedax* and Vestimentifera broadly show different genomic traits. *Osedax* has a small, AT-rich genome, with many gene losses and a reduced immune repertoire. Vestimentifera tends to show larger genomes, with a more extensive gene complement and richer innate immunity (although there is inter-specific variation for these traits, highlighted with a lighter green arrow). We hypothesise that the different nutritional relationships between the hosts and symbionts in these two groups might explain, at least partially, these genomic differences. *Osedax* and its endosymbiont co-depend on and compete to exploit the finite, nutritionally unbalanced diet obtained from bones, which might have favoured the evolution of an energetically "cheaper" genome in *Osedax*. In Vestimentifera, however, the endosymbiont acts as a primary producer, which might be able to sustain larger host genomes. Drawings are not to scale.

repertoire, with gene losses and expansions in key functional groups that support metabolic adaptations to its symbiotic lifestyle (Figs. 2, 3b, 4a; Supplementary Fig. 5d, f). As observed in the marine microbial assemblages on bone surfaces[48], the expansion of secreted matrix metalloproteases[33] (Fig. 5a) combined with the active secretion of acid in the root tissue[32] are the most probable mechanisms of bone digestion by the host (Fig. 7a). The *Osedax*-microbe association, however, entails further molecular and metabolic interactions to overcome a nutritionally unbalanced diet that is deficient in carbohydrates but enriched in (hydroxy)proline- and glycine-rich proteins and, in some cases, lipids[49,50]. Most notably, our findings suggest that the *Oceanospirillales* endosymbionts might be able to provide *Osedax* with glucose through the glyoxylate cycle (Fig. 3b) and that *Osedax* and the endosymbionts cooperate to maintain a physiological status of hyperammonemia (Fig. 4d). The former allows the catabolism of fatty acids to produce carbohydrates, which the host could take up by digesting the endosymbionts and store as glycogen (e.g., as seen in *Osedax*'s oocytes[74]), whereas the latter could stimulate the biosynthesis of amino acids, ultimately counterbalancing the lack of carbohydrates and skewed amino acid composition in bone. Notably, the use and occurrence of the glyoxylate cycle in animals is controversial[75,76] and only reported in a handful of taxa[77,78], likely as a consequence of horizontal gene transfer[79] and often concerning stress and a metabolic diapause, such as in the Dauer larva of nematodes[51], hibernating mammals[80] and bleached coral[81]. Indeed, *Osedax*, like Vestimentiferan hosts and their endosymbionts, lacks isocitrate lyase, but this enzyme is present in *Osedax*'s endosymbiont[18] (Fig. 3b; Supplementary Data 3, 7). Therefore, the

metabolic diversity of the *Oceanospirillales* endosymbiont may be critical to maximising the use of the imbalanced resources derived from the bones and ultimately acts as a selective pressure to acquire and maintain this microbe as the primary symbiont.

Symbiotic interactions can impose selective pressures that direct genome evolution—most notably in symbionts[82] but also occasionally in hosts[83]—triggering changes in genome size (e.g., genome erosion)[84], gene content[85] and even DNA base composition in favour of AT-rich genomes[86]. Most of these changes, however, are known for strictly vertically transmitted obligate endosymbionts of insects. Our study shows that Vestimentifera and *Osedax*, two annelid lineages within Siboglinidae that establish environmentally acquired symbioses, show differences in genome structure and composition (Fig. 2a–c; Supplementary Fig. 2m). While Vestimentifera tends to have larger genome sizes, similar GC content to asymbiotic annelids[37,38], and larger gene repertoires, *O. frankpressi* has a small, AT-rich genome, with a reduced gene content (Fig. 7b). In addition, these Siboglinidae crucially differ in their nutritional symbioses—chemoautotrophic in Vestimentifera and heterotrophic in *Osedax*—which enable them as adults to thrive in different ecological niches with different nutritional pressures. In hydrothermal vents and methane seeps, Vestimentifera relies on virtually unlimited inorganic nutrients that are exploited by the endosymbionts, which in their role as primary producers sustain long-lasting collaborative co-dependencies with their hosts[3,5]. Decaying bones are, however, nutritionally finite, and thus *Osedax* and their endosymbionts may establish a competitive co-dependency to exploit those nutritionally unbalanced resources (Fig. 7b). Moreover, the potential use of the glyoxylate cycle for energy production would be

less energetically efficient than the sole catabolism of fatty acids[87]. Therefore, we hypothesise that the interaction between *Osedax* and its endosymbiont might, in turn, favour the genomic streamlining of the annelid host (Fig. 7b) so that it becomes metabolically and energetically "cheaper" and can sustain larger endosymbiotic populations for longer periods. Our findings thus suggest that incipient genome erosion can occur in hosts with horizontally acquired symbionts and that adaptive genome evolution may differ based on the type of nutritional interactions between the host and symbiont. In the future, dissecting the metabolic co-dependencies between Siboglinidae and their endosymbionts, including the Frenulata and *Sclerolinum*–the other two major lineages within Sibogliniade–will help to disentangle the role of neutral and adaptive selective pressures in the evolution of these fascinating, but still poorly understood, animal symbioses.

# Methods

## Specimen collections, gDNA extraction and sequencing
Live adult specimens of *O. frankpressi*, *Oasisia alvinae* and *R. pachyptila* were obtained with deep-sea specialised robots off the coasts of California and Mexico (Supplementary Fig. 1c, d). Mexican samples were collected under CONAPESCA permit PPFE/DGOPA-200/18. Ultra-high molecular weight genomic DNA (gDNA) was extracted following the Bionano Genomics IrysPrep agar-based, animal tissue protocol (Catalogue # 80002) from an entire *O. frankpressi* adult female, a piece of the trunk (including trophosome) of *Oasisia alvinae*, and a piece of the vestimentum of *R. pachyptila*. Long-read PacBio sequencing and short-read Illumina sequencing was performed at the Genome Centre of the University of California Berkeley in a PacBio Sequel II and Illumina Novaseq platforms (Supplementary Table 1).

## Transcriptome sequencing
Total RNA from dissected tissues and body parts of *Oasisia alvinae* (crown, opisthosome and trophosome), and *R. pachyptila* (crown and trunk wall) was extracted with an NEB totalRNA Monarch kit and used for standard strand-specific RNA Illumina library prep. Libraries were sequenced to a depth of 40–50 million paired reads of 150 bases length in a NovaSeq platform (Supplementary Table 1). Publicly available datasets for *O. frankpressi* (NCBI short read archive accession numbers SRR2017399 and SRR2017400) were used in this study (Supplementary Table 1).

## Host genome assembly and quality check
PacBio reads were used to generate an initial genome assembly with Canu v.1.8[88] with options 'batOptions = "-dg 3 -db 3 -dr 1 -ca 500 -cp 50'. Two rounds of polishing using PacBio reads were performed using Pbmm2 v.1.1.0 (https://github.com/PacificBiosciences/pbmm2) and Arrow (pbgcpp v.1.9.0)[89]. Short genomic Illumina reads were quality filtered with FastQC v.0.11.8 and Cutadapt v.2.5[90], mapped to the polished assembly with BWA v.0.7.17[91] and used for final polishing with Pilon v.1.23[92]. The polished versions of the genomes of *O. frankpressi*, *Oasisia alvinae* and *R. pachyptila* were used as input to BlobTools v.2.1[93] to identify and remove contigs with high similarity to bacteria. After decontamination, the haplotypes were purged with Purge_Dups v.1.0.1[94]. Quality check was performed with BUSCO v.3.0.2[95], to estimate gene completeness of the assembly (Supplementary Table 3), QUAST v.5.0.2[96], and KAT v.2.4.2[97] to assess haplotype removal (Supplementary Fig. 2b–d) and potential bacterial remnants.

## Genome size estimations
Short Illumina reads were mapped to the reference host genome assembly with BWA v.0.7.17 and KAT v.2.4.2[97] to count and generate a histogram of canonical 21-mers. GenomeScope2[98] was used to estimate the genome size and heterozygosity (Supplementary Fig. 2e–g).

## Symbiont genome assembly and annotation
For *O. frankpressi* and *Oasisia alvinae*, we used Kraken2 v.2.1.0[99] and Krakentools v.0.1[99] to isolate long PacBio reads of bacterial origin. After error correction with Canu v.1.8[88], these PacBio reads were assembled using Metaflye v.2.9[100] followed by ten polishing iterations with options "–pacbio-corr –meta –keep-haplotypes –iterations 10" and final polishing with NextPolish v.1.4.0[101]. The resulting assemblies were manually inspected using Bandage v.0.9.0[102], binned with Max-Bin2 v.2.2.7[103] and quality checked with CheckM v.1.0.8[104] and Meta-Quast v.5.2.0[105]. Gene annotation was performed with Prokka v.1.14.5[106] with the "–compliant" option and proteins involved in secretion systems were identified by scanning for unordered replicons using the curated HMM profiles of TXSscan in MacSyFinder v.2[107]. The bacterial genomes were checked for secreted proteins with eukaryotic-like domains using EffectiveELD through EffectiveDB, on default settings[108]. All coding sequences of the main endosymbiont ribotype for *O. frankpressi*, Vestimentifera and Frenulata were assigned KO numbers using BlastKOALA v.2.2[109], which were used as input for KEGG Mapper v.5[110] to analyse the metabolic capabilities of each symbiont. The NCBI COG database[111] was used to tag functional categories to the annotated genes. Enrichment analyses of functional categories and Gene Ontology terms were performed with GSEA v.4.2.3[112] and OrthoVenn2 v.2[113]. To compute the *p*-values for enriched Gene Ontology terms in a protein cluster (Supplementary Data 6, 7), a hypergeometric distribution was used to identify significantly enriched terms within each cluster of orthologous/paralogous genes. GTDB-Tk v.1.6.0[114] was used for whole genome phylogenetic placement and identification of neighbouring available genomes isolated from free-living deep-sea bacteria. Circos v.0.69-9[115] was used for genome assembly visualisation.

## Annotation of repeats in host genomes
RepeatModeler v.2.0.1[116] and Repbase[117] were used to build a de novo library of repeats for the host genome of *O. frankpressi*, *Oasisia alvinae* and *R. pachyptila*. The predicted genes of *Owenia fusiformis*[37] and DIAMOND v.0.8.22[118] were used to filter out bona fide genes in the predicted repeats with an e-value threshold of 1e-10. Subsequently, RepeatMasker v.4.1.0[119] (Supplementary Tables 4–6) and LTR-finder v.1.07[120] were used to identify and annotate repeats, and RepeatCraft[121] to generate a consensus annotation that was used to soft-mask the genome assemblies of the three annelid species. To explore the transposable element landscape, we used the online tool TEclass[122] to annotate the TEs identified by RepeatModeler and the scripts "calc-DivergenceFromAlign.pl" and a custom-modified version of "create-RepeatLandscape.pl", both from RepeatMasker v.4.1.0, to estimate Kimura substitution levels, which were plotted using ggplot2 v.3.3.0[123]. Previously published TE landscapes were included for comparisons[37].

## Functional annotation of host genomes
Individual RNA-seq Illumina libraries (Supplementary Table 1) were de novo assembled with Trinity v.2.9.1[124] after quality trimming with Trimmomatic v.0.35[125]. GMAP v.2017.09.30[126] and STAR v.2.7.5a[127] were used to map transcripts and quality-filtered Illumina reads to the soft-masked genome assemblies of the corresponding species. For *R. pachyptila*, publicly available datasets (SRA accession numbers SRR8949056 to SRR8949077) were also mapped to the soft-masked genome assembly. In addition, gene transfer format (GTF) files from the mapped reads and curated intron junctions were inferred with StringTie v.2.1.2[128] and Portcullis v.1.2.2[129]. All RNA-seq-based gene evidence was merged with Mikado v.2.Orc2[130], which produced a curated transcriptome-based genome annotation. Full-length Mikado transcripts were used to train Augustus v.3.3.3[131], which was then used to generate ab initio gene predictions that incorporate the intron hints of Portcullis and the exon hints of Mikado. In addition, Exonerate v.2.4.0[132] was used to produce spliced alignments of the curated

proteomes of *Owenia fusiformis*, *C. teleta* and *L. luymesi* that were used as further exon hints for Augustus. Finally, the Mikado RNA-seq-based gene evidence and the ab initio predicted Augustus gene models were merged with PASA v.2.4.1[133]. A final, curated gene set was obtained after removing spurious gene models and genes with high similarity to transposable elements. Gene completeness and annotation quality were assessed with BUSCO v.3.0.2[95]. Trinotate v.3.2.1[134], PANTHER v.1.0.10[47] and the online tool KAAS[135] were used to functionally annotate the curated gene sets.

### Comparison of *R. pachyptila* assemblies
Overall genomic stats were obtained with BUSCO v.3.0.2[95], QUAST v.5.0.2[96] and AGAT v.0.5.0[136]. We used minimap2 v.2.17 to align our *R. pachyptila* assembly with the assembly previously reported[10] and the R package pafr to generate a dot-plot representation of the sequence similarity between the two versions. In addition, we reassembled all transcriptomic evidence published elsewhere[12] using Trinity v.2.9.1[124] and cd-hit v.4.8.1[137]. To identify one-to-one orthologs between genomic and transcriptomic resources, we used a reciprocal best BLAST hit approach with BLAST v.2.12.0+ [138]. Finally, we used PFAMscan v.1.6[139] to identify and quantify distinct Pfam domains in the different assemblies.

### Gene family evolutionary analyses
The non-redundant proteomes of *O. frankpressi*, *Oasisia alvinae* and *R. pachyptila* together with 25 high-quality genomes spanning major groups of the animal tree (Supplementary Data 2) were used to construct orthogroups with OrthoFinder v.2.5.2[140] using DIAMOND v.2.0.9[118] with "−ultra-sensitive" option. The OrthoFinder output and a published Python script[36] were used to infer gene family evolutionary dynamics at each node and tip of the tree. Gene Ontology term enrichment analyses for expanded and lost gene families were performed with the R package "TopGO" v.2.42.0. The number of orthologs per gene family and species as generated by OrthoFinder was used to perform a Principal Component Analysis with R built-in functions.

### Reconstruction of host metabolic pathways and developmental gene sets
PANTHER and Pfam annotations obtained through PANTHER v.1.0.10[47] and Trinotate v.3.2.1[134], respectively, were used to assess for the presence of each enzyme involved in the synthesis of amino acids, vitamin Bs, nitrogen metabolism, glycine degradation, matrix metalloproteases, transcription factors and DNA repair pathways in an array of annelid species. A combination of BlastKOALA[109] and KofamKOALA[141] was used to annotate the host and endosymbiont genomes for the analysis of the lipid and carbohydrate metabolism. Information about each step in a pathway was collected from MetaCyc[142], KEGG[143] and PANTHER[47] databases. To analyse the tissue-specific expression of candidate genes in *O. frankpressi*, *Oasisia alvinae* and *R. pachyptila*, quality-filtered short Illumina reads were pseudo-mapped to the filtered gene models of each species with Kallisto v.0.46.2[144] to quantify transcript abundances as Transcripts per Kilobase Million (TPM) values. The R libraries ggplot2 v.3.3.0[123] and pheatmap v.1.0.12 (https://cran.r-project.org/web/packages/pheatmap/index.html) were used to plot expression and abundance heatmaps.

### Reconstruction of innate immune repertoires
The OrthoFinder output was used to identify gene families of innate immune pattern recognition receptors of *O. frankpressi*, Vestimetifera and two asymbiotic annelids, *Owenia fusiformis* and *C. teleta*, with the published pattern recognition receptors of Vestimetifera[9] as baits (Supplementary Data 11–17). PANTHER and Pfam annotations (see above) of the target proteins were further used to remove sequences that were too short or lacked target domains. TPM expression values

(see above) and TBtools v.1.042[145] were used to plot gene expression heatmaps.

### Reconstruction of the G protein-coupled receptor (GPCR) repertoire
Transcriptomes of the focal species were downloaded and processed as described elsewhere[146]. Multiple sequence alignments of rhodopsin type GPCRs (PF00001), secretin type GPCRs (PF00002), glutamate type GPCRs (PF00003) and frizzled type GPCRs (PF01534) were downloaded from the Pfam webpage (https://pfam.xfam.org) and used to create HMM profiles using hmmer-3.1b2[147]. HMMer search was performed with an e-value cut-off of 1e-10. The online version of CLANS (https://toolkit.tuebingen.mpg.de/tools/clans) was used for the initial BLAST comparison for the cluster analysis and edges below 1e-10 for secretin, glutamate and frizzled type GPCRs and 1e-20 for rhodopsin type GPCRs were removed. The java offline version of CLANS[148] was then used for the cluster analysis. The *p*-value for clustering was set to 1e-25. Singletons and group-specific sequence clusters with less than five sequences and no annotation (using Linkage clustering for identification) were removed. The highly vertebrate-specific expanded olfactory GPCR type-A receptors were also deleted as these showed no connections and strongly repulsed all other sequences. Gene clusters were annotated according to the presence of characterized sequences of *Drosophila melanogaster*, *Homo sapiens*, *Danio rerio* and *Platynereis dumerilii*.

### Orthology assignments
MAFFT[149] with default options was used to align candidate sequences to a curated set of proteins that we obtained either from previous studies[36,150] or manually from UniProt[151]. Conserved protein domains were retained by trimming by hand the alignment in Jalview[152] and the resulting sequences were re-aligned in MAFFT with the "L-INS-I" algorithm[149]. After a final trim to further remove spurious regions with trimAI v.1.4.rev15[153], FastTree v.2.1.10[154] with default options and IQ-Tree v.2.2.0-beta[155] (for matrix metalloproteases) using the options "-m MFP -B 1000", were used to infer orthology relationships. In addition, for the matrix metalloproteases, the posterior probabilities were obtained from Bayesian reconstructions in MrBayes v.3.2.7a[156], which were performed using as a prior the LG matrix[157] with a gamma model[158] with four categories to describe sites' evolution rate. Four runs with eight chains were run for 20,000,000 generations. FigTree v.1.4.4 (https://github.com/rambaut/figtree) and Adobe Illustrator were used to edit the final trees. CD-Search[159] with default options and the Conserved Domain Database (CDD)[160] were used to annotate protein domains in the predicted matrix metalloproteases.

### Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

## References
1. Archibald, J. *One Plus One Equals One*. (Royal Society of Biology, 2013).
2. Sagan, L. On the origin of mitosing cells. *J. Theor. Biol.* **14**, 225–274 (1967).

3.  Dubilier, N., Bergin, C. & Lott, C. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat. Rev. Microbiol.* **6**, 725–740 (2008).

4.  Grassle, J. F. in *Advances in Marine Biology* Vol. 23 (eds J. H. S. Blaxter & A. J. Southward) 301–362 (Academic Press, 1987).

5.  Hilario, A. et al. New perspectives on the ecology and evolution of siboglinid tubeworms. *PLoS ONE* **6**, e16309 (2011).

6.  Petersen, J. M. & Dubilier, N. Methanotrophic symbioses in marine invertebrates. *Environ. Microbiol. Rep.* **1**, 319–335 (2009).

7.  Nussbaumer, A. D., Fisher, C. R. & Bright, M. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**, 345–348 (2006).

8.  Li, Y. et al. Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm *Lamellibrachia luymesi. BMC Biol.* **17**, 91 (2019).

9.  Sun, Y. et al. Genomic signatures supporting the symbiosis and formation of chitinous tube in the deep-sea tubeworm *Paraescarpia echinospica. Mol. Biol. Evol.* **38**, 4116–4134 (2021).

10. de Oliveira, A. L., Mitchell, J., Girguis, P. & Bright, M. Novel insights on obligate symbiont lifestyle and adaptation to chemosynthetic environment as revealed by the giant tubeworm genome. *Mol. Biol. Evol.* **39**, msab347 (2022).

11. Wang, M. et al. The genome of a vestimentiferan tubeworm (*Ridgeia piscesae*) provides insights into its adaptation to a deep-sea environment. *BMC Genomics* **24**, 72 (2023).

12. Hinzke, T. et al. Host-microbe interactions in the chemosynthetic *Riftia pachyptila* symbiosis. *mBio* **10**, e02243–19 (2019).

13. Bailly, X. et al. Evolution of the sulfide-binding function within the globin multigenic family of the deep-sea hydrothermal vent tubeworm *Riftia pachyptila. Mol. Biol. Evol.* **19**, 1421–1433 (2002).

14. Zal, F., Lallier, F. H., Wall, J. S., Vinogradov, S. N. & Toulmond, A. The multi-hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila*. I. Reexamination of the number and masses of its constituents. *J. Biol. Chem.* **271**, 8869–8874 (1996).

15. Nyholm, S. V., Song, P., Dang, J., Bunce, C. & Girguis, P. R. Expression and putative function of innate immunity genes under in situ conditions in the symbiotic hydrothermal vent tubeworm *Ridgeia piscesae. PLoS One* **7**, e38267 (2012).

16. Li, Y., Liles, M. R. & Halanych, K. M. Endosymbiont genomes yield clues of tubeworm success. *ISME J.* **12**, 2785–2795 (2018).

17. Yang, Y. et al. Genomic, transcriptomic, and proteomic insights into the symbiosis of deep-sea tubeworm holobionts. *ISME J.* **14**, 135–150 (2020).

18. Goffredi, S. K. et al. Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea *Osedax* worms. *ISME J.* **8**, 908–924 (2014).

19. Robidart, J. C. et al. Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. *Environ. Microbiol.* **10**, 727–737 (2008).

20. Sun, J. et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* **1**, 121 (2017).

21. Lan, Y. et al. Hologenome analysis reveals dual symbiosis in the deep-sea hydrothermal vent snail *Gigantopelta aegis. Nat. Commun.* **12**, 1165 (2021).

22. Wippler, J. et al. Transcriptomic and proteomic insights into innate immunity and adaptations to a symbiotic lifestyle in the gutless marine worm *Olavius algarvensis. BMC Genomics* **17**, 942 (2016).

23. Rouse, G. W., Goffredi, S. K., Johnson, S. B. & Vrijenhoek, R. C. Not whale-fall specialists, *Osedax* worms also consume fishbones. *Biol. Lett.* **7**, 736–739 (2011).

24. Rouse, G. W., Goffredi, S. K. & Vrijenhoek, R. C. *Osedax*: bone-eating marine worms with dwarf males. *Science* **305**, 668–671 (2004).

25. Vrijenhoek, R. C., Johnson, S. B. & Rouse, G. W. A remarkable diversity of bone-eating worms (*Osedax*; Siboglinidae; Annelida). *BMC Biol.* **7**, 74 (2009).

26. Goffredi, S. K., Paull, C. K., Fulton-Bennett, K., Hurtado, L. A. & Vrijenhoek, R. C. Unusual benthic fauna associated with a whale fall in Monterey Canyon, California. *Deep Sea Res. Part I: Oceanographic Res. Pap.* **51**, 1295–1306 (2004).

27. Shimabukuro, M. & Sumida, P. Y. G. Diversity of bone-eating *Osedax* worms on the deep Atlantic whale falls—bathymetric variation and inter-basin distributions. *Mar. Biodivers.* **49**, 2587–2599 (2019).

28. Goffredi, S. K., Johnson, S. B. & Vrijenhoek, R. C. Genetic diversity and potential function of microbial symbionts associated with newly discovered species of *Osedax* polychaete worms. *Appl. Environ. Microbiol.* **73**, 2314–2323 (2007).

29. Goffredi, S. K. et al. Evolutionary innovation: a bone-eating marine symbiosis. *Environ. Microbiol.* **7**, 1369–1378 (2005).

30. Katz, S., Klepal, W. & Bright, M. The *Osedax* trophosome: organization and ultrastructure. *Biol. Bull.* **220**, 128–139 (2011).

31. Rouse, G. W., Pleijel, F. & Tilic, E. *Annelida* (Oxford University Press, 2022).

32. Tresguerres, M., Katz, S. & Rouse, G. W. How to get into bones: proton pump and carbonic anhydrase in *Osedax* boneworms. *Proc. Biol. Sci.* **280**, 20130625 (2013).

33. Miyamoto, N., Yoshida, M. A., Koga, H. & Fujiwara, Y. Genetic mechanisms of bone digestion and nutrient absorption in the bone-eating worm *Osedax japonicus* inferred from transcriptome and gene expression analyses. *BMC Evol. Biol.* **17**, 17 (2017).

34. Smith, C. R., Glover, A. G., Treude, T., Higgs, N. D. & Amon, D. J. Whale-fall ecosystems: recent insights into ecology, paleoecology, and evolution. *Ann. Rev. Mar. Sci.* **7**, 571–596 (2015).

35. Bonnivard, E., Catrice, O., Ravaux, J., Brown, S. C. & Higuet, D. Survey of genome size in 28 hydrothermal vent species covering 10 families. *Genome* **52**, 524–536 (2009).

36. Martin-Duran, J. M. et al. Conservative route to genome compaction in a miniature annelid. *Nat. Ecol. Evol.* **5**, 231–242 (2021).

37. Martin-Zamora, F. M. et al. Annelid functional genomics reveal the origins of bilaterian life cycles. *Nature* **615**, 105–110 (2023).

38. Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531 (2013).

39. Martin-Duran, J. M., Ryan, J. F., Vellutini, B. C., Pang, K. & Hejnol, A. Increased taxon sampling reveals thousands of hidden orthologs in flatworms. *Genome Res.* **27**, 1263–1272 (2017).

40. Bright, M. & Giere, O. Microbial symbiosis in Annelida. *Symbiosis* **38**, 1–45 (2005).

41. Lund, M. B., Kjeldsen, K. U. & Schramm, A. The earthworm-Verminephrobacter symbiosis: an emerging experimental system to study extracellular symbiosis. *Front. Microbiol.* **5**, 128 (2014).

42. Graf, J., Kikuchi, Y. & Rio, R. V. Leeches and their microbiota: naturally simple symbiosis models. *Trends Microbiol.* **14**, 365–371 (2006).

43. Hewitt, O. H., Díez-Vives, C. & Taboada, S. Microbial insights from Antarctic and Mediterranean shallow-water bone-eating worms. *Polar Biol.* **43**, 1605–1621 (2020).

44. Miyazaki, M. et al. *Neptunomonas japonica* sp. nov., an *Osedax japonicus* symbiont-like bacterium isolated from sediment adjacent to sperm whale carcasses off Kagoshima, Japan. *Int. J. Syst. Evolut. Microbiol.* **58**, 866–871 (2008).

45. Polzin, J., Arevalo, P., Nussbaumer, T., Polz, M. F. & Bright, M. Polyclonal symbiont populations in hydrothermal vent tubeworms and the environment. *Proc. Biol. Sci.* **286**, 20181281 (2019).

46. Reynolds, D. & Thomas, T. Evolution and function of eukaryotic-like proteins from sponge symbionts. *Mol. Ecol.* **25**, 5242–5253 (2016).

47. Thomas, P. D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).

48. Borchert, E. et al. Deciphering a marine bone-degrading micro-biome reveals a complex community effort. *mSystems* **6**, e01218-20 (2021).

49. Ricard-Blum, S. The Collagen Family. *Cold Spring Harb. Perspect. Biol.* **3**, a004978 (2011).

50. Higgs, N. D., Little, C. T. S. & Glover, A. G. Bones as biofuel: a review of whale bone composition with implications for deep-sea biology and palaeoanthropology. *Proc. R. Soc. B: Biol. Sci.* **278**, 9–17 (2011).

51. O'Riordan, V. B. & Burnell, A. M. Intermediary metabolism in the dauer larva of the nematode *Caenorhabditis elegans*—II. The glyoxylate cycle and fatty-acid oxidation. *Comp. Biochem. Physiol. Part B: Comp. Biochem.* **95**, 125–130 (1990).

52. Liu, F., Thatcher, J. D. & Epstein, H. F. Induction of Glyoxylate Cycle Expression in *Caenorhabditis elegans*: A Fasting Response throughout Larval Development. *Biochemistry* **36**, 255–260 (1997).

53. Saz, H. J. The enzymic formation of glyoxylate and succinate form tricarboxylic acids. *Biochem. J.* **58**, xx–xxi (1954).

54. Kornberg, H. L. & Krebs, H. A. Synthesis of cell constituents from C2-units by a modified tricarboxylic acid cycle. *Nature* **179**, 988–991 (1957).

55. Melendez-Hevia, E. & Paz-Lugo, P. D. Branch-point stoichiometry can generate weak links in metabolism: the case of glycine bio-synthesis. *J. Biosci.* **33**, 771–780 (2008).

56. Wright, P. A. Nitrogen excretion: three end products, many phy-siological roles. *J. Exp. Biol.* **198**, 273–281 (1995).

57. Natesan, S., Jayasundaramma, B., Ramamurthi, R. & Reddy, S. R. R. Presence of a partial urea cycle in the leech, *Poecilobdella gran-ulosa*. *Experientia* **48**, 729–731 (1992).

58. Jones, P., Patel, K. & Rakheja, D. in *A Quick Guide to Metabolic Disease Testing Interpretation* (Second Edition) (eds Jones, P., Patel, K. & Rakheja, D.) 75–78 (Academic Press, 2020).

59. Majumdar, R., Shao, L., Minocha, R., Long, S. & Minocha, S. C. Ornithine: the overlooked molecule in the regulation of polyamine metabolism. *Plant Cell Physiol.* **54**, 990–1004 (2013).

60. Das, S., Mandal, M., Chakraborti, T., Mandal, A. & Chakraborti, S. Structure and evolutionary aspects of matrix metalloproteinases: a brief overview. *Mol. Cell Biochem.* **253**, 31–40 (2003).

61. Everts, V. et al. Degradation of collagen in the bone-resorbing compartment underlying the osteoclast involves both cysteine-proteinases and matrix metalloproteinases. *J. Cell Physiol.* **150**, 221–231 (1992).

62. Janeway, C. A. Jr. & Medzhitov, R. Innate immune recognition. *Annu Rev. Immunol.* **20**, 197–216 (2002).

63. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140**, 805–820 (2010).

64. Worsaae, K., Rimskaya-Korsakova, N. N. & Rouse, G. W. Neural reconstruction of bone-eating *Osedax* spp. (Annelida) and evolu-tion of the siboglinid nervous system. *BMC Evolut. Biol.* **16**, 83 (2016).

65. de Mendoza, A., Sebe-Pedros, A. & Ruiz-Trillo, I. The evolution of the GPCR signaling system in eukaryotes: modularity, conserva-tion, and the transition to metazoan multicellularity. *Genome Biol. Evol.* **6**, 606–619 (2014).

66. Bockaert, J. & Pin, J. P. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* **18**, 1723–1729 (1999).

67. Frobius, A. C., Matus, D. Q. & Seaver, E. C. Genomic organization and expression demonstrate spatial and temporal Hox gene coli-nearity in the lophotrochozoan *Capitella* sp. I. *PLoS ONE* **3**, e4004 (2008).

68. Petrov, D. A. & Hartl, D. L. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl Acad. Sci. USA* **96**, 1475–1479 (1999).

69. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115 (2010).

70. Deng, W., Henriet, S. & Chourrout, D. Prevalence of mutation-prone microhomology-mediated end joining in a chordate lacking the c-NHEJ DNA repair pathway. *Curr. Biol.* **28**, 3337–3341.e3334 (2018).

71. Krokan, H. E. & Bjørås, M. Base excision repair. *Cold Spring Harb. Perspec. Biol.* **5**, a012583 (2013).

72. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Bio-chem.* **79**, 181–211 (2010).

73. Sfeir, A. & Symington, L. S. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem. Sci.* **40**, 701–714 (2015).

74. Taboada, S. et al. Bone-eating worms spread: insights into shallow-water *Osedax* (Annelida, Siboglinidae) from Antarctic, Subantarctic, and Mediterranean waters. *PLoS ONE* **10**, e0140341 (2015).

75. Holmes, R. P. The absence of glyoxylate cycle enzymes in rodent and embryonic chick liver. *Biochim. Biophys. Acta* **1158**, 47–51 (1993).

76. Jones, J. D., Burnett, P. & Zollman, P. The glyoxylate cycle: does it function in the dormant or active bear? *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **124**, 177–179 (1999).

77. Cioni, M., Pinzauti, G. & Vanni, P. Comparative biochemistry of the glyoxylate cycle. *Comp. Biochem. Physiol. Part B: Comp. Biochem.* **70**, 1–26 (1981).

78. Popov, V. N., Moskalev, E. A., Shevchenko, M. & Eprintsev, A. T. Comparative analysis of the glyoxylate cycle clue enzyme iso-citrate lyases from organisms of different systemic groups. *Zh . Evol. Biokhim Fiziol.* **41**, 507–513 (2005).

79. Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V. & Kondrashova, M. N. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol. Direct* **1**, 31 (2006).

80. Davis, W. L., Goodman, D. B., Crawford, L. A., Cooper, O. J. & Matthews, J. L. Hibernation activates glyoxylate cycle and gluco-neogenesis in black bear brown adipose tissue. *Biochim. Biophys. Acta* **1051**, 276–278 (1990).

81. DeSalvo, M. K., Sunagawa, S., Voolstra, C. R. & Medina, M. Tran-scriptomic responses to heat stress and bleaching in the elkhorn coral *Acropora palmata*. *Mar. Ecol. Prog. Ser.* **402**, 97–113 (2010).

82. Moran, N. A., McLaughlin, H. J. & Sorek, R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**, 379–382 (2009).

83. Nygaard, S. et al. Reciprocal genomic evolution in the ant-fungus agricultural symbiosis. *Nat. Commun.* **7**, 12233 (2016).

84. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).

85. Wilson, A. C. & Duncan, R. P. Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses. *Proc. Natl Acad. Sci. USA* **112**, 10255–10261 (2015).

86. Dietel, A. K., Merker, H., Kaltenpoth, M. & Kost, C. Selective advantages favour high genomic AT-contents in intracellular ele-ments. *PLoS Genet.* **15**, e1007778 (2019).

87. Itamar Luís, G., Albanin Aparecida, M.-P., Ana Claudia Piovezan, B. & Alice Teresa, V. Metabolic modeling and comparative bio-chemistry in glyoxylate cycle. *Acta Scientiarum. Biol. Sci.* **38** 1–6 (2016).

88. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).

89. Kingan, S. B. et al. A high-quality genome assembly from a single, field-collected spotted lanternfly (*Lycorma delicatula*) using the PacBio Sequel II system. *Gigascience* **8**, giz122 (2019).

90. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 3 (2011).

91. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

92. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).

93. Laetsch, D. & Blaxter, M. BlobTools: interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. *F1000Research* **6**, 1287 (2017).

94. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

95. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

96. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

97. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).

98. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. Genome-Scope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

99. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

100. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

101. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).

102. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

103. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

104. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

105. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).

106. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

107. Abby, S. S. & Rocha, E. P. C. Identification of protein secretion systems in bacterial genomes using MacSyFinder. *Methods Mol. Biol.* **1615**, 1–21 (2017).

108. Eichinger, V. et al. EffectiveDB–updates and novel features for a better annotation of bacterial secreted proteins and type III, IV, VI secretion systems. *Nucleic Acids Res.* **44**, D669–D674 (2016).

109. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and Ghost-KOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).

110. Kanehisa, M. & Sato, Y. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Sci.* **29**, 28–35 (2020).

111. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinforma.* **4**, 41 (2003).

112. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007).

113. Xu, L. et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, W52–W58 (2019).

114. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

115. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

116. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).

117. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

118. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

119. Smith, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013–2015).

120. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).

121. Wong, W. Y. & Simakov, O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. *Bioinformatics* **35**, 1051–1052 (2019).

122. Abrusan, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass–a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).

123. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).

124. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

125. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

126. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

127. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

128. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

129. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *Gigascience* **7**, giy131 (2018).

130. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* **7**, giy093 (2018).

131. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

132. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **6**, 31 (2005).

133. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

134. Bryant, D. M. et al. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Rep.* **18**, 762–776 (2017).

135. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).

136. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (Version v0.5.0) (Zenodo) (2022)

137. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

138. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma*. **10**, 421 (2009).

139. Mistry, J., Bateman, A. & Finn, R. D. Predicting active site residue annotations in the Pfam database. *BMC Bioinforma*. **8**, 298 (2007).

140. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. **20**, 238 (2019).

141. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

142. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*. **42**, D459–D471 (2014).

143. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. **44**, D457–D462 (2015).

144. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*. **34**, 525–527 (2016).

145. Chen, C. et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* **13**, 1194–1202 (2020).

146. Thiel, D., Yanez-Guerra, L. A., Franz-Wachtel, M., Hejnol, A. & Jekely, G. Nemertean, brachiopod, and phoronid neuropeptidomics reveals ancestral spiralian signaling systems. *Mol. Biol. Evol*. **38**, 4847–4866 (2021).

147. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol*. **7**, e1002195 (2011).

148. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).

149. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*. **30**, 772–780 (2013).

150. Seudre, O., Carrillo-Baltodano, A. M., Liang, Y. & Martin-Duran, J. M. ERK1/2 is an ancestral organising signal in spiral cleavage. *Nat. Commun*. **13**, 2286 (2022).

151. Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. **32**, D115–D119 (2004).

152. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

153. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

154. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

155. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol*. **32**, 268–274 (2015).

156. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).

157. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol*. **25**, 1307–1320 (2008).

158. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol*. **39**, 306–314 (1994).

159. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*. **32**, W327–W331 (2004).

160. Lu, S. et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. **48**, D265–D268 (2019).

161. Loffek, S., Schilling, O. & Franzke, C. W. Series "matrix metalloproteinases in lung health and disease": Biological role of matrix metalloproteinases: a critical balance. *Eur. Respir. J*. **38**, 191–208 (2011).

## Acknowledgements

## Author contributions

J.M.M.-D., L.H., G.R., and G.M. conceived and designed the study. G.R., N.R.-K. and S.G. collected the samples; G.M. assembled and annotated all genomes, and performed gene family evolution and metabolic complementarity analyses; B.P. assembled and annotated the symbiont genomes and contributed to metabolic complementarity analyses; Y.S., P.Q. and J.Q. performed analyses on PRR evolution; D.T. and G.J. did GPCR evolutionary analyses; F.M.M.-Z. performed Bayesian phylogenetic analyses; M.T. performed genomic extractions; A.M.C. and Martin Tresguerres contributed to host-symbiont metabolic analyses; G.M., L.H., and J.M.M.-D. drafted the manuscript and all authors critically read and commented on the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-38521-6.

**Correspondence** and requests for materials should be addressed to Lee M. Henry or José M. Martín-Durán.

**Peer review information** *Nature Communications* thanks Maxim Rubin-Blum and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.