Université de Montréal

Evolution, connectivity, and resilience in deep-sea chemosynthetic-based ecosystems

Par

Maeva Perez

Département des Sciences Biologiques, Arts et Sciences

Thèse présentée en vue de l'obtention du grade de docteur

en Sciences Biologiques

Mai 2023

Université de Montréal

Unité académique : Département des Sciences Biologiques, Arts et Sciences

Cette thèse intitulée

Evolution, connectivity, and resilience in deep-sea chemosynthetic-based ecosystems

Présenté par

Maeva Perez

A été évalué(e) par un jury composé des personnes suivantes

Sophie Breton

Président-rapporteur

Bernard Angers

Directeur de recherche

S. Kim Juniper

Codirecteur

C. Robert Young

Codirecteur

Pei-Yuan Qian

Codirecteur

Sandra Ann Binning

Membre

Marie-Anne Cambon

Examinateur externe

Résumé

La machinerie pour l'exploitation des ressources minérales qui se trouvent au fond des océans est déjà opérationnelle et la délivrance de permis miniers est imminente malgré d'inquiétantes lacunes de connaissances sur ces écosystèmes. En effet, dans une optique de sauvegarde, il est particulièrement important de mieux connaître les processus clés de l'écologie des profondeurs. Quelles sont les conséquences évolutives de la symbiose si répandue dans les écosystèmes profonds? Comment les sources hydrothermales sont-elles connectées? Quelles adaptations permettent la résilience des espèces endémiques de ces milieux? Dans le Pacifique est, la faune hydrothermale est caractérisée par de denses populations de palourdes de la famille Vesicomyidae et de vers polychètes tubicoles qui servent de niche à une multitude d'autres espèces. Ces deux groupes d'invertébrés dépendent pour leur nutrition de bactéries symbiotiques chimiolithotrophes. Celles-ci sont directement transmises de génération en génération chez les palourdes, et acquises de novo à partir de sources environnementales chez les vers. De plus, les vers possèdent une grande plasticité phénotypique associée aux conditions environnementales très variées (en terme de température, d'oxygène et de concentration de minéraux) dans lesquelles ils se retrouvent. Du fait du contraste dans leurs mode de transmission des symbiotes, de leur distribution étendue, et de leur rôle écologique important, ces deux groupes taxonomiques sont un excellent modèle pour étudier l'évolution, la connectivité et la résilience dans les écosystèmes marins profonds basés sur la chimiosynthèse. Ainsi, les objectifs de ma thèse sont de 1) déterminer les conséquences du mode de transmission des symbiotes sur leur évolution, 2) comparer la connectivité inter-sources entre les populations d'hôtes et de symbiotes, et 3) caractériser la méthylation de l'ADN chez les polychètes des sources hydrothermales et déterminer si ce mécanisme épigénétique joue un rôle adaptatif important. Ces objectifs sont abordés dans trois études indépendantes qui révèlent que 1) des processus à la fois neutres et sélectifs façonnent les génomes des symbiotes bactériens, 2) les populations de symbiotes bactériens dans les cheminées hydrothermales ne sont pas panmictiques mais sont influencées par des modèles locaux de connectivité, et 3) la méthylation de l'ADN est un mécanisme important d'adaptation dans les grands fonds marins. Ultimement, ces études permettent d'établir des lignes directrices en matière de conservation pour les opérations minières, et aident à l'établissement d'aires marines protégées.

Mots-clés : génomique, épigénomique, microbiologie, symbiose, évolution

Abstract

The mining industry is ready to exploit the mineral resources lying on the seafloor and the issuing of mining permits is imminent despite concerning knowledge gaps about the key evolutionary and ecological processes at play in these ecosystems. What are the evolutionary consequences of symbioses which are ubiquitous in deep-sea benthic ecosystems? How are chemosynthetic-based ecosystems connected? What kind of adaptations enable the resilience of vent endemic species to their extreme environment? In the eastern Pacific, chemosynthetic-based communities are characterized by dense aggregations of vesicomyid clams (in hydrocarbon seeps) or tubeworms (in hydrothermal vents) both of which offer habitat for many other species. Both invertebrates rely on chemolithotrophic bacteria for their nutrition. In the clams these symbionts are transmitted directly to the next generation through the eggs whereas in the tubeworms the symbionts are acquired *de novo* from the environment. The tubeworms also display striking phenotypic plasticity according to the physico-chemical conditions of their habitat. Because of their contrasting symbiont transmission mode, extended distribution, and ecological significance, these two taxonomic groups constitute an excellent model to study evolution, connectivity, and resilience in deep-sea chemosynthetic-based ecosystems. Thus, the objective of my thesis are to 1) identify the consequences of symbiont transmission mode on their evolution, 2) compare host and symbiont populations connectivity, and 3) characterize DNA methylation in deep-sea polychaetes and assess whether this epigenetic mechanism could explain their resilience. These objectives were addressed in three independent studies which revealed that 1) both neutral and selective processes participate in shaping the genomes of bacterial symbionts, 2) the populations of bacterial symbionts in hydrothermal vents are not panmictic but are influenced by local patterns of connectivity, and 3) DNA methylation is an important mechanism of adaptation in the deep-sea. Ultimately, these studies provide conservation guidelines for mining operations and help with the establishment of marine protected areas.

Keywords : genomic, epigenomic, microbiology, symbiosis, evolution

Table of contents

List of Tables

List of Figures

List of abbreviations

CBE: chemosynthetic-based ecosystem

SMS: seafloor massive sulfide

RGE: reductive genome evolution

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

MLSA: multi-locus sequence analyses

MEF: Main Endeavour Field

CB: Clam-Bed

MV: Middle Valley

OTU: operational taxonomic unit

BBNJ: biological diversity in areas beyond national jurisdiction

UNCLOS: United Nations convention of the law of the sea

ABNJ: areas beyond national jurisdiction

Thanks

What an adventure this thesis has been! I first want to acknowledge the funding agencies that made this work possible and enable me to travel and work in three different countries. Awards from the Université de Montréal's Arts et Science faculty and the Biological Sciences department helped me relocate from British Columbia to Québec at the beginning of my thesis. NSERC's doctoral fellowship gave me the opportunity to pursue my own projects and take risks. NSERC's internship supplement allowed me to join Dr. Young at the National Oceanography Center in Southampton (UK) for an incredibly formative six months where I picked up important computational genomics skills. The FRQNT/CSBQ's international internship award later allowed me to move to Hong-Kong to start a whole new and overly ambitious project in the laboratory of Prof. Qian. Lastly, the three papers presented in this thesis demanded a huge amount of computing resources; equivalent to 15 years of continuously running analyses on my little computer! These resources were provided by the Digital Research Alliance of Canada with top-notch user support. Special thanks to the clusters Cedar and Beluga for letting me flare up their CPUs.

Words cannot express my gratitude towards my co-supervisors. In addition to mentoring me throughout the epigenomics project, Prof. Qian gave me many opportunities to further my career beyond of the scope of this thesis. Professor, I am grateful for your financial and academic support and will strive to match your leadership and managerial skills. Dr. Robert Young motivated me to pursue studying the tubeworm's epigenetics and introduced me to the fascinating headache that is reductive genome evolution. Rob, your bleeding-edge expertise and infinite patience gave me the skills and confidence to call myself a computational biologist. As my Master's advisor, Prof. Kim Juniper introduced me to deep-sea research and showed me how cool microbes are. He persuaded me to follow-up on the idea of CRISPR-typing and gave me the resources to do so during my PhD. Kim, j'aspire à avoir ta vision avant-gardiste de la science de demain. Prof. Bernard Angers's direction has been the cornerstone of this thesis. He trusted we with the helm and steered the ship back when I was getting lost. Bernard, c'est au final dans ton cours de génomique évolutive que j'ai trouvé ma voie. Au doctorat, tu m'as poussé à de profondes réflexions en théorie écologique et philosophie et celles-ci me hanteront avec ravissement pour toujours.

I am also deeply indebted to my other mentors, Dr. Corinna Breusing, Prof. Jin Sun, Prof. Sun Yannan, and Dr. Lan Yi whose critical evaluations, constructive criticisms, and intellectual contributions have been instrumental in shaping the quality of my work. I would like to mention the Deep-Sea Biology Society for been instrumental in fostering these (and more) academic friendships and collaborations. And how could I not thank my fantastic friends and lab mates. Hinatea, Romain, Tatiana, Erik et Vincent vous m'avez supporté à travers toutes les étapes charnières de ce doctorat et dans les moments les plus durs. Jack, Jinping, Emma, Liu Xuan, Lan Yi, Xiao Yao and Wei Tong you made me feel at home in Hong-Kong and introduced me to delicious food that totally blew my mind.

Enfin, j'aimerais remercier ma famille d'enseignants qui m'ont fait aimé l'École depuis l'enfance et mes parents qui ont nourrit ma passion pour la science, la mer et le dépaysement. And Connor. You are sitting next to me as I write this in your office and you know what I have to say. There were plenty of cross-roads, and many, many lows. I would have quit a hundred times if you hadn't been by my side. We had to be separated again and again and for so long, but you made it easy. Wait. No. Not in that sense. You know what I mean. It has been a pleasure experiencing the two-body problem with you and I can't wait for us to overcome the next challenge.

Chapter 1 General introduction

Background information on chemosynthetic-based ecosystems

A collection of diverse extreme environments

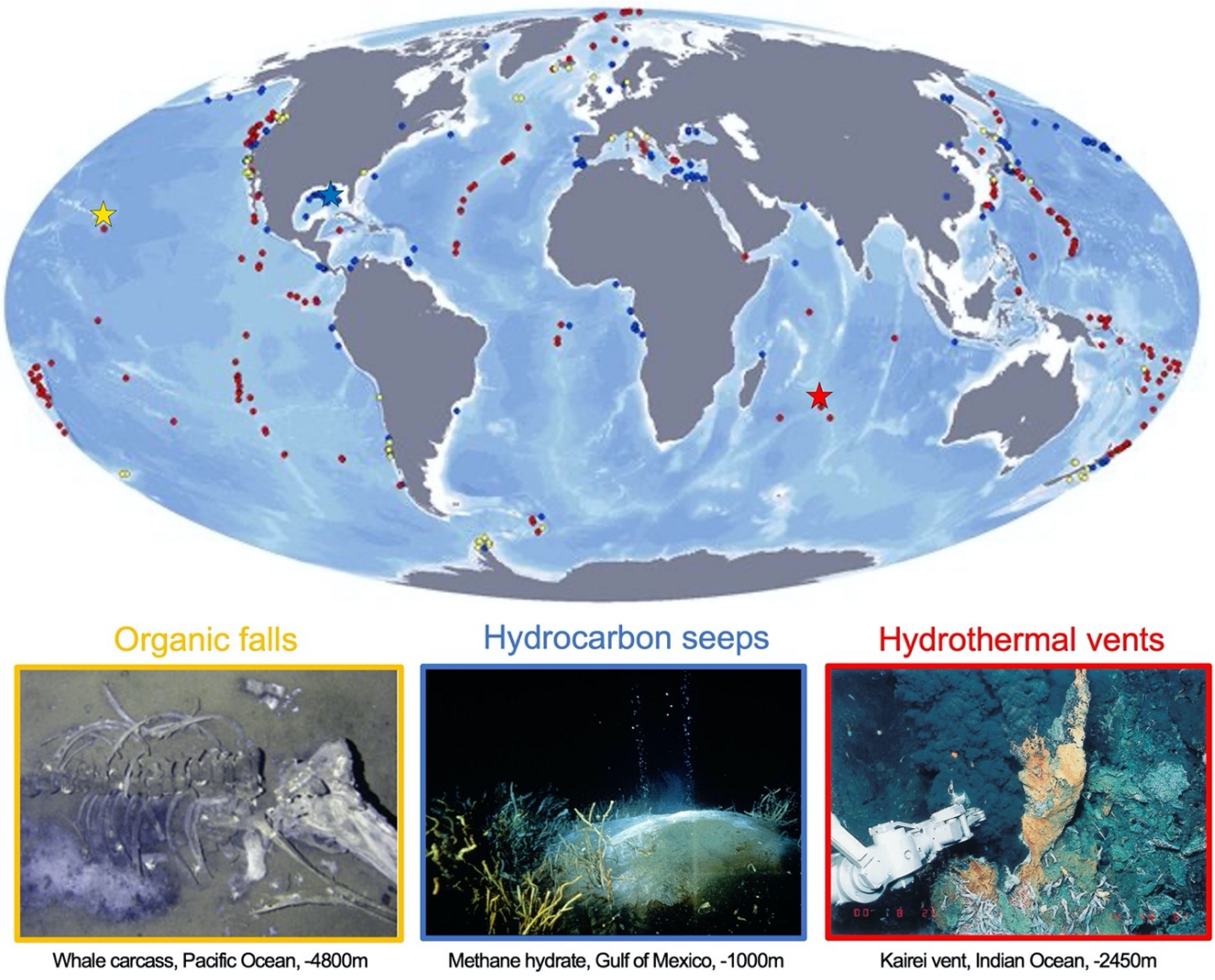

Figure 1.1 Global distribution and diversity of deep-sea chemosynthetic-based ecosystems. Top: Map of known deep-sea chemosynthetic based ecosystems in 2010 from German *et al.* (2011). Colored stars indicate the location of the images below. **Bottom**: Examples of chemosynthetic-based ecosystems. Photo credit from left to right: Hashimoto *et al.* (2001), Ian McDonald, Smith *et al.* (2015).

Deep-sea chemosynthetic-based ecosystems (CBEs) are benthic ecosystems found world-wide below the photic zone (*i.e.* below 200m). Such ecosystems include organic falls, hydrocarbon seeps

and hydrothermal vents (Figure 1.1). All are characterized by extreme hydrostatic pressure, a complete absence of light and photosynthetically-derived food, and extreme spatial physicochemical gradients.

Organic falls are sporadic remains of rather large organisms (*e.g.* whales) decomposing on the seafloor. Over time, multiple communities succeed each other on the organic carcass. In the third stage of decomposition (*i.e.* the sulfophilic stage), a CBE which can last for years or decades depending on the animal size is formed (Smith *et al.* 2015). In hydrocarbon seeps it is reduced chemical emanations from local hydrocarbon pockets or subsurface reservoirs which fuel chemosynthesis. These ecosystems are often found along continental margins where deep deposits of oil and gas can seep through faults in sedimentary rocks and under the right conditions of pressure and temperature, gas issued from the anoxic decomposition of organic matter by bacteria can accumulate as solid clumps at the sediment-water interface (Joye 2020). Finally, hydrothermal vents are submarine thermal springs. They are found anywhere there is volcanic activity and thus, are mostly located in the deep-sea (below 1000m) at the boundary of tectonic plates such as along mid-ocean ridges, and back arcs spreading centers which are associated with subduction zones (Beaulieu 2015). The hydrothermal fluids originate from seawater that seeps into the porous oceanic crust, exchange minerals with the surrounding rocks while being geothermally heated, and finally percolate upwards to a zone of discharge (Fontaine *et al.* 2009). At the discharge site, they form plumes that can reach temperatures up to 400°C and are rich in carbon dioxide and reduced compounds (Figure 1.2, A). Mixing with the cold surrounding seawater (~2°C) stimulates the precipitation of the minerals dissolved in these plumes creating chimney structures that can be several meters in height (Spiess *et al.* 1980).

A unique biology

Most of the food that reaches the bottom of the oceans comes in the form of marine snow; small detritus of organic matter that sink down from the euphotic zone. However, due to remineralisation on the way, less than one percent of the carbon fixed by the phytoplankton reaches the seafloor (De La Rocha and Passow 2014). This means an insufficient amount of photosynthesis-derived food is available away from the continental margins and in parts of the oceans with low surface primary productivity (Corliss *et al.* 1979). Yet, CBEs are able to sustain huge biomasses compared to the

surrounding benthic and pelagic deep-sea environments thanks to high concentrations of reduced compounds which enables chemosynthesis (Corliss *et al.* 1979).

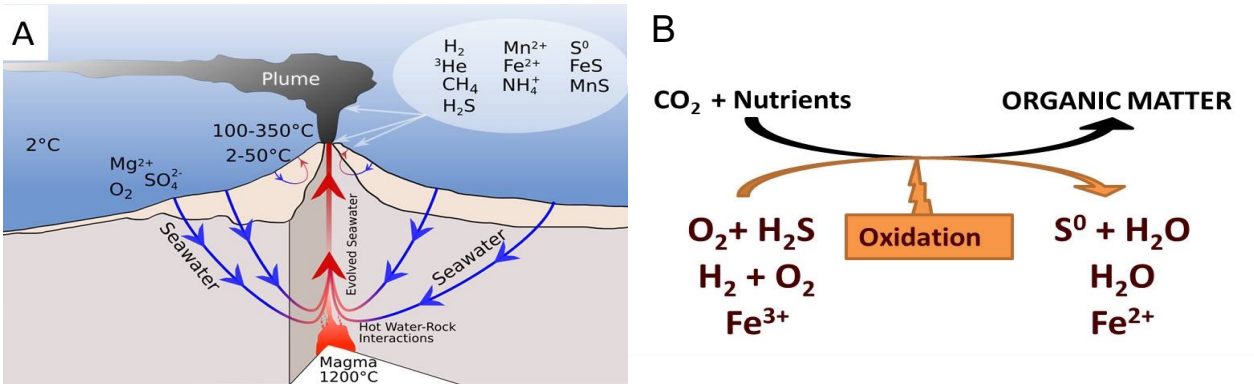

Figure 1.2 Reducing compounds in the hydrothermal fluids fuel chemosynthesis. A) Schematic cross section of a hydrothermal vent showing the reduced compounds present in the hydrothermal fluids; **B)** Schematic representation of chemosynthesis with three different examples of electrons donor; H_2S: hydrogen sulfide; H_2: dihydrogen; Fe^{3+}: iron III ions.

At the base of the food chain, there are lithoautotrophic bacteria and archaea that are able to oxidize the reduced compounds in the sediments, diffuse or hydrothermal fluids (*e.g.* hydrogen sulfide, methane, dihydrogen, iron) and use the resulting energy to fix carbon dioxide into organic molecules (Figure 1.2, B). For example, sulfur oxidizing bacteria (mostly belonging to the phyla Gamma- and Epsilon-proteobacteria) can harvest energy from hydrogen sulfide (H_2S) and are usually dominating the vent microbial assemblages on the surface of the sulfide chimneys and amongst the bacterioplankton of vent fluids (Meier *et al.* 2017; Olins *et al.* 2017).

The fauna in deep-sea CBEs is dominated by invertebrates and is characterized by endemism. For instance, the vast majority of vent animals are only found in vent ecosystems, with some species also living within hydrocarbon seeps and whale falls (Dubilier *et al.* 2008; Smith and Baco 2003). One remarkable adaptation these animals possess to deal with food scarcity is symbiosis which is diverse and ubiquitous (Figure 1.3). In fact, the dominant taxa in all deep-sea chemosynthetic-based communities are always associated with chemoautotrophic bacteria. Furthermore, deep-sea microbes and animals possess special adaptations to deal with high pressure, and low or high temperature (homeoviscous adaptations). For example, the hydrothermal vent worm *Alvinella pompejana* which lives on sulfide chimneys close to superheated fluid emanations possess a highly

thermos stable collagen that maintains proper stiffness of its body wall under high temperature conditions; up to 46ºC (Le Bris and Gaill 2007).

Last but not least, these environments are sources of great novelty. Organisms from CBEs possess many new genes thus offering great potential for biotechnology in the domains of agriculture, pharmaceutics and bioremediation (Thornburg *et al.* 2010). A beautiful example was brought recently by Ruan *et al.* (2017) who found that a particular molecule produced by the tubeworm species *Ridgeia piscesae* (a species that will be used as a model in this thesis) demonstrated promising anti-tumor properties when tested against HeLa and fibrosarcoma cell lines *in vitro*. Future discoveries are guaranteed to have important impacts on our understanding of ecology and evolution. In light of this emerging research, it is clear that CBEs represent a very promising but yet untapped pharmacopeia.

To sum up, because of their extreme environmental conditions, CBEs constitute unique and valuable models to investigate fundamental questions about the origins of biodiversity, and to understand the processes that give rise to novelty.

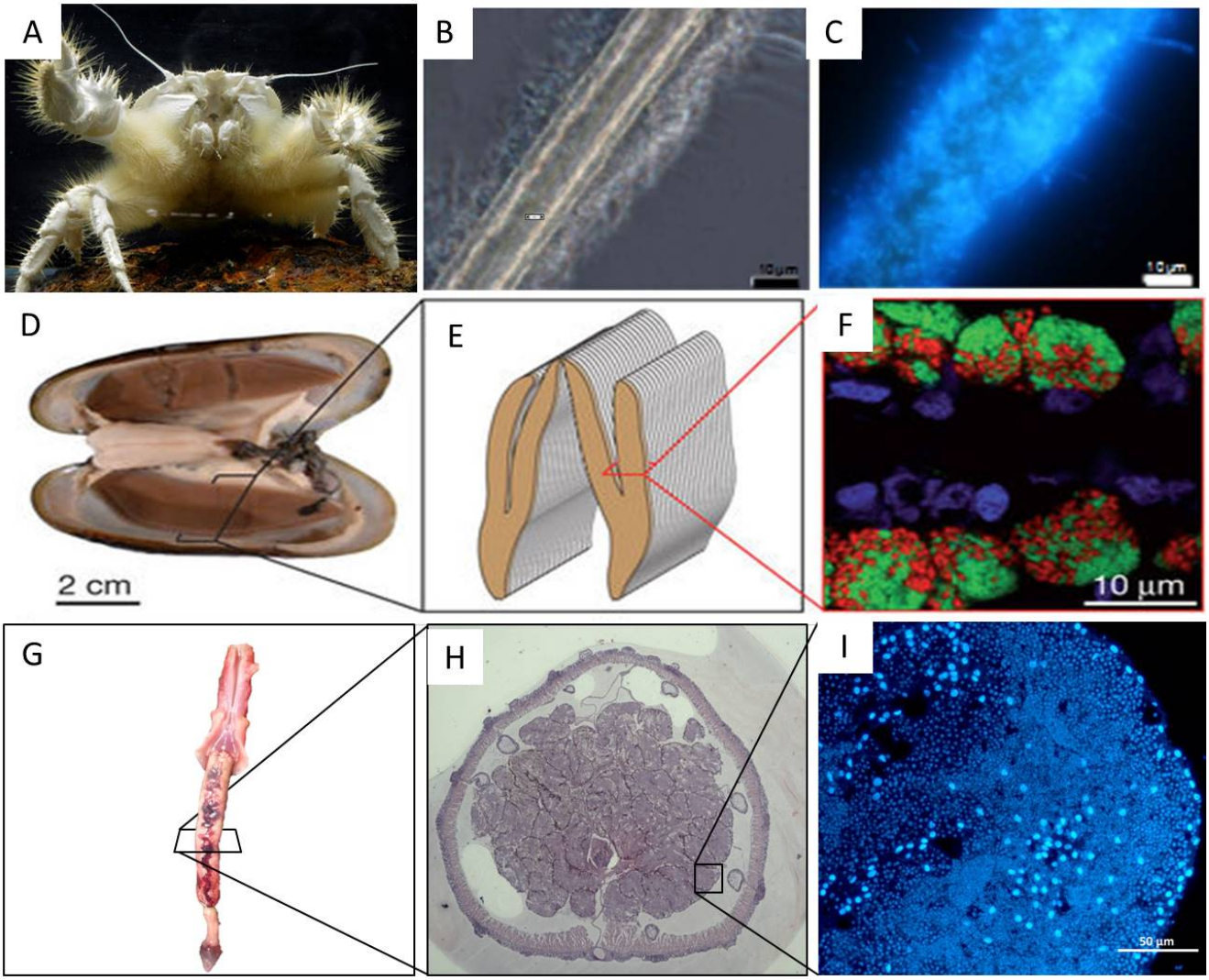

Figure 1.3 Three examples of symbioses with chemolithotrophic bacteria at hydrothermal vents. The crab *Shinkaia crosnieri* **(A)** possesses long and dense setae **(B)** on the ventral surface of its body which are colonized by epibiotic filamentous bacteria that consume hydrogen sulfide *(fluorescently labelled with DAPI in **C**). In this symbiotic association, the host cultivates the symbionts it feeds on by actively ventilating them with hydrothermal fluids* (Watsuji *et al.* 2012, 2017). The symbionts of *Bathymodiolus* sp. **(D)** are found within the cells of their gills **(E)** and have the particularity of using either methane [bacteria fluorescently labelled in red in **F**]) of sulfide [bacteria fluorescently labelled in green in **F**]) as energy source. This allows the mussels to survive in a broad range of environmental conditions (Duperron *et al.* 2006). Finally, the polychaete worm *Ridgeia piscesae* **(G)** does not have a digestive system but possesses instead a massive organ called trophosome **(H)** within which it hosts a single phylotype of sulfur-oxidizing symbionts (fluorescently labelled by DAPI in **I**). This worm is completely dependent on its symbionts for its nutrition. Photo credits: A: Japan Agency for Marin-Earth Science and Technology Science; B and C: Konishi *et al.* 2013; D, E, and F: Petersen *et al.* 2011; H: courtesy of Candice St-Germain.

An imminent threat

While CBEs are rich in genetic resources, they are also rich in mineral resources. On the one hand, gas hydrates commonly found in hydrocarbon seep environments are increasingly considered as an economically viable alternative source of natural gas that could be used to supplement or substitute depleting underground reservoirs. The first field tests for gas hydrate exploitation were conducted in 2002 in Canada in the Mallik permafrost region (Beaufort Sea). The extreme conditions favorable to gas hydrate formation makes their exploitation technically challenging, but over 80 countries have already begun research activities (Yang *et al.* 2019). As of 2020, nearly 4000 patents for gas hydrate extraction technology have been attributed worldwide (Wang *et al.* 2022). On the other hand, hydrothermal vent sediments form polymetallic sulfides also called seafloor massive sulfide (SMS) deposits with various concentrations of metals (mostly zinc, copper, silver, and gold) depending on the composition of the hydrothermal fluids (Fouquet *et al.* 1991). These mineral deposits are scattered and represent a much smaller tonnage than the terrestrial deposits (Van Dover *et al.* 2018). Yet, increased demand and resource depletion on land along with new technological advances make the mining of SMS deposits in the deep-sea an economically viable and attractive option (Figure 1.4).

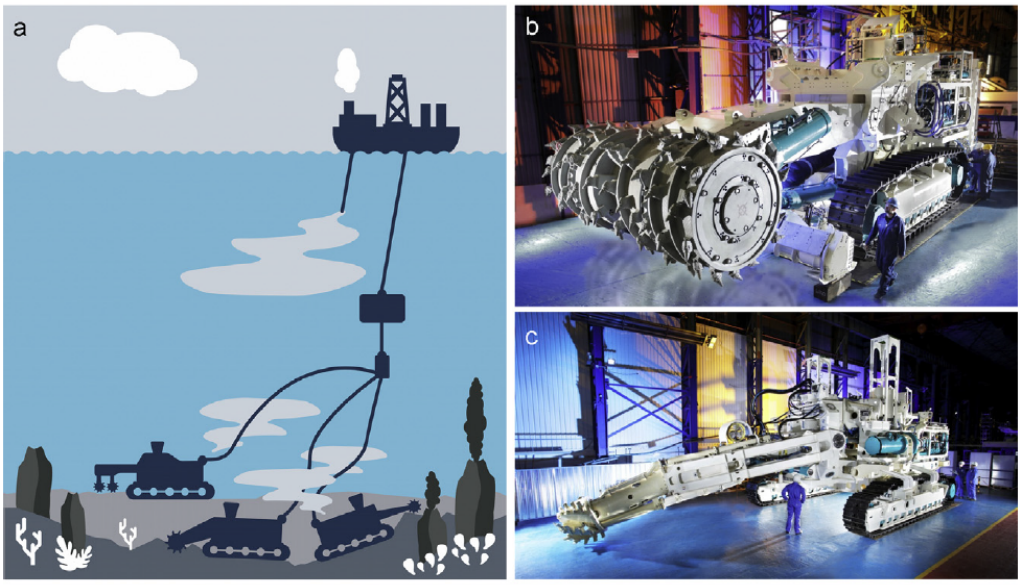

Figure 1.4 Mining hydrothermal vents. A) Graphical representation of a polymetallic sulphide mining operation. SMS mining activity would consist of crushing the polymetallic sulfides into small transportable sizes with remotely operated vehicles, pumping the resulting slurry to the ship, de-watering the ore-water mixture, and finally discarding the waste waters; **B)** Canadian company Nautilus Mineral's Bulk cutter machine; **C)** Nautilus Minerals's collection machine. From Gollner *et al.* (2017).

Under current exploitation plans, gas hydrates and SMS mining are foreseen to be disastrous for their respective ecosystems (Gollner *et al.* 2017). Mining would directly cause the death of all the animals growing on or near the deposits. Furthermore, the traffic of the crawling vehicles and resulting sediment resuspension plumes is likely to affect larval dispersal and settlement and may release toxic chemicals in the vicinity of the exploited area and beyond affecting both the microbial and animal communities. Changes in chemical concentrations and local circulation are expected to cause habitat loss and habitat fragmentation and thus have lasting effects on CBEs (Ellis *et al.* 2017; Gollner *et al.* 2017; Reed *et al.* 2015; Van Dover *et al.* 2018).

Another issue specific to hydrothermal vent ecosystems is that most are located in international waters. Therefore their SMS resources are part of "the common heritage of mankind" (*UNCLOS - Part XI, Section 2* 1982), and are currently under the jurisdiction of the International Seabed Authority (ISA) that, in absence of a legal framework for the regulation and monitoring of mining operations, has not issued exploitation permits yet. The ISA is working with stakeholders and lawmakers to draft those regulations by 2023 (International Seabed Authority 2018). Unfortunately, the hydrothermal vents that are located in the various countries' exclusive economic zone (EEZ) and Extended Continental Shelves (ECF) are not protected by these precautionary measures. In October 2017, the Japanese government were the first to proceed with a mining test at an inactive hydrothermal vent site in their EEZ (METI 2017), and in Papua New Guinea, negotiations for commercial operations have been conducted for several years (Nautilus Minerals 2018).

For better or worse, global economic and societal interests have shifted the focus of deep-sea scientific research towards questions of conservation. Within this framework, deep-sea scientists, notably biologists, have been tasked with the responsibility of delivering evidence-based assessments of the resilience (or lack thereof) of CBEs, and defining effective measures for safeguarding these ecosystems against the impacts of human activities.

Thesis objectives and hypotheses

This thesis is composed of three independent studies that contribute to the global effort for understanding the key processes shaping CBEs and developing conservation guidelines for the upcoming mining operations.

Conservation efforts face significant challenges due to the lack of fundamental understanding regarding the evolution and ecology of deep-sea species; 1) How have deep-sea taxa have evolved?, and 2) How connected are the diverse deep-sea ecosystems today? **Our general objective is to determine what factors influence the evolution and connectivity of symbiotic bacteria, and what the mechanisms of resilience in deep-sea chemosynthetic ecosystems are** at the level of the populations and species.

The first (Chapter 2) and second (Chapter 3) studies respectively focus on the evolution and connectivity of endosymbionts associated with two vent and seeps taxa: the clams of the family Vesicomyidae and the siboglinid tubeworms. The clams' symbionts are maternally transmitted while these of the tubeworms are environmentally acquired.

The symbiont transmission mode strongly influences the evolution of the holobiont. Vertically transmitted symbionts co-disperse and co-evolve with their host. Consequently, their genomes are strongly affected by reductive evolution (see Key concepts); a phenomenon which remains to be fully understood. In the first study, I therefore used the vesicomyid symbiosis as a model to better parse the interplay between the neutral and selective factors that affect the evolution of vertically-transmitted symbioses.

On the contrary, environmentally acquired symbionts do not co-disperse with their hosts. In CBEs, the structure and connectivity of environmentally acquired symbiont populations remains largely undescribed because these bacteria cannot be easily cultured and their strains differentiated. Thus, the second study's aim was to assess if the recently discovered "CRISPR" gene can be used as a marker to uncover the regional genetic diversity and population structure of uncultured CBE bacteria. To do so, I used the model symbiont species *Candidatus* Endoriftia persephone which associates with the siboglinid tubeworm *Ridgeia piscesae*.

Finally, the third study (Chapter 4) focuses on animal resilience. Specifically, I investigate if and how DNA methylation (an epigenetic mechanism that allows many species to cope with changing environments) is used by vent and seep polychaete species. The functional roles of DNA methylation are poorly understood in invertebrates and no genome-wide surveys of DNA methylation in polychaetes have been conducted so far. This study's objective is to fill this knowledge gap by sequencing the first methylomes for three polychaetes species inhabiting CBEs: two species of siboglinid tubeworms, *R. piscesae* and *Paraescarpia echinospica*, and the alvinellid *Paralvinella palmiformis*.

Table 1.1 summarizes the objectives and hypotheses pertaining to each chapter of this thesis. Fulfilling these objectives posed significant challenges. Indeed, CBEs are inherent difficulty to access and study *in situ*. Furthermore, most deep-sea animals have yet to be successfully kept outside of their natural habitat for more than a few days and therefore, experiments in controlled conditions typically have to be executed on board scientific ships using expensive equipment (*e.g.* hyperbaric aquaria). As for deep-sea bacteria, most remain uncultured (not for lack of trying). To reach my goals, I therefore turned to genetic and genomic approaches. The molecule of DNA is a formidable historical record and by leveraging its power, one can learn much about the biology of an organism.

By comparing the whole genomes of vesicomyid-associated bacteria and their free-living relatives in the first study (chapter 2), I shed light on different aspects of the evolutive history, and current ecology of the holobionts. In the second study (chapter 3), a new genetic marker allowed me recognize the sheer diversity of tubeworm symbionts and infer their patterns of connectivity at a relatively small regional scale. Lastly, in the third study (chapter 4) the genomes of three deep-sea worm species, enabled me to understand how these worms can regulate their metabolism to comply with the extreme conditions of their environment.

The subsequent section, titled "Key Concepts," presents the fundamental principles pertaining to each of my studies and describes how DNA sequences are analyzed and utilized to extract valuable biological information. The next section (The models) introduces the three animal models used for these studies: vesicomyid clams, siboglinid tubeworms and alvinellids.

Table 1.1 Summary of thesis objectives, hypotheses and expected results.

STUDY 1: COMPARATIVE GENOMICS OF THE SYMBIONTS OF VESICOMYID CLAMS		
Theme: *How have deep-sea taxa evolved?*	*Keywords:* *Molecular evolution, symbiosis*	*Model:* *Vertically-transmitted endosymbionts of vesicomyid clams*
Main Objective	**Determine the relative contribution of neutral and selective processes to the evolution of vesicomyid clams endosymbionts**	
Sub-Objective 1 : **Determine if the vesicomyid symbiosis is typical of that of vertically transmitted symbionts**		
Hypothesis	The vesicomyid symbionts represent an intermediate state of reductive genome evolution	
Expected results	· Reduced size, low GC, gene loss etc. · Contrasting differences in genomic features between symbionts	
Methods	*Computation of genome characteristics and comparisons*	
Sub-Objective 2 : **Determine whether the symbionts genome evolution is affected by neutral processes**		
Hypothesis	Genetic drift strongly affects the genome evolution of clam symbionts	
Expected results	· Shift of symbiont dN/dS towards neutrality, reduction of codon usage bias	
Methods	*Gene-wise comparison of dN/dS between symbionts and free-living*	
Sub-Objective 3 : **Determine if selective processes also play an important role in symbiont genome evolution**		
Hypothesis	Arms race evolution affects the clam symbionts	
Expected results	· Episodic positive selection in genes that affect host-symbiont communications	
Methods	*Detection of episodic selection through dN/dS analysis of all core genes, comparisons of genes presence/absence*	

STUDY 2: DIVERSITY AND POPULATION STRUCTURE OF *R. PISCESAE* SYMBIONTS		
Theme: *How connected are hydrothermal vents?*	*Keywords:* *Bacterial population genetics, connectivity*	*Model:* *Populations of Ridgeia piscesae symbionts on the Juan de Fuca ridge*
Main Objective	**Characterize the structure of the populations of *R. piscesae* symbionts on the Juan de Fuca ridge**	
Sub-Objective 1 : **Determine the extent of genetic diversity in the symbionts**		
Hypothesis	There are multiple strains of symbionts in the environment	
Expected results	· Same symbiont genotypes across multiple hosts	
Methods	*Identification of symbiont genotypes through MLSA and CRISPR typing*	
Sub-Objective 2 : **Determine how the symbiont populations are partitioned**		
Hypothesis	Symbionts cluster into multiple ecotypes according to environment	
Expected results	· Correlation genetic structure of populations and environmental parameters	
Methods	*Analyses of partition of variance to determine if habitat type (HF, LF) significantly contributes to symbiont partitioning*	

STUDY 3: FUNCTIONAL ROLES OF DNA METHYLATION IN THREE POLYCHAETE SPECIES		
Theme: *How resilient are CBE species to environmental changes?*	*Keywords:* *Epigenomes, DNA methylation, Evolution,*	*Model:* *Paraescarpia echinospica, Ridgeia piscesae and Paralvinella palmiformis*
Main Objective	**Assess if DNA methylation is an important epigenetic mechanism in deep-sea worms**	
Sub-Objective 1 : **Assess whether the worms possess a functional metabolism for DNA methylation**		
Hypothesis	*The worms possess a functional DNA methylation metabolism*	
Expected results	· Key genes of the DNA methylation metabolism are present in the worms' genomes · Key genes of the DNA methylation metabolism are present in the worms' transcriptomes	
Methods	*Genomes and transcriptomes sequencing, assembly, and annotation*	
Sub-Objective 2 : **Evaluate if the methylome can be obtained from third generation sequencing technology**		
Hypothesis	*DNA methylation is accurately detected by third generation sequencing*	
Expected results	· Methylation levels estimates similar between Nanopore sequencing technology and Whole-Genome Bisulfite Sequencing	
Methods	*Comparative analyses of DNA methylation profiles obtained through Nanopore and WGBS technologies*	
Sub-Objective 3 : **Test established hypotheses about the roles of DNA methylatio**n		
Hypothesis 1	*Gene-body methylation upregulates gene expression*	
Expected results	· Expression of hypermethylated genes significantly higher than hypomethylated genes	

Hypothesis 2	*Promoter methylation down regulates gene expression*
Expected results	· Expression of hypermethylated genes significantly lower than hypomethylated genes
Hypothesis 3	*Transposable elements are silenced by DNA methylation*
Expected results	· Younger, presumably more mobile transposable elements are more methylated than older ones
Methods	*Determination of genome-wide methylation patterns of Ridgeia piscesae and Paralvinella palmiformis Comparative genomic and epigenomic analyses on orthologous genes*

Key concepts

Reductive genome evolution and selective processes at play in host-restricted symbioses

Regardless of their phylogenetic origin, host, or habitat, the genomes of vertically transmitted symbionts all share the same characteristics; compared to environmentally acquired symbionts or free-living bacteria, they are smaller, contain fewer genes, almost no non-coding sequences, and are enriched in AT (depleted in GC). The symbionts also display accelerated mutation rates (Moran 1996; Wernegreen *et al.* 2001; Itoh *et al.* 2002; Herbeck *et al.* 2003; Moran *et al.* 2009) and reduced codon bias (Wernegreen and Moran 1999; Rispe *et al.* 2004). These unique characteristics are thought to result from the same phenomenon of reductive genome evolution (RGE).

An important body of research conducted on the pea aphid/*Buchnera* and a few other insect/bacteria models has suggested that RGE could be explained solely by neutral processes (Figure 1.5). Captured symbionts are contained within their host's cells and therefore isolated from their free-living relatives limiting opportunities for horizontal gene transfer (Bordenstein and Reznikoff 2005). They also experience successive bottleneck events during their transmission which reduce their effective population size and increase the relative effect of genetic drift over selection in determining the fate of alleles in their population (Kuo *et al.* 2009). As a consequence, the symbionts experience rapid fixation of slightly deleterious mutations and reduction of intra-host genetic diversity (Wernegreen and Moran 1999; Funk *et al.* 2001).These mutations accumulate overtime because in asexual populations, they cannot be purged by homologous recombination; a phenomenon coined the Muller ratchet (Felsenstein 1974). Mutations in bacteria are naturally biased towards deletions (Mira *et al.* 2001) and spontaneous base-substitution mutations are biased towards AT (Hershberg and Petrov 2010) explaining respectively, the erosion and shift in nucleotide composition observed in symbiont genomes. Non-synonymous substitutions on the one hand, may result in the loss of a start codon or the gain of premature stop codons in coding sequences which, along with reading frame disruption caused by insertions and deletions, eventually lead to the inactivation of genes (pseudogenization). These genes are then slowly eroded and removed from the genome because of the aforementioned deletional bias of mutations (Mira *et al.* 2001; Kuo *et al.* 2009). On the other hand, synonymous substitutions contribute to reducing

the codon bias for translational efficiency (Sharp and Li 1986a; Wernegreen and Moran 1999). Furthermore, AT enrichment increases the likelihood of generating homopolymer stretches of A and Ts and thus the probability of additional deletions by slippage of the replication machinery (Moran *et al.* 2009). Itoh *et al.* (2002) suggested that the characteristics of symbiont genomes could also be explained by increased mutation rates (caused for example by the loss of genes involved in DNA repair), or by genome-wide relaxation of selection (the host likely provides its symbionts with optimal growth conditions). Relaxed selection would also enable the proliferation of mobile elements which would further aggravate gene loss through pseudogenization (Lawrence *et al.* 2001); a hypothesis supported by multiple observations of increased proportion of mobile elements in the genomes of recently acquired symbionts (Gil and Belda 2008; Plague *et al.* 2008; Koga and Moran 2014).

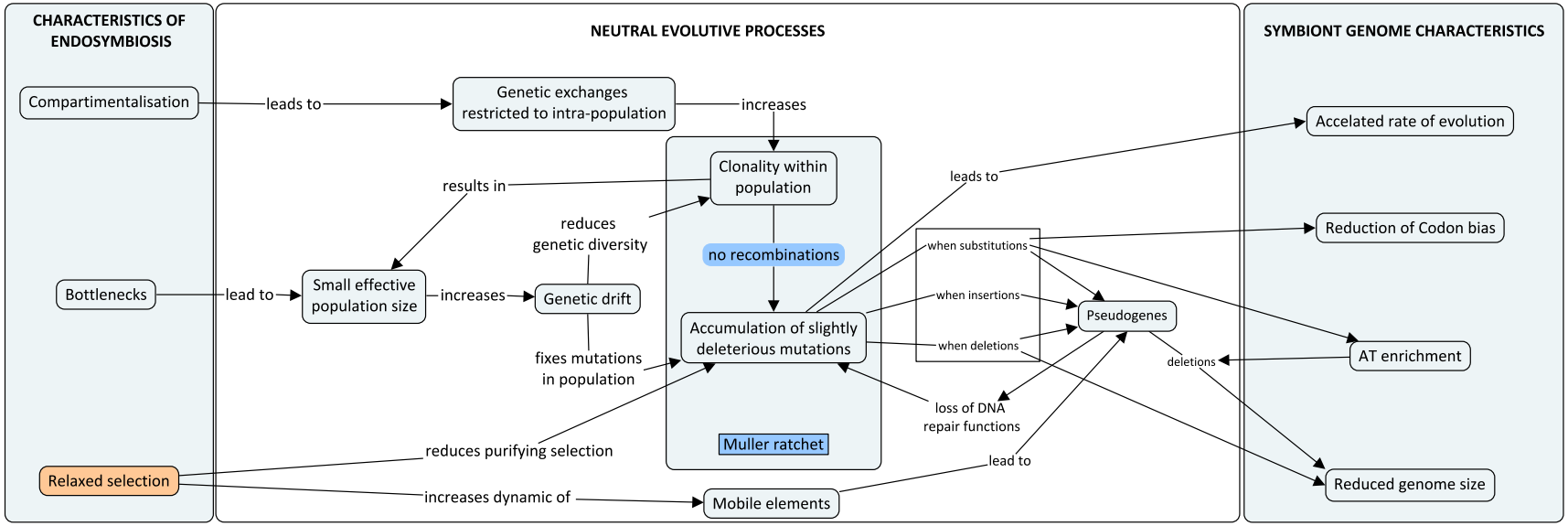

Figure 1.5 Symbiont reductive genome evolution explained by neutral processes.

If symbiont genome reduction depends on neutral processes alone, it is expected that they will lead to symbionts with increasingly poor fitness and eventually prompt the extinction of their hosts (Bennett and Moran 2015). Because of this, RGE is sometimes referred to as the 'genome reduction syndrome' (Latorre and Manzano-Marín 2017). Paradoxically, there is little evidence the symbionts lose their fitness overtime (Baumann 2005). Symbiont replacement was observed in some models (Pérez-Brocal *et al.* 2006; Koga and Moran 2014; Sudakaran *et al.* 2017; Chong and Moran 2018) and has been interpreted as a mechanism for escaping this evolutionary rabbit hole thus supporting the neutral theory of symbiont evolution, but it is unclear if these replacements happened as a result of symbiont degeneration.

Conversely, the reductive genome evolution observed in endosymbionts can also be explained by a number of selective processes. For instance, selection could favor genome streamlining because it allows for faster cell proliferation at lower cost (Giovannoni *et al.* 2014). Indeed, for free-living bacteria in nutrient poor waters, reduced genome size and higher AT content lead to significant reduction of the Nitrogen budget (Grzymski and Dussaq 2012). In the case of symbiosis, having low nutrients demands and being able to rapidly expand when the environmental conditions are favorable would be advantageous for the survival of a bacterial population and its host. Gene loss in symbionts may also be explained by selective pressure against redundancy; *i.e.* the Black Queen hypothesis (Morris *et al.* 2012). This hypothesis posits that a selective advantage exists for bacteria that lose functions already provided by the community because it allows them to allocate less resource to protein synthesis without scarifying functions. The community can be a minority of the symbionts or the host itself (McCutcheon and Moran 2007; Rio *et al.* 2016). Endosymbiont gene transfer to the nuclear genome of the host is very well documented for animal and plant organelles (Timmis *et al.* 2004) and has been observed in the endosymbionts *Wolbachia* (Hotopp *et al.* 2007) and *Paulinella* (Nowack *et al.* 2011). However, for several other models of Eukaryote/ bacteria intracellular symbioses such as the mealybug/*Tremblaya princeps* (Husnik *et al.* 2013), the human body louse/*Riesia* (Kirkness *et al.* 2010), or the pea aphid/*Buchnera* (Nikoh *et al.* 2010; McCutcheon and Moran 2012), no evidence of lateral gene transfer from the symbiont to the host were found.

Finally, arms race dynamics are expected to occur in co-evolving host and symbionts in order to maintain the host-symbiont specificity and the functioning of cyto-nuclear interactions through

speciation (Bennett and Moran 2015). Indeed, diversifying selection acting on genes involved in the mediation of host-symbiont interactions such as liposaccharides, and peptidoglycans was observed in divergent clades of *Wolbachia* (Brownlie *et al.* 2007) and many facultative endosymbionts (Dale and Moran 2006). In a recent study, Chong et al (2019) performed the first genome-wide screens for selection the *Buchnera* of the aphid subfamily Aphidinae. Of the 371 protein-coding genes tested, the authors detected 29 positively selected genes representing a variety of metabolic functions including notably two outer membrane porins (OmpF and OmpA) which are suspected to be important for host interaction (Chong *et al.* 2019).

In summary, it is clear that further studies and a greater number and diversity of models are needed to decipher the relative contribution of neutral and selective processes to the symbiont molecular evolution. In Chapter 2's study, I investigated symbiont evolution in the vesicomyid clams (see the model description at page 32).

Patterns of connectivity within meta-populations and the case of bacteria

CBEs are patchy hot spots of biodiversity separated by long stretches of desertic ocean floor (Corliss *et al.* 1979). Hence, vent and seeps species are presumably composed of multiple interconnected sub-populations (Vrijenhoek 1997). The structure of, and flux within these networks are closely linked to the persistence of these metapopulations and is therefore of great interest for conservation.

In a nutshell, the connectivity patterns at hydrothermal vents is expected to follow two main schemes: the 'stepping stone model' or the 'island model' (Wright 1943; Kimura and Weiss 1964; Vrijenhoek 1997) (Figure 1.6). In the steppingstone model, venting sites are well connected to their close neighbours leading to populations that are isolated by increasing distances. In this scenario, the destruction of a habitat can lead to cuts in the gene flow leaving some populations completely isolated and therefore more vulnerable. In the island model, species that have high dispersal capabilities are not affected by distance but differences in habitat quality or biases in the direction of the dispersal (*e.g.* from oceanic currents) may result in patches of 'source' populations producing a lot of dispersing migrants and 'sink' populations sustained by constant import of new recruits (Vrijenhoek 1997). In this scenario, oceanic currents emerge as a crucial determinant of the genetic structure of populations. Accurately estimating the direction and magnitude of gene flow thus

becomes paramount in safeguarding these ecosystems, as the extinction of a "source" population due to marine operations would detrimentally impact the "sink" sites. Estimations of macrofaunal connectivity are already an integral part of several frameworks aiming at assessing the resilience of proposed mining sites and developing preservation areas (International Seabed Authority 2010; Boschen *et al.* 2016; Ellis *et al.* 2017) . However, it is important to consider microbial connectivity as well, especially for bacterial species that have an important ecological role such as these forming symbioses with vent keystone species.

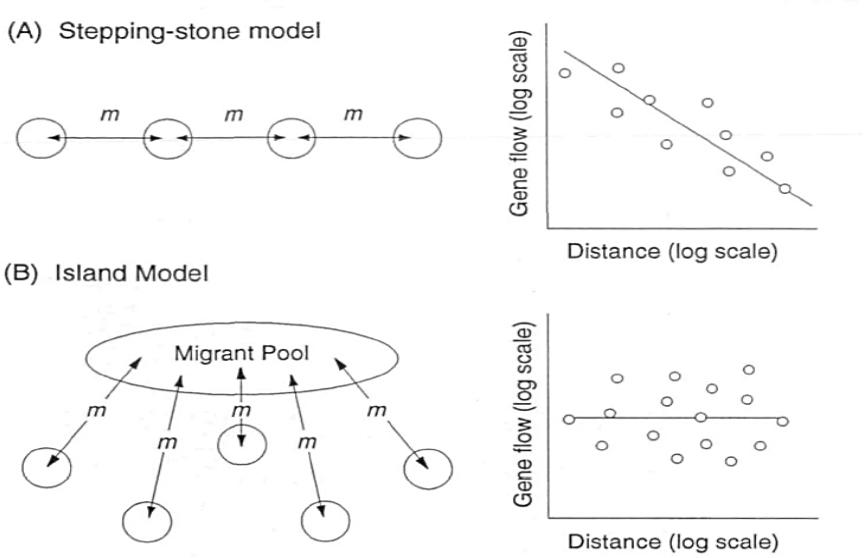

Figure 1.6 Two common models of connectivity between hydrothermal vents. A) The stepping stone model: populations are isolated by increasing distances. **B)** In the island model, populations have high dispersal capabilities and gene flow is not affected by distance. From Vrijenhoek (1997).

The persistent idea formulated by Baas-Becking (1934) that when it comes to bacteria “everything is everywhere, and the environment selects” probably explains why bacterial species have so far not been included in the design of marine protected areas. Due to their small size, extensive population sizes (Fenchel 2003), and capability for dormancy (Stevenson 1977; Sachidanandham and Yew-Hoong Gin 2008), prokaryotes were initially believed to possess exceptional dispersal capacities, enabling them to establish themselves successfully in any environment where local conditions, both biotic and abiotic, are favorable (Finlay and Esteban 2004). This paradigm was initially supported by the cosmopolitan distribution of bacteria at different taxonomic levels (and down to the species) in both terrestrial and aquatic systems (Roberts and Cohan 1995; Wirsen *et al.* 1998; Staley and Gosink 1999; Glöckner *et al.* 2000; Zwart *et al.* 2002; Ward and O’Mullan

2002; Morris *et al.* 2002; Finlay and Esteban 2004; Jones *et al.* 2016). For example, the sulfur-oxidizing chemolithoautotroph *Hydrogenovibrio thermophilus* was isolated from hydrothermal vents in both the mid-Atlantic and Juan de Fuca (Pacific Ocean) ridges, and the two isolated clones possessed identical sequences for their 16S rRNA gene (Boden *et al.* 2017). However, the visible structure of bacterial populations is limited by the genetic markers used to describe it; the higher gene conservation (the lower the diversity on a locus), the lower the resolution of the population structure. Thus, the very conserved 16S rRNA gene used in many studies claiming Bacteria were globally distributed is probably not variable enough to detect genetic diversity at the population level (Cho and Tiedje 2000). The two aforementioned clones of *Hydrogenovibrio thermophilus* for example (Boden *et al.* 2017), diverge by about 3.5% (at the nucleotide level) when their whole genome is considered (Table 1.2) suggesting allopatric differentiation.

Table 1.2 16S rDNA identity and genome-wide average nucleotide identity (gANI) comparisons between two bacterial species isolated in hydrothermal vents in different oceans. gANI performed through JGI's-IMG Pairwise ANI tool

Species	***Hydrogenovibrio* sp. MA2-6**	***Hydrogenovibrio* thermophilus JR-2**
Location	Snake Pit hydrothermal vents (Atlantic Ocean)	Juan de Fuca hydrothermal vents (Pacific Ocean)
IMG genome ID	2571042363	2836772259
NCBI BioSample	SAMN02745443	SAMN10724091
Genome size	2676715	2612894
16S identity		100 %
mean gANI		96.5 %

These observations indicate that new and more variable markers must be used to uncover the strain-level diversity of bacteria and assess the structure of their populations. In the study of Chapter 3, I conducted an inquiry of the population structure of the symbionts of *R. piscesae* (see the model description at page 34)

The epigenome and the roles of DNA methylation

In its broader sense, the epigenome is the level of information between the genome and the phenotype that integrates information from the both the genetic sequence and the environment (Waddington 1942). More specifically, the epigenome is the state of expression of a genome that

is stably heritable through mitoses but not immutable (Berger *et al.* 2009). A specific epigenetic state is initiated and maintained through cell division by distinct but interconnected molecular processes which without ever changing the underlying DNA sequence, affect the phenotype (Russo *et al.* 1996). These intracellular pathways end up modulating transcription (*e.g.* histone modification restricting the accessibility of genes [Bártová *et al.* 2008], DNA methylation preventing the recruitment of the transcription machinery [Mandrioli 2007; Massicotte *et al.* 2011]), translation (*e.g.* microRNAs hybridizing mRNAs [Filipowicz *et al.* 2008]), and protein conformation (prions [Halfmann and Lindquist 2010]). Transition between different epigenetic states can happen actively via deactivation or activation of new epigenetic effectors, or passively via dilution of epigenetic effectors across cell division (Hitchins 2015). Thereupon, multiple phenotypes can arise from a single genotype leading, for example, to the development of different tissues in metazoa (Waddington 1957). Importantly, because the epigenetics is influenced not only by intrinsic (encoded in the genome) but also extrinsic (from the environment) signals (Jaenisch and Bird 2003), it allows for the generation of non-genetic variation (Bossdorf *et al.* 2008). Thus, epigenetic mechanisms allow organisms to respond to environmental shifts at both the level of the individual (*i.e.* acclimation) and the population (*i.e.* adaptation). The processes by which epigenetic variation allows populations to persist through environmental changes and thus increase their resilience, coined 'epigenetic buffering' by O'Dea *et al.* (2016), is of particular ecological significance in the face of climate change and anthropogenic disturbances (Morris 2014; Jeremias *et al.* 2018).

A key challenge in detecting epigenetic variation is to target the appropriate epigenetic marker. DNA methylation is the best studied epigenetic mechanism (Suzuki and Bird 2008). The methylation of the genome consists in the addition of methyl groups to the cytosines or the adenines. DNA methylation negatively affects the functioning of the transcription machinery by, amongst others, preventing the binding of its recruiters to the DNA (Klose and Bird 2006). This way, heavy methylation of the promotor region results in gene silencing whereas hyper-methylation of the gene body may prevent the mispairing of the RNA polymerase and therefore the production of aberrant transcripts (Keller *et al.* 2016). Furthermore, there is evidence that variation in gene body methylation can lead to exon skipping (Lyko *et al.* 2010).
Consequently, DNA methylation can provide phenotypic variation through differential activation/repression of genes and production of alternative transcripts (Roberts and Gavery 2012).

DNA methylation has been well described in vertebrate models but its evolutive history and functional role in invertebrates, the dominant macrofauna in deep-sea benthic habitats, has hardly been examined (de Mendoza *et al.* 2020). Still, DNA methylation is found in many invertebrates (Keller *et al.* 2016). In CBEs which are characterized by strong environmental gradients, DNA methylation may be particularly important. In Chapter 4's study, I characterized the methylome of siboglinid and alvinellid worm species (see the descriptions of models at pages 34 and 37).

The models

Vesicomyid clam holobionts

The deep-sea vesicomyid clams (Bivalvia: Vesicomyidae: Pliocardiinae) constitute the most diverse group of deep-sea bivalves (Johnson *et al.* 2017). They represent 173 described species that are found worldwide and occupy a vast array of habitats from continental margins to deep trenches (Audzijonyte *et al.* 2012; Krylova and Sahling 2010; MolluscaBase 2019). These clams are typically found partially buried in the sediments of a variety of reducing habitats, from hydrocarbon seeps (Figure 1.7) to hydrothermal vents (Audzijonyte *et al.* 2012). They often form dense beds providing structural habitat for other species (Guillon *et al.* 2017). They all possess sulfur oxidizing symbionts within the epithelial cells of their gills, which provide them with chemosynthetically derived food. Reciprocally, the hosts provide to the symbionts a steady supply of nutrients for their metabolism (including chemosynthesis) and shelter. The symbionts are vertically transmitted to the next generation through the eggs (Cary and Giovannoni 1993; Ikuta *et al.* 2016) but there is evidence this transmission is leaky (Stewart *et al.* 2008, 2009).

Figure 1.7 *Calyptogena sp*. from Sagami Bay, Japan, -1100m. Photo credit: Takao Yoshida, JAMSTEC

The endosymbionts are close relatives to those environmentally acquired by bathymodiolin mussels and to the free living bacteria of the SUP05 group (Anantharaman *et al.* 2013; Roeselers *et al.* 2010). They generally show co-speciation with their hosts and form two phylogenetic groups (Peek *et al.* 1998) which are sometimes referred to as the Gigas (or Clade I) and Ruthia (Clade II) clades (Figure 1.8). Fossil and molecular information suggest the symbionts were before the radiation of the groups about 45 Mya (Peek *et al.* 1997). Since it appears to be more recent than that of other well-studied models such as the aphid/*Buchnera* (~ 200 Mya, Moran *et al.* 1993) and nematode/*Wolbachia* (~100Mya, Ferri *et al.* 2011), the symbiosis between the vesicomyid clams and their bacteria is likely to shed light on the evolutionary processes that follow host capture. (Peek *et al.* 1998)

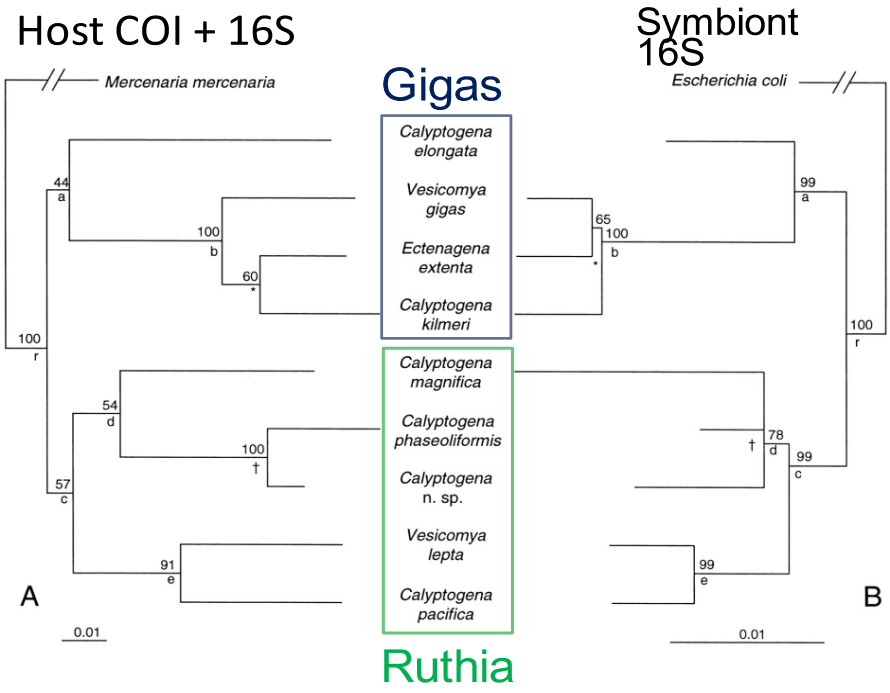

Figure 1.8 The congruent host and symbiont phylogenies in the vesicomyid symbiosis form two monophyletic clades. **A)** Host phylogeny based of COI and 16S rDNA. **B)** Symbiont phylogeny based of 16S rDNA. Adapted from Peek *et al.* (1998).

Siboglinid tubeworms holobionts

The tubeworm species that inhabits HTVs and seeps, also called vestimentiferans, are veritable ecosystem engineers. These polychaetes (phylum Annelida) form dense aggregations that provide

habitat for many other species (Tsurumi and Tunnicliffe 2003). Their success lies in a key adaptation it possesses; they host chemosynthetic bacteria (Figure 1.9, G, Scott *et al.* 1998, Scott *et al.* 1999). The symbionts are hosted in the vacuoles of specialized cells (bacteriocytes) contained within an organ called the trophosome (Figure 1.9).

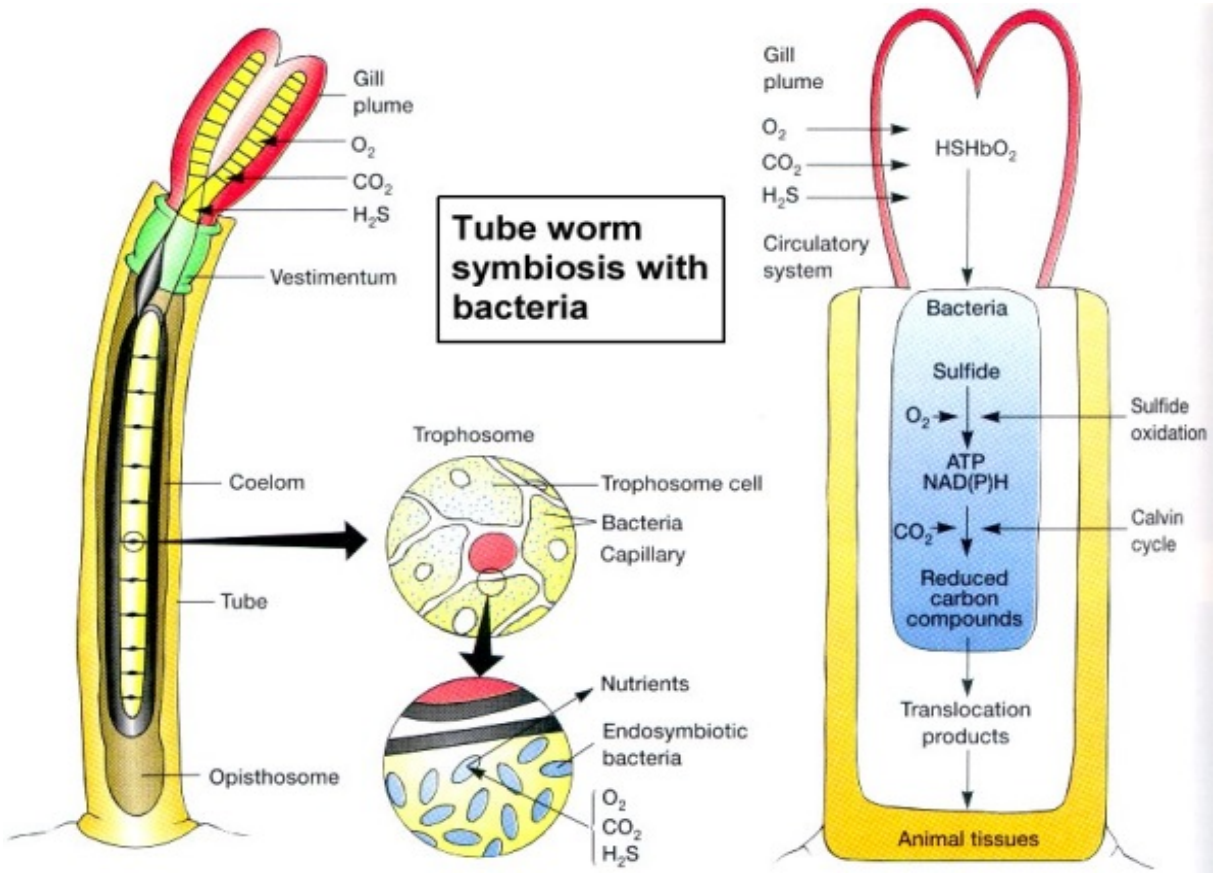

Figure 1.9 Trophosome structure and metabolism of vestimentiferans. From Prescott *et al.*,(2003)

In this mutualistic association, the worm supplies the bacteria with the inorganic compounds necessary for sulfide oxidation and CO_2 fixation: dioxygen, carbon dioxide and hydrogen sulfide. In return, the endosymbionts provide the tubeworm with the organic molecules necessary for its metabolism either by excretion or by being directly digested (Felbeck and Jarchow 1998; Bright *et al.* 2000). The symbionts associated with all hot vent tubeworms from the east Pacific ocean belong to a single species (according average nucleotide identity threshold [Kim *et al.* 2014]) of Gammaproteobacteria coined *Candidatus* Endoriftia persephone (Robidart *et al.* 2008), but symbionts associated with the species of the East Pacific Rise (*Riftia pachyptila* and *Tevnia jerichonana*, *Oasisia alvinae*) and those associated with worms of the Juan de Fuca Ridge (*R.*

piscesae, *Lamellibrachia* sp.) seem to belong to two vicariant populations (Di Meo *et al.* 2000; McMullin *et al.* 2003; Perez and Juniper 2016). The symbiotic bacteria are horizontally transmitted, that is to say, acquired *de novo* from the surrounding environment at each generation (Harmer *et al.* 2008). There is also evidence that the symbionts can escape the tissues of dead hosts and return to a free-living stage (Klose *et al.* 2015). As such, the bacterial symbionts may use the hosts as opportunistic breeding habitat. For the hosts however, the symbiosis is obligate. While the trophosome develops, its digestive system disappears rendering it completely dependent on its endosymbionts for its nutrition.

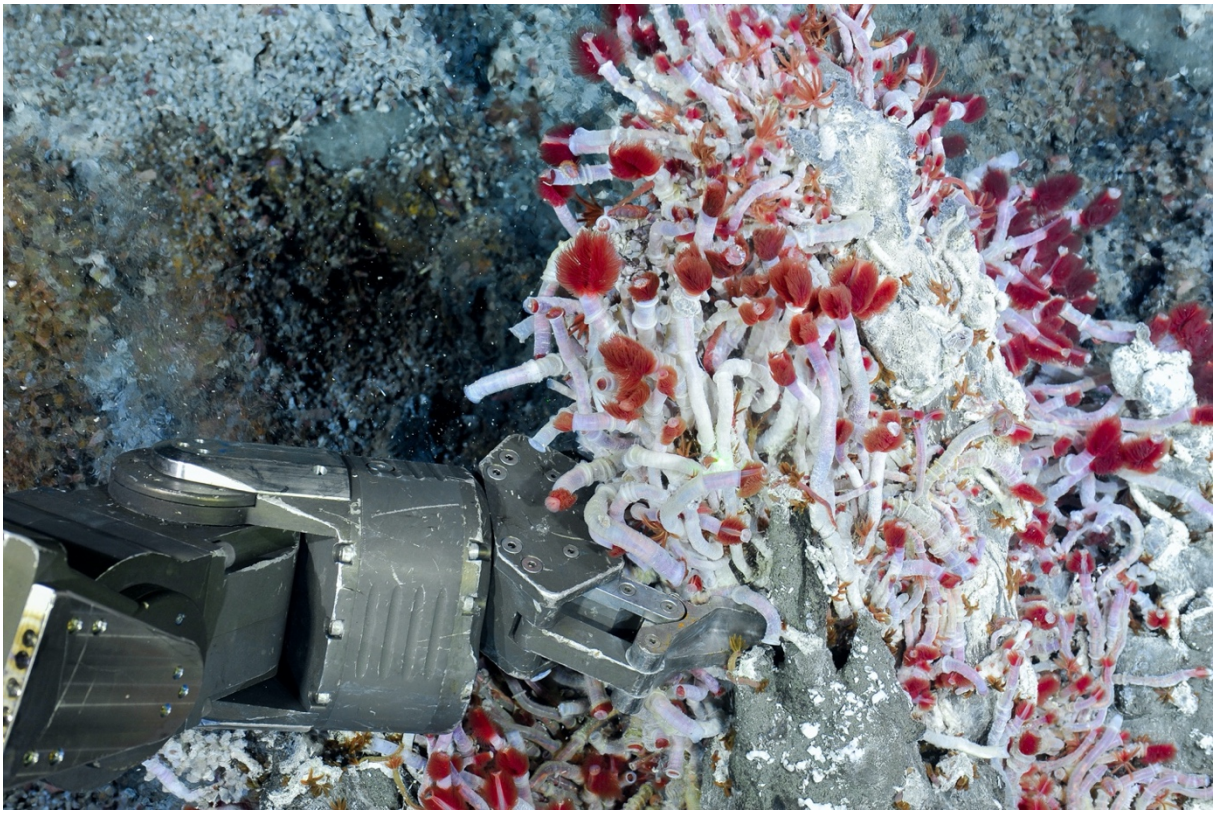

Figure 1.10 Aggregation of the short-fat morphotype of *R. piscesae* at Endeavour vents, Canada, -2100m.
Photo credit: ONC/CSSF-ROPOS.

On the Juan de Fuca ridge the species *R. piscesae* is the most wide-spread macrofauna animal. Its individuals disperse passively during their larval stage (Hilário *et al.* 2005) but as adults, they are sessile and organized in discrete and extremely dense patches; studies report more than 10 000 and up to 250 000 ind.m^{-2} (Sarrazin and Juniper 1999; Urcuyo *et al.* 2007). *R. piscesae* displays a great degree of phenotypic variation (Carney *et al.* 2002). These range from short (up to 20cm) with wide white tubes and well-developed, bright red gills (Short-fat morphotype, Figure 1.10) in

habitats of intense hydrothermal fluid discharge to long (>1m) with rusty-coloured narrow tubes and reduced branchial plumes (Long-skinny morphotype) in habitats located further away from hydrothermal discharge. The phenotypic plasticity of *R. piscesae* would allow its populations to not only colonize the surfaces of sulfide edifices but also the basaltic substrates in their surroundings where the hydrothermal fluids are much more diffused. Worms in these environments have lower growth and reproduction rates (Tunnicliffe *et al.* 2014; Urcuyo *et al.* 2007) but tend to live longer than their congeners in the high flow habitats (Tunnicliffe *et al.* 1990; Urcuyo *et al.* 2007). As such, low-flow habitats could act as refuges for *R. piscesae*'s populations; if a change in hydrothermal activity causes the mortality of the colonies living on the chimneys, the population as a whole survives because it has also colonized these more stable environments (Chevin and Lande 2011).

Many siboglinid species can also be found in hydrocarbon seeps (Bright and Lallier 2010). Less is known about the biology about these worms because they have historically been less studied than their charismatic vent congeners. One such species studied in this thesis, *Paraescarpia echinospica*, is commonly found in methane seeps of the South China Sea in the western Pacific Ocean (Sun *et al.* 2021). Like the long-skinny morphotypes of *R. piscesae*, these worms appear to extract most of their hydrogen sulfide from the substrate they attach to (Urcuyo *et al.* 2003) and have the largest genomes amongst that of the siboglinids sequenced to date (Sun *et al.* 2021). Another recent study has shown that that siboglinids from low-energy habitats have extremely low growth rates and may live for centuries (Durkin *et al.* 2017)! Such unique life history traits would render these species particularly fragile to rapid environmental disturbances. The bacterial chemosymbionts of seep siboglinids belong to a different species than these of their vent congeners and appear to be more genetically heterogeneous (Breusing *et al.* 2020; Di Meo *et al.* 2000; Duperron *et al.* 2009; Kimura *et al.* 2003; Patra *et al.* 2016; Rubin-Blum *et al.* 2014; Zimmermann *et al.* 2014). Perhaps in these worms, microbial diversity allows for a certain degree of metabolic plasticity to face gradual shifts in their environments (Li *et al.* 2018).

Alvinellids

Polychaete worms of the family Alvinellidae are vent endemics and live exclusively on hydrothermal edifices. Eight of the eleven described species live in the eastern Pacific

hydrothermal vents but some species were described in the western Pacific and Indian Ocean (Desbruyères and Laubier 1989; Han *et al.* 2021). Phylogenetic analyses based on proteins (Fontanillas *et al.* 2017) and mitochondrial genes (Perez *et al.* 2022) indicate these worms are grouped in three divergent clades. The worms are thought to have originated in the eastern Pacific and subsequently colonized habitats in the western Pacific and Indian Ocean (Han *et al.* 2021) but to fully describe their biogeography, more extensive surveys of hydrothermal vent systems worldwide would be needed. Stable isotope analyses and behavioral observations have revealed alvinellids occupy high trophic level and have a scavenger/predatory lifestyle (Grelon *et al.* 2006; Lelièvre *et al.* 2017; Levesque *et al.* 2003; Tunnicliffe *et al.* 1993; Van Dover 2002).

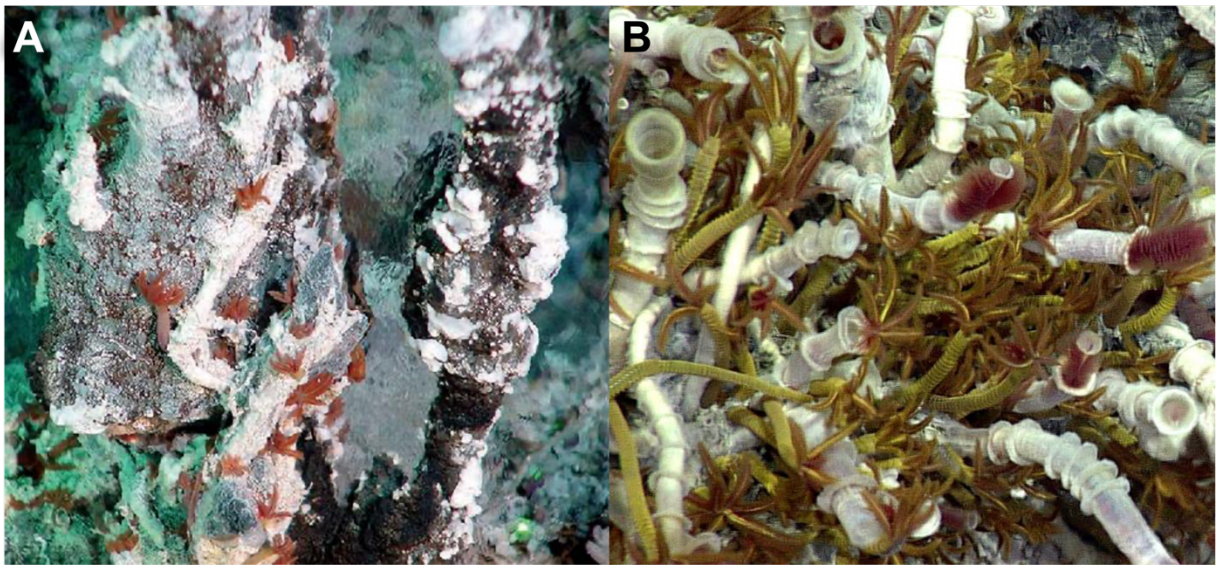

Figure 1.11 *Paralvinella* sp. from the northeastern Pacific vents. A) *P. sulfincola* on sulfide chimney at Endeavour, Canada, -2250m. Note the shimmering of hots hydrothermal fluids; **B)** *P. palmiformis* amongst *R. piscesae* worms at Axial seamount, USA, -1500m. Photo credits: A: S.K. Juniper, B: NSF-OOI/UW.

Remarkably, these worms are amongst the most thermotolerant animals on the planets making them prime models to study thermal plasticity and adaptation. *Paralvinella sulfincola* (Figure 1.11, A) for instance builds mucous houses high on the sulfide accretions (Tunnicliffe *et al.* 1993) and has thermal preference at 40-50ºC when given the choice to position itself on a thermal gradient (Girguis and Lee 2006). This species can however survive a chronic exposure (*i.e.* 12 hours or more) from 2ºC to 65ºC (Dilly *et al.* 2012). On the other hand, *Paralvinella palmiformis* (Figure 1.11, B), a cousin species commonly found at the same sites as *P. sulfincola* preferentially occupy

cooler positions on the sulfide chimneys and do not form tubes (Sarrazin *et al.* 1999). The thermal preference of adults is 30-40ºC (Girguis and Lee 2006) and they are often found within aggregations of *R. piscesae* short-fat morphotypes. Juveniles appear to occupy a lower thermal niche than the adults (McHugh 1989). Experiments have shown *P. palmiformis* have a lower chronic thermal tolerance (0-40ºC) than *P. sulfincola* (Dilly *et al.* 2012). Hyperthermophily appears to be an ancestral trait of the group which was then independently lost in multiple species (Fontanillas *et al.* 2017).

Chapter 2 – Divergent paths in the evolutionary history of maternally-transmitted clam symbionts

Maëva Perez[1†*], Corinna Breusing[2†], Bernard Angers[1], Roxanne A. Beinart[2], Yong-Jin Won[3], and C. Robert Young[4]

[†]Authors contributed equally
[*]Corresponding author
[1]Department of Biological Sciences, Université de Montréal, Montreal, Canada
[2]Graduate School of Oceanography, University of Rhode Island, Narragansett, USA
[3]Division of EcoScience, Ewha Womans University, Seoul, South Korea
[4]National Oceanography Centre, Southampton, UK

Status:

Published in Proceedings of the Royal Society B: Biological Sciences in 2022. doi: 10.1098/rspb.2021.2137

Contributions:

In this article which follows a preliminary study of Dr. Young, I share first co-authorship with Dr. Corinna Breusing. Dr. Breusing investigated the physiological differences between the two clam symbiont clades while I focused on the evolutive processes at play in their genome evolution. On the one hand, Dr. Breusing assembled and annotated the genomes of *Ca.* V. gigas 1 and *Ca.* V. soyoae 1. She also detected functional orthologs (*i.e.* orthogroups) with OrthoFinder, performed the comparative analyses of gene content, the Fubar and MEME analyses on genes of interest, and produced the figures S2.1, S2.4, S2.6 and the tables S2.1, S2.4, S2.6 and S2.10. On the other hand, I assembled the remaining symbiont and mitochondrial genomes, performed the phylogenetic, structural variation, codon bias, RELAX, aBSREL, and Bucky analyses, and produced the rest of the figures. The manuscript was written and edited jointly with input from the other co-authors.

Résumé

La transmission verticale des bactéries endosymbiotiques est accompagnée d'une perte de gènes quasi-irréversible et d'une réduction progressive de la taille de leurs génomes. Bien que les processus évolutifs de réduction génomique aient été bien décris, ils sont beaucoup moins compris dans les systèmes marins où la transmission verticale de symbiotes est rarement observée. L'association entre les palourdes des profondeurs et leurs Gamma-protéobactéries chimiosymbiotiques est un exemple de symbiose verticale dans l'océan. Ici, nous évaluons les contributions de la dérive génétique, de la recombinaison, et de la sélection sur l'évolution du génome de deux clades de symbiontes de vesicomidés en comparant le génome de 15 espèces de symbiotes (1.017–1.586 Mb) à ceux des mitochondries de leurs hôtes, et de bactéries libres qui leur sont phylogénétiquement proches. Nos analyses suggèrent que la dérive est une pression majeure dans l'évolution de ces bactéries. Toutefois la sélection et les évènements de recombinaison interspécifiques semblent être critiques au maintien de l'intégrité fonctionnelle des symbiotes et à la mise en place de patrons de gènes conservés différents entre les clades. Ceux-ci diffèrent notamment par leur métabolisme des sulfures (ou du soufre), la respiration anaérobique, et leur dépendance à la vitamine B12 environnementale, reflétant probablement des adaptations à différents habitats. Globalement, ces résultats contribuent à notre compréhension des processus éco-évolutifs qui affectent l'évolution réductive des génomes dans le cadre des symbioses à transmission verticale.

Abstract

Vertical transmission of bacterial endosymbionts is accompanied by virtually irreversible gene loss that results in a progressive reduction in genome size. While the evolutionary processes of genome reduction have been well described in some terrestrial symbioses, they are less understood in marine systems where vertical transmission is rarely observed. The association between deep-sea vesicomyid clams and chemosynthetic Gammaproteobacteria is one example of maternally inherited symbioses in the ocean. Here, we assessed the contributions of drift, recombination, and selection to genome evolution in two extant vesicomyid symbiont clades by comparing 15 representative symbiont genomes (1.017–1.586 Mb) to those of closely related bacteria and the hosts' mitochondria. Our analyses suggest that drift is a significant force driving genome evolution in vesicomyid symbionts, though selection and inter-specific recombination appear to be critical for maintaining symbiont functional integrity and creating divergent patterns of gene conservation. Notably, the two symbiont clades possess putative functional differences in sulfide physiology, anaerobic respiration, and dependency on environmental vitamin B12, which likely reflect adaptations to different ecological habitats available to each symbiont group. Overall, these results contribute to our understanding of the eco-evolutionary processes shaping reductive genome evolution in vertically transmitted symbioses.

Introduction

Heritable symbioses with intracellular bacteria are observed across the eukaryotic domain of life (Russell 2019). These symbioses have profound consequences for both host and symbiont, by altering sex-ratios in a population, providing nutrients that are otherwise unavailable in the host's habitat, or enhancing resistance to predators and pathogens (Bennett and Moran 2015; Russell and Cavanaugh 2017). Vertical transmission of bacterial lineages from parent to off-spring inevitably leads to reductive genome evolution (RGE) in the symbionts (Vrijenhoek 2010; Bennett and Moran 2015). This process results from successive bottleneck events during transovarial transmission, which decrease the effective population size and genetic diversity of endosymbiont populations (Moran 1996). The genetic homogeneity of vertically transmitted symbionts is further amplified by reduced rates of horizontal gene transfer (*i.e.*, homologous recombination between bacterial lineages), which decrease with higher degrees of host restriction (Russell *et al.* 2020).

Consequently, genetic drift increases relative to selection in these taxa, favoring the accumulation of slightly deleterious mutations (Muller's ratchet) (Muller 1964). The pea aphid/*Buchnera* symbiosis and several other well studied insect/bacteria models support this neutral hypothesis (Wernegreen 2011), whereas other metazoan/microbial symbioses highlight the importance of selection in shaping RGE. For instance, Red Queen/King dynamics are predicted to maintain specificity and the functioning of cyto-nuclear interactions between host and symbiont (Bennett and Moran 2015; Veller *et al.* 2017). Furthermore, symbiont traits that are beneficial for the host might experience increased selective pressures for maintenance of function and/or diversification, while selection may be relaxed on genes that are functionally redundant or no longer necessary for symbiont survival. Thus, differences in gene content among related symbionts can reveal how host-symbiont pairs diverged in their ecological niches over evolutionary time (Hansen and Moran 2014). Ultimately, niche differentiation mediated by differential gene loss has the potential to influence host community structure through habitat partitioning and host evolution through ecological speciation. Despite its importance for ecological and evolutionary processes, there is still a significant gap in our understanding of the selective processes influencing patterns of genome reduction in vertically transmitted bacteria. This is especially true for the heritable endosymbionts of marine organisms, since vertical transmission is less common in aquatic symbioses (Russell 2019).

Relatively strict vertical transmission of bacterial endosymbionts has been observed in deep-sea clams of the family Vesicomyidae (subfamily Pliocardiinae) (Peek *et al.* 1998), providing an opportunity to examine neutral and selective processes shaping RGE in the marine environment. Vesicomyid clams represent the most diverse group of deep-sea bivalves, with 173 described species present in reducing habitats ranging from hydrocarbon seeps on continental margins to hydrothermal vents on mid-ocean ridges (Audzijonyte *et al.* 2012; Johnson *et al.* 2017). All symbiont-bearing taxa are nutritionally dependent on their chemosynthetic gammaproteobacterial partners, which derive chemical energy from the oxidation of reduced sulfur compounds to produce nutrition for their hosts (Childress *et al.* 1991; Newton *et al.* 2008). Symbiont capture was likely a single event that happened before their radiation about 45 Mya (Peek *et al.* 1997; Johnson *et al.* 2017), an acquisition that is much more recent than that of well-studied terrestrial symbioses (~ 100–200 Mya) (Moran *et al.* 1993; Ferri *et al.* 2011). Based on ribosomal sequence data, vesicomyid symbionts are classified into two divergent clades: Clade I (associated with hosts of

the *gigas*-group), and Clade II (associated with all other vesicomyid hosts) (Kuwahara *et al.* 2011; Johnson *et al.* 2017). Topological congruences between host mitochondrial and symbiont phylogenies indicate that symbionts co-evolve with their hosts (Peek *et al.* 1998), although disruptions of these relationships occur through infrequent horizontal transmission events that allow for recombination between bacterial lineages (Stewart *et al.* 2008, 2009; Decker *et al.* 2013; Ozawa *et al.* 2017a; Breusing *et al.* 2019). Previous analyses of one representative symbiont lineage from each clade (*Candidatus* Ruthia *magnifica* for Clade I and *Ca.* Vesicomyosocius okutanii for Clade II) suggest that RGE is ongoing in vesicomyid symbionts and that Clade I is in a more advanced state of genome reduction than Clade II (Kuwahara *et al.* 2008). *Ca.* Ruthia magnifica and *Ca.* Vesicomyosocius okutanii possess intermediate genome sizes (1.16 Mbp *versus* 1.02 Mbp) and levels of AT enrichment (66% *versus* 68%) compared to other host-restricted symbionts, while contrasting levels of gene decay and GC content for 10 housekeeping genes were observed across their respective clades (Shimamura *et al.* 2017).Variations in host affiliation and genome reduction between symbiont clades do not appear to be driven by adaptation to different broad-scale habitat types, as host species of both clades have been found at hydrothermal vents and hydrocarbon seeps and often co-occur at the same locality (Goffredi and Barry 2002; Newton *et al.* 2008; Johnson *et al.* 2017; Cruaud *et al.* 2019). However, limited genetic data suggest that the two symbiont clades differ in physiological characteristics related to nitrate reduction and sulfur metabolism, which may affect microhabitat exploitation (Goffredi and Barry 2002; Newton *et al.* 2008), and could, thus, influence patterns of gene conservation. In fact, niche partitioning has been linked to patterns of gene loss in a variety of marine and freshwater bacteria (Luo *et al.* 2017; Baumgartner *et al.* 2017).

In this study, we assessed the contributions of neutral and selective processes to RGE in vesicomyid symbionts by comparing their genome characteristics to those of outgroup bacterial relatives and the hosts' mitochondria. We tested the hypothesis that genetic drift is a significant force driving RGE in these symbionts and determined to what extent selection has shaped their genetic makeup over evolutionary times.

Materials and methods

Detailed methods are available in the Supplementary Methods in Annex I and online (https://royalsocietypublishing.org/doi/suppl/10.1098/rspb.2021.2137).

Genome analyses

New mitochondrial and symbiont genomes for eleven lineages of vesicomyid clams were assembled and annotated in this study, while genomes for another four species were retrieved from previous publications (Kuwahara *et al.* 2007; Liu *et al.* 2016; Ozawa *et al.* 2017b; Tillich *et al.* 2017; Yang *et al.* 2019; Lee *et al.* 2019; Russell *et al.* 2020; Ip *et al.* 2020) (Figure 2.1, S2.1). Bacterial relatives of the SUP05 clade that comprised lower degrees of host restriction (*Bathymodiolus thermophilus* symbiont [Won *et al.* unpubl.[1]], *Ca.* Thioglobus autotrophicus [Shah and Morris 2015]) were selected as outgroups (Figure 2.1, Table S2.1, S2.2). Similarities and taxonomic affiliations among genomes were assessed with FastANI (Jain *et al.* 2018) and GTDB-Tk (Chaumeil *et al.* 2019).

Comparative genomics

Sequence homology between symbiont genomes was inferred through assessment of positional orthology and orthogroup identification with OrthoFinder (Emms and Kelly 2019) (Table S2.3, S2.4). ProgressiveMauve (Darling *et al.* 2010) and Grimm (Tesler 2002) were used to identify large-scale structural differences among mitochondrial and symbiont genomes based on 13 and 716 conserved protein-coding genes, respectively. Phylogenetic trees were produced from these gene sets in MrBayes (Ronquist *et al.* 2012). Concordance among tree topologies was assessed with Bucky (Larget *et al.* 2010). Pairwise synonymous substitution rates for the mitochondrial and symbiont core genomes were computed following the method in Goldman and Yang (1994).

[1] Later published in Patra *et al.* (2022)

Selection analyses

Branch-specific episodic diversifying selection was identified based on non-recombining core protein-coding genes using aBSREL (Smith *et al.* 2015). Changes in the strength of selection on core protein-coding genes were inferred through quantifications of codon usage bias (Zhang *et al.* 2012) and phylogenetic hypothesis testing with RELAX (Wertheim *et al.* 2015). To strengthen the inferences from these analyses, we performed all tests with additional metagenome-assembled genomes representative of the diversity of free-living and horizontally transmitted SUP05 bacteria. FUBAR (Murrell *et al.* 2013) and MEME (Murrell *et al.* 2012) were used to assess signatures of pervasive and episodic site-specific positive selection in 17 candidate genes that showed marked differences in presence/absence or duplication patterns between the two symbiont clades.

Data availability

Symbiont genomes (CP060680–CP060688, JACRUR000000000, JACRUS000000000) and raw reads are available at the National Center for Biotechnology Information under BioProject PRJNA641445. Genome annotations and metabolic reconstructions can be found on the RAST webserver. Host mitochondrial *COI* sequences and genomes have been deposited in GenBank under accession numbers MT894120–MT894130 and MT947381–MT947391, respectively. Genome alignment files and all scripts used in this study are available at https://github.com/maepz/VesicSymb_Evolution.

Results

Host mitochondrial and symbiont genomes and phylogenies

The genome-wide mitochondrial phylogeny is congruent with host phylogenetic relationships based on multilocus and *COI* sequence data (Johnson *et al.* 2017) (Figure 2.1). Host mitochondrial genomes examined in this study possess identical gene orders and contents, though structural variation is evident among taxa between the $tRNA^{Trp}$/$tRNA^{His\_2}$ and *ND6* loci and in the *COX2* gene (1005–1452 bp, Figure S2.2).

Intra-host symbiont populations were genetically homogeneous with frequency distributions of genetic variants typical of monoclonal populations (Table S2.5, Figure S2.3). Genome size, GC content and number of intact protein-coding genes for the 15 vesicomyid symbiont assemblies ranged from 1.02–1.59 Mb, 31–37% and 896–1455, respectively (Table S2.5), with Clade I having consistently lower values for these genomic characteristics than Clade II. Following initial nomenclature, the symbiont lineages are referred to by the previously erected genera for the two groups, *Candidatus* Vesicomyosocius for Clade I, and *Candidatus* Ruthia for Clade II, accompanied by host species name (Newton *et al.* 2007; Kuwahara *et al.* 2007; Ip *et al.* 2020). This classification at the genus level is consistent with 16S rRNA similarity (< 95%) (Stackebrandt and Goebel 1994), clustering based on average nucleotide identity and alignment fraction (Barco *et al.* 2020) (Figure S2.4), taxonomic assignment (Table S2.6) and genetic isolation between the two symbiont clades (see below).

Mitochondrial and symbiont phylogenies show good concordance for all lineages except one (Figure 2.1). *Ca*. V. diagonalis and *Ca*. V. extenta are nearly identical based on genome size, GC content and gene composition (Figure S2.5, S2.6, Table S2.7), whereas the respective host mitochondrial lineages are divergent. The donor lineage in this recent symbiont replacement appears to be *Archivesica diagonalis*, which co-occurs with *Phreagena extenta* at hydrocarbon seeps in Monterey Canyon. Signatures of elevated substitution rates are evident on the branch leading to Clade I, which is notably longer than the corresponding branch for Clade II, in contrast to patterns in the host phylogeny (Figure 2.1). The symbiont pairs across the Clade I-Clade II bipartition are also significantly more divergent than the others even when controlled for host divergence ($1 < dS_{mito} < 2$; Figure S2.7).

Symbiont genome structure and recombination

The *B. thermophilus* symbiont and *Ca.* T. autotrophicus share about 1 Mbp of their genomes with the clam symbionts, with at least 22 and 3 inversion events being present relative to the *Ca.* R. *magnifica* reference, respectively (Figure S2.8).

Genome structure also differs among the clam symbionts (Figure 2.1, S2.8), with intra-host structural variation being particularly evident within *Ca.* R. phaseoliformis and *Ca.* R. southwardae. Three distinct inversions compared to the *Ca.* R. *magnifica* genome were found in

the genomes of *Ca*. V. okutanii, *Ca*. V. gigas, and the monophyletic group composed of *Ca.* R. fausta, *Ca.* R. pacifica and *Ca.* R. rectimargo. Inversions between the *tufA* and *tufB* paralogs, hotspots for chromosomal inversions (Hughes 2000), seem to have happened multiple times throughout the symbiont phylogeny (Figure 2.1, S2.8).

Bayesian concordance analysis detected substantial recombination within, but not among symbiont clades (Figure 2.1). Relatively little topological concordance was found in Clade II, with 37 different topologies being necessary to fully represent the diversity of conflicting phylogenetic signals compared to only 11 in Clade I (Figure S2.9). Within Clade I, conflict arises from the uncertainty of the position of *Ca.* V. gigas and *Ca.* V. marissinica (Figure 2.1). Within Clade II, only the grouping of *Ca.* R. fausta, *Ca.* R. rectimargo and *Ca.* R. pacifica is supported by all gene trees, while the positions of all other species have low support.

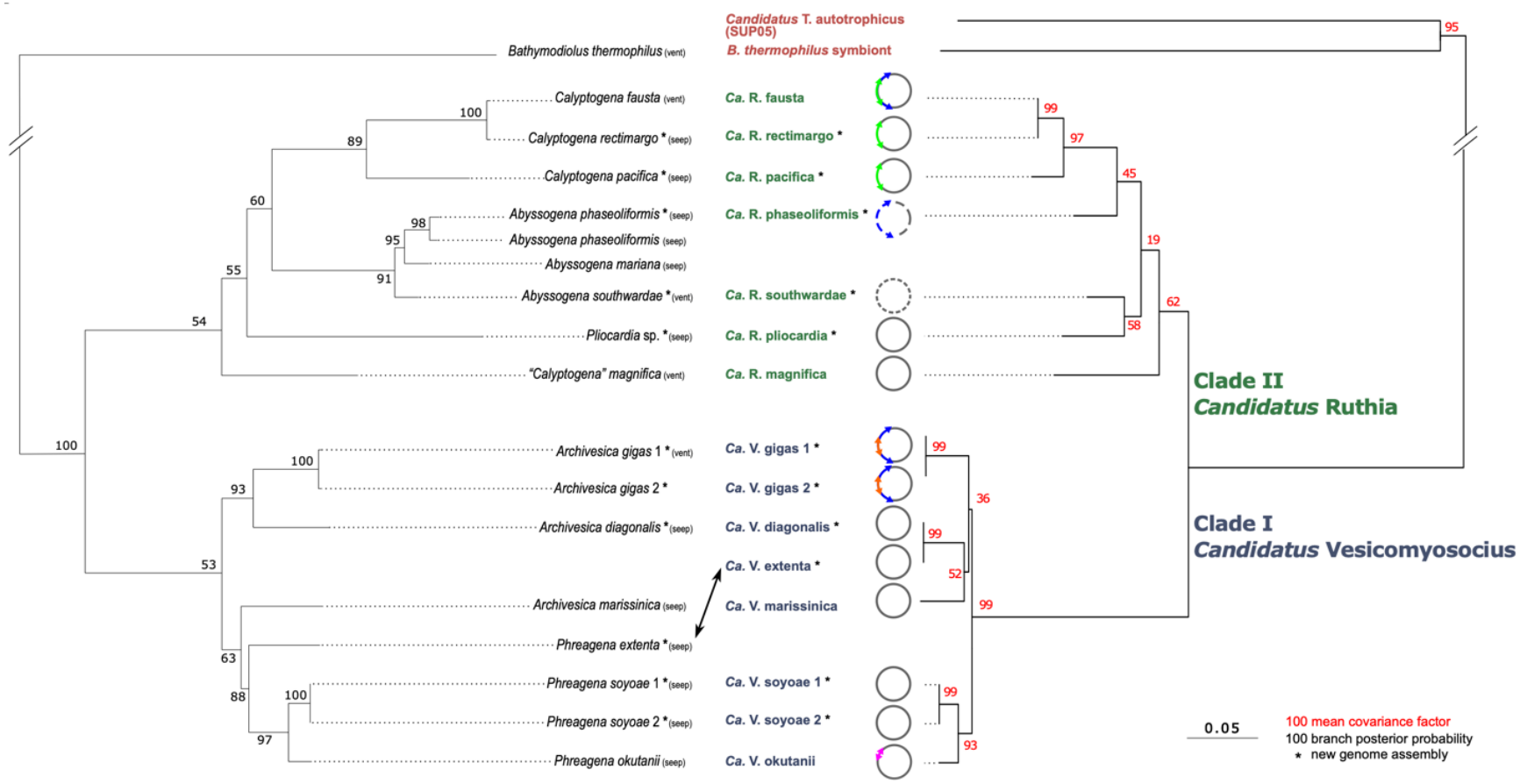

Figure 2.1 Genome-wide host mitochondrial (left) and symbiont (right) trees. Phylogenies represent the Bayesian majority-rule consensus of 2000 independent trees (GTR + G + I model). Left: consensus tree and branch posterior probabilities from the concatenated alignment of 13 core mitochondrial genes. Right: consensus tree from the concatenated alignment of syntenic blocs shared between symbiont (Clade I: blue; Clade II: green) and outgroup (light red) genomes. Genome inversions and assembly fragmentation are displayed at the end of the branches (blue: inversions between TufA/B paralogues; green, orange and magenta: other inversions). Numbers in red are the genome-wide mean covariance factors, which represent the percentage of non-recombining syntenic blocs supporting each split in the phylogeny. *Genomes newly sequenced in this study.

Patterns of gene conservation

The free-living bacterial and environmentally acquired symbiont genomes contained many large (> 5kb) contiguous sections that were absent in the clam symbionts. These genomic islands comprised mostly unannotated genes and mobile elements, but also genes related to heavy metal tolerance and anti-viral defense (in the *B. thermophilus* symbiont) as well as motility and nitrogen metabolism (in *Ca*. T. autotrophicus) (Table S2.4).

The clam symbionts possessed essentially a subset of the genes found in the outgroup lineages, with *Ca.* R. southwardae, *Ca.* R. phaseoliformis and *Ca.* R. pliocardia showing the highest degree of gene conservation. Many genes unique to the vesicomyid symbionts appeared to be pseudogenes resulting from the degeneration of ancestral homologs. Patterns of pseudogenization were relatively prevalent and variable in Clade II (Figure S2.6), while homologous regions within genomes of Clade I were usually characterized by large deletions. Among the Clade II symbionts, gene degeneration was most pronounced in *Ca.* R. *magnifica*, which possessed a conservation pattern closer to that of Clade I (Figure S2.5, S2.6).

Both symbiont clades shared a core genome related to chemoautotrophic metabolism, but showed differences in presence/absence, duplication and degeneration patterns for genes related to a diversity of other metabolic processes (Supplementary Results in Annex I, Figure S2.6, Table S2.4). For instance, the genomes of Clade I and Clade II symbionts encoded different types of methionine synthase. While Clade I contained genes for the cobalamin-dependent homocysteine methyltransferase *metH* and associated genes for cobalamin (precursor) transport and conversion (*btuM*, *btuR/cobA*), Clade II contained the cobalamin-independent version of this enzyme (*metE*) along with its transcriptional activator (*metR*). However, all symbiont lineages lacked pathways for *de novo* cobalamin biosynthesis. Genomes of both symbiont clades also differed in the presence of operons for dissimilatory (*narGHIJ*: Clade I) and assimilatory (*nasA*: Clade II) nitrate reductases, genes for putative nickel transporters (*hupE*) and nickel-dependent enzymes (*gloA*), as well as genes involved in glyoxylate regeneration (*icl*) and transcriptional repression of certain ribonucleotide reductases (*nrdR*) (only in Clade II). Surprisingly, *nasA* was annotated as pseudogene in almost all Clade II lineages and *Ca.* T. autotrophicus. This is possibly a misclassification as functional expression of *nasA* is observed in deep-sea SUP05 populations

(Anantharaman *et al.* 2013). Alternatively, this gene might be in an early stage of pseudogenization as all variants encompassed over 74% of the intact protein length. An operon encoding cysteine dioxygenase type I (*cdo*) and an aspartate aminotransferase superfamily protein, which has homology to cysteine sulfinic acid decarboxylase (*csad*) from *B. azoricus* (GenBank: SEH86284), was exclusively found in Clade I. Unlike their Clade II congeners, the genomes of almost all Clade I symbionts were characterized by a duplication of the sulfide:quinone oxidoreductase type I gene (*sqrI*).

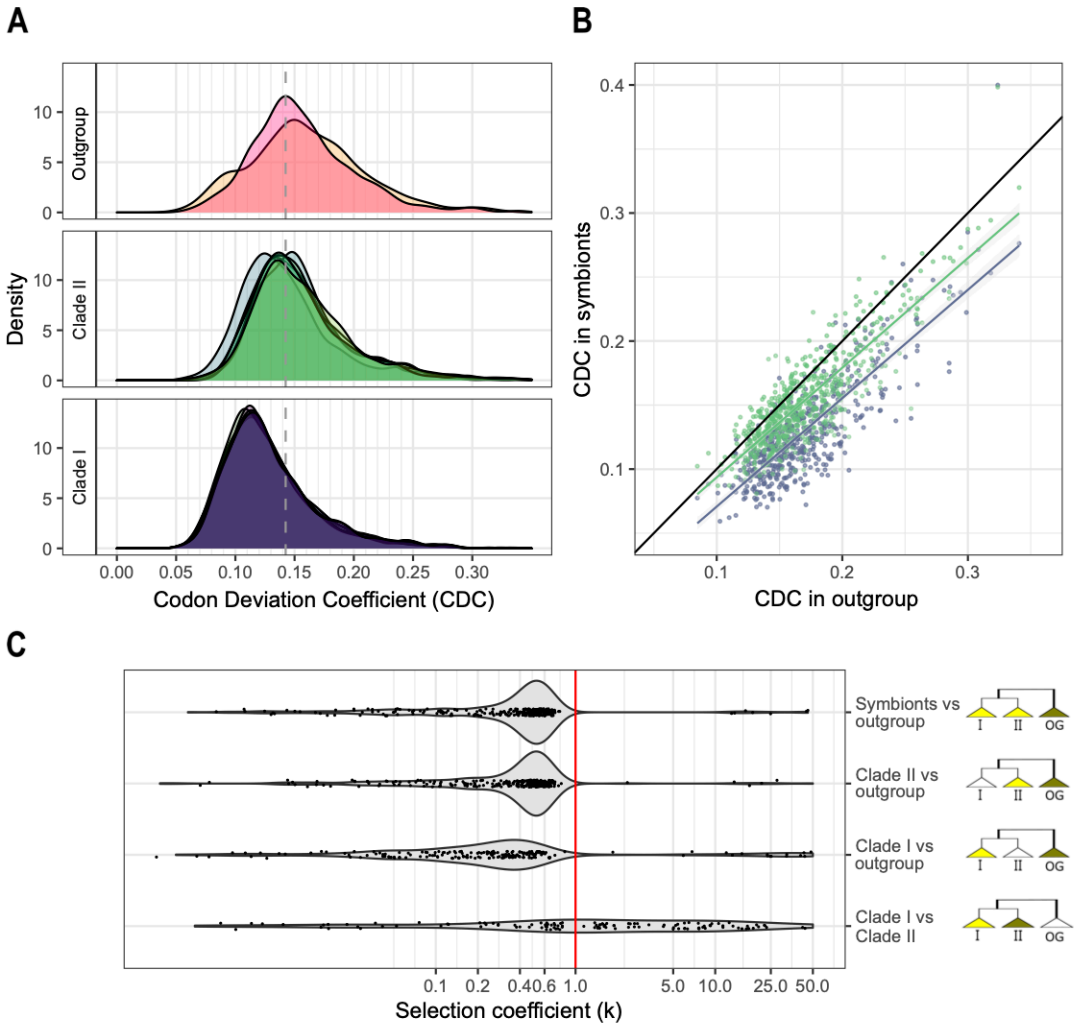

Figure 2.2 Codon usage bias in clam symbionts and outgroup bacterial relatives. A) Codon deviation coefficient (CDC) spectra for each genome within the outgroup; yellow: *B. thermophilus* symbiont; red: *Ca.* T. autotrophicus. **B)** Correlation between the average CDC of the outgroup, Clade I (blue) and Clade II (green) based on 555 core genes. CDC values vary from 0 (no bias) to 1 (maximum bias). **C)** Log-scaled selection parameter (k) spectra of core genes for which a significant change in selection was detected. CDC values were significantly lower in Clade I than Clade II, and CDC and k values were significantly lower in both symbiont clades than the outgroup (paired Wilcoxon–Mann–Whitney test p-value < 0.01).

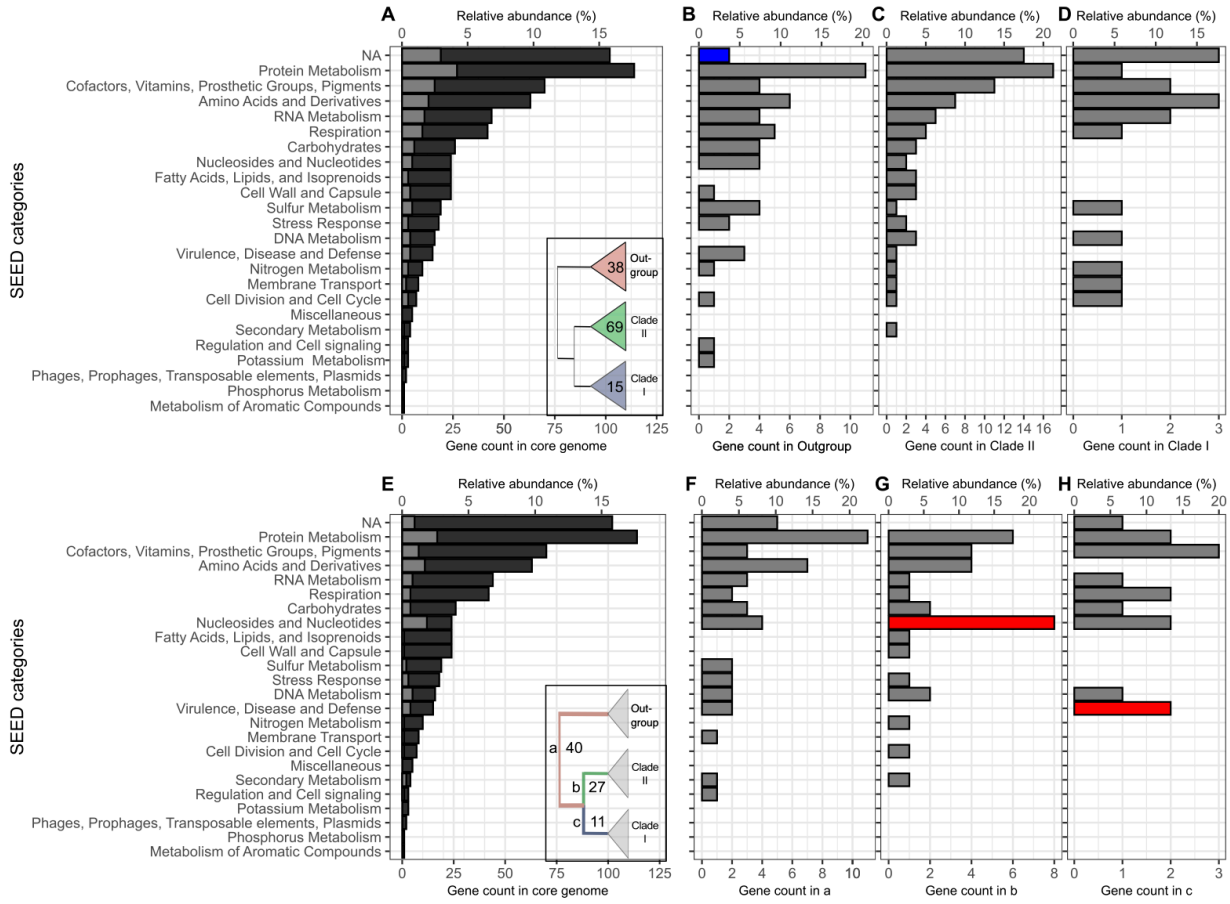

Figure 2.3 SEED category distribution of core genes under episodic diversifying selection within phylogenetic clades (A-D), and on partitioning branches (D-H). **A)** Distribution of all non-recombining core genes (dark grey, 555 loci) and loci under selection within the outgroup, Clade I and Clade II (light grey, 114 loci). **B)** Genes under selection within the outgroup. **C)** Genes under selection within Clade II. **D)** Genes under selection within Clade I. **E)** Distribution of all non-recombining core genes (dark grey, 555 loci) and loci under selection on all partitioning branches (light grey, 71 loci). **F)** Genes under selection on branch a. **G)** Genes under selection on branch b. **H)** Genes under selection on branch c. Note that genes may be present in multiple functional categories, clades and/or branches. Insets show number of loci selected within clades or along branches. SEED categories significantly overrepresented (in red) and underrepresented (in blue) compared to the core genome are highlighted. NA, no functional annotation.

Genome-wide patterns of relaxed and intensified selection

Both symbiont clades showed reduced codon usage bias (Figure 2.2A, B) and dN/dS rate-class extremes (Figure 2.2C) compared to the outgroup (Figure 2.2C), indicating a genome-wide decline in the efficacy of natural selection, *i.e.*, a reduction in selective constraint (Table S2.7, S2.8). Codon usage bias and selection intensity analyses including incomplete genomes from free-living SUP05 bacteria and closely related horizontally transmitted symbionts associated with deep-sea mussels

(*Bathymodiolus* sp.) and sponges (*Suberites* sp.) further support the inference of drift-driven RGE in the vesicomyid symbionts (Figure S2.10, Supplementary Results). Relaxation was comparable to that observed in insect endosymbionts and appeared to be exacerbated in Clade I (Figure 2.2B, C) (Wertheim *et al.* 2015). Genes exhibiting intensified and relaxed selection in the clam symbionts represented a multitude of metabolic categories, although some functions were predominantly affected by directional shifts in selection regimes (Figure S2.11). Genes under relaxed selection were mostly involved in protein, amino acid and nucleoside/nucleotide metabolism, cell division and cell cycle, whereas genes under intensified selection were largely associated with respiration and sulfur metabolism.

Patterns of positive selection in core and clade-specific genes

114 protein-coding core genes exhibited evidence for episodic diversifying selection along branches in the phylogeny (Table S2.9). Selection appears to be distributed throughout the evolutionary history of the symbionts (Figure 2.3), acting mostly on the outgroup branches and the branches discriminating the outgroup, Clade I, and Clade II. Eighty-five percent of loci that showed signs of selection were classified into SEED categories (Figure 2.3). These loci were equally represented among cellular functions of the core genome except for a few categories (*e.g.*, nucleotide synthesis and defense) along the branches separating the two symbiont clades (Figure 2.3).

Apart from *gloA, narI*, and *narJ*, all investigated metabolic genes that were differentially preserved between clades showed evidence of pervasive or episodic site-specific diversifying selection that affected structural or functional regions in the encoded proteins (Table S2.10). Pervasive positive selection was observed at 1–3 sites across the entire phylogeny in ten of the 17 genes tested: *btuM*, *btuR*, *csad*, *hupE*, *icl*, *metR*, *narG*, *narH, nasA*, *sqrI*. In addition, episodic positive selection was detected at 1–7 sites along a proportion of branches in all tested genes except for *btuR*, *gloA, narI*, and *narJ*. In the case of *cdo*, *csad* and *nasA*, these episodes of site-specific selection seemed to have mostly occurred in the ancestral lineages as no evidence for selection was found along the extant symbiont branches (Table S2.10).

Discussion

RGE is ongoing and driven by neutral processes

Current insights into the evolutionary processes shaping RGE in maternally inherited symbionts stem mostly from well-studied terrestrial insect-bacteria associations, where genetic drift has been shown to be the dominant force driving patterns of endosymbiont gene loss (Wernegreen 2011; Martínez-Cano *et al.* 2015). Our analyses of 15 vesicomyid symbiont genomes suggest that neutral processes play an equally important role in marine vertically transmitted symbioses. As in other models of recently acquired bacteria (Andersson and Andersson 1999; Burke and Moran 2011), gene content differed substantially between vesicomyid symbiont genomes, indicating that the different lineages are independently losing genes. The presence of structural variation and varying degrees of gene degeneration imply that vesicomyid symbionts have not yet reached a stable streamlined state compared to many insect endosymbionts (Tamas *et al.* 2002), as suggested previously (Kuwahara *et al.* 2008). All clam symbionts exhibited a reduced GC%, decrease in codon usage bias, and a genome-wide trend of relaxation in selective pressures relative to the outgroup. Overall, these observations support the nearly neutral theory of RGE, driven by a reduction of effective population size in these taxa (Moran 1996).

In agreement with previous findings (Stewart *et al.* 2008, 2009; Decker *et al.* 2013; Ozawa *et al.* 2017a), we detected no recombination between Clade I and Clade II, even though some of the host taxa co-occur (Goffredi and Barry 2002; Decker *et al.* 2013). This implies that there is enough molecular and ecological divergence between the two clades for clonal interference and/or strong host-symbiont epistatic interactions to constrain symbiont exchange. Clade I and Clade II are also discriminated based on measures of genomic relatedness and functional genomic traits, all of which support a classification of these symbionts into two distinct bacterial genera, *Ca.* Vesicomyosocius and *Ca.* Ruthia (Ozawa *et al.* 2017a; Baumgartner *et al.* 2017).

RGE is exacerbated in non-recombining symbionts

Symbionts of Clade I appear to be in a more advanced state of RGE than those of Clade II, as their genomes are smaller and lower in GC%, possess fewer genes and pseudogenes, exhibit less codon usage bias and are in general more homogeneous. Patterns of gene conservation suggest that much

of the loss in this group happened after its speciation but before its radiation, a period of roughly 20 Myrs (Peek *et al.* 1998; Johnson *et al.* 2017). Fossil informed phylogenetic inference places the radiation of vesicomyid clams at 47 ± 2 Mya and that of Clade I at 22 ± 5 Mya (Johnson *et al.* 2017), resulting in a rate of genome reduction of about 20Kbp/Myrs in the Clade I symbionts. Together with increased substitution rates on its diverging branch, these results imply that the ancestral Clade I lineage experienced an acute episodic acceleration of RGE. Based on genome-wide levels of topological disagreement, inter-specific homologous recombination is widespread among symbionts of Clade II but almost absent in Clade I. A reduction of the rate of infection by environmental symbionts and/or drift-driven loss of the recombination machinery (Kuwahara *et al.* 2011) may have strongly reduced the rate of genetic exchange across Clade I symbionts, thereby setting this genus on a divergent evolutionary path.

Recombination can alter rates of evolution due to Hill-Robertson interference (Hill and Robertson 1966) by randomizing the associations between mutations that otherwise would be in linkage disequilibrium. In small populations, deleterious alleles fix through drift, reducing the mean fitness of the population (Muller 1964). Additionally, background selection against deleterious alleles can purge linked beneficial alleles from the population (Charlesworth 1994), whereas hitchhiking effects can retain linked deleterious mutations (Gillespie 2000). Low rates of recombination typically increase Hill-Robertson effects (Hill and Robertson 1966), whereas absence of recombination can reduce the rate of adaptation through clonal interference (Gerrish and Lenski 1998).

Strong linkage disequilibrium forces whole genomes to sweep in populations that lack capabilities for genetic exchange. Hence, loss of recombination should favor symbiont replacement in cases where the divergence between native and foreign symbionts is low enough to avoid host-symbiont incompatibilities. We find multiple examples of symbiont replacement among lineages of Clade I. For instance, individual *P. extenta* clams have acquired the symbionts of the sympatric species *A. diagonalis.* Likewise, some *A. gigas* populations have been found to carry the symbionts of the host species *P. soyoae* (Stewart *et al.* 2008). Symbiont replacement occurs in several vertically transmitted symbioses (Sudakaran *et al.* 2017) and is speculated to constitute a mechanism for escaping the evolutionary rabbit hole caused by Muller's ratchet (Bennett and Moran 2015).

Despite the lack of recombination machinery in Clade I, *Ca.* V. gigas and *Ca.* V. marissinica showed signs of genetic exchange. Perhaps recombination in these species is mediated via symbiont-derived host-encoded proteins. Evidence for transfer of ancestral symbiont gene homologs to the host nuclear genome was recently found in *A. marissinica* clams (Ip *et al.* 2020). Overall, these observations support a crucial role of recombination in maintaining symbiont genome integrity (Russell *et al.* 2020) and moderating the ecological consequences of increased clonality.

Selective processes might be tied to genetic and environmental contexts

Although our data indicate that genetic drift is a major force mediating RGE in vesicomyid symbionts, significant fractions of the symbiont genomes are affected by natural selection. Selection might act on genes involved in host-symbiont interactions, as these genes are expected to experience reciprocal adaptations through speciation and niche exploitation. Diversifying selection affecting genes that play a role in host-symbiont interactions, such as lipopolysaccharides and peptidoglycans, is observed in several terrestrial obligate and facultative endosymbionts (Dale and Moran 2006; Brownlie *et al.* 2007). Surprisingly, our data do not confirm these predictions and instead suggest a pervasive pattern of positive selection affecting a broad range of cellular functions. This could indicate that the accumulation of slightly deleterious mutations in the symbiont genomes enhances selective pressures for compensatory mutations, as described for cellular organelles and other bacterial endosymbionts (Lambert and Moran 1998; Howe and Denver 2008). Alternatively, these patterns might reflect ongoing adaptations to the host intracellular environment (Martínez-Cano *et al.* 2015).

Strong functional contrasts in gene loss between outgroup and clam symbionts as well as between both symbiont clades further suggest a role of niche differentiation in shaping symbiont genome composition through RGE, which has consequences for ecological processes, like habitat use, and evolutionary processes, like host speciation. Genes enabling bacteria to face the challenges of a free-living environment, such as metal detoxification, anti-viral defense and inter-species competition, were not conserved in the clam symbionts, while both clades differed in retention of genes affecting diverse metabolic traits, including the dependency on enzyme cofactors, anaerobic respiration, and sulfur physiology.

Symbiont clades show putative differences in physiology and ecological niche

Clade I and II symbionts encoded different, convergently evolved types of methionine synthase (González *et al.* 1996), which vary in their requirement for vitamin B12. Measurements of cobalamin in the deep-sea are challenging, and we, therefore, do not currently have informative data on vitamin B12 concentrations experienced by the clams. However, as both clades appeared unable to synthesize cobalamin *de novo*, these findings indicate that the environmental availability of vitamin B12 has the potential to be an important factor influencing the distribution of these taxa. Cobalamin independence in Clade II may offer a selective advantage by allowing these symbioses to exploit habitats that would otherwise be inaccessible. By contrast, the requirement for exogeneous vitamin B12 (or its derivatives) might limit the range of (micro)habitats Clade I-based associations can colonize, unless cobalamin is acquired from a secondary symbiont. Despite this potential cost, the retention of a cobalamin-dependent methionine synthase in Clade I likely provides an evolutionary benefit, given that MetH has a fifty-fold higher catalytic rate constant than MetE and thus enables faster growth (González *et al.* 1996).

Comparative measurements of vesicomyid growth rates suggest that species hosting Clade I symbionts typically grow faster than species with symbionts of Clade II, despite a less efficient sulfur physiology (Barry and Kochevar 1998). Since growth is influenced by a variety of factors (Barry and Kochevar 1998), it is possible that the enzymatic differences in methionine biosynthesis among symbiont clades contribute to an accelerated anabolism in Clade I-based associations, although this remains to be experimentally tested. The preservation of cobalamin-dependent enzymes as a result of conferred physiological advantages appears to be common across the eubacterial domain (Shelton *et al.* 2019). About 86% of bacterial lineages seem to have at least one cobalamin-dependent enzyme despite the existence of a cobalamin-independent alternative, and many of these lineages rely on vitamin B12 production from other microbes in their environment (Shelton *et al.* 2019). The importance of vitamin B12 for the biology of the two symbiont groups is also evident in the fact that only Clade II symbionts encode a transcriptional repressor (NrdR) for the ribonucleotide reductase NrdAB, a key enzyme that controls the synthesis of DNA (Torrents 2014). In Clade I, expression of NrdAB is probably regulated by cobalamin, which has been shown to repress NrdAB transcription through riboswitches (Borovok *et al.* 2006). There is evidence that the two symbiont clades differ in their requirements for other enzyme cofactors, such as nickel.

The genomes of Clade II symbionts encoded a specific transporter for nickel uptake, and most of these lineages contained at least one confirmed nickel-dependent enzyme (glyoxalase I (Boer *et al.* 2014) all of which were absent in Clade I.

Our data extend previous findings that Clade I and Clade II symbionts show differences in encoded gene clusters for nitrate reduction (Newton *et al.* 2008), confirming that these patterns are truly clade-specific characteristics. The use of NarGHIJ for nitrate reduction in Clade I likely enables these symbioses to inhabit hypoxic environments (Newton *et al.* 2008), since the use of nitrate as an electron acceptor would reduce the symbiont's requirement for oxygen and, consequently, avoid competition with the host. These assumptions agree with field observations showing that clam species hosting Clade I symbionts typically occupy microhabitats with higher levels of hydrogen sulfide (and, thus, presumably lower oxygen) than those hosting Clade II symbionts (Goffredi and Barry 2002). Niche partitioning based on environmental sulfide levels has also been suggested by physiological comparisons of *P. soyoae* and *C. pacifica*, which imply that *P. soyoae* symbionts require higher H_2S concentrations for chemosynthesis (Goffredi and Barry 2002). This could be due to a less efficient sulfide metabolism in *Ca.* V. soyoae (Clade I) resulting from an increased load of deleterious mutations in accordance with its more advance state of genome reduction compared to *Ca.* R. pacifica (Clade II). The presence of two tandem copies of *sqrI* displaying evidence of concerted evolution suggests increasing gene dosage as compensating mechanism in Clade I. Our data indicate that there might be additional adaptations (or ancestral restrictions) to contrasting sulfide environments between symbiont clades. Only Clade I symbionts encode genes for CDO and putatively CSAD, which are key enzymes in the biosynthesis of taurine/hypotaurine. These non-proteinogenic amino acids are important for sulfide detoxification in many symbiont-bearing invertebrates that inhabit sulfidic environments (Brand *et al.* 2007). Since Clade I-based associations appear to be prevalent in high-sulfide habitats, it is possible that these symbionts directly or indirectly contribute to H_2S tolerance of their hosts. Another function of CDO and CSAD could involve replenishment of metabolic intermediates. Vesicomyid symbionts do not possess a complete TCA cycle and must recycle succinate through other means (Newton *et al.* 2008). In Clade II symbionts succinate regeneration occurs via the glyoxylate shunt, while the mechanism in Clade I symbionts is unclear (Newton *et al.* 2008). Perhaps taurine is further metabolized via taurine dioxygenase, which would generate succinate as end product. If this

pathway can be confirmed through physiological experiments, autonomous recycling of succinate by the symbiont could make important contributions to the holobiont's carbon budget.

Conclusions

Our analyses show that patterns of genome reduction in vesicomyid symbionts are mostly shaped by genetic drift and that factors affecting symbiont clonality strongly influence the rate of RGE. A first period of intensified RGE likely occurred during the shift from horizontal to vertical transmission mode in the early to mid-Eocene, and was followed by a second period in Clade I after transition to complete host restriction in the late Eocene/early Oligocene. The pervasive nature of episodic diversifying selection across functional traits in the vesicomyid symbiont genomes, however, suggests that neutral evolutionary processes (drift, mutation, and recombination) are not the sole drivers of molecular evolution in these taxa. Differential patterns of gene loss between Clade I and Clade II reiterate that RGE does not follow a universal trajectory, but is a reflection of the eco-evolutionary context of the respective host-symbiont association. Convergent gene loss and pseudogenization imply common evolutionary pressures for some genes, whereas selection and lineage-specific gene retention imply niche-specific adaptation in others. Future studies linking environmental data with symbiont genomic information at the population level will help to decipher the contributions of host eco-physiology, symbiont fitness, cytonuclear incompatibilities, and rates of lateral transfer to symbiont evolution.

Acknowledgements

We thank the captains, crews and submersible pilots of the R/Vs *Atlantis*, *Western Flyer* and *Point Lobos* for supporting the sample collections, and Bob Vrijenhoek for providing clam specimens for this study. We further thank N. Pratt and A. Baylay for their contributions to library preparation and sequencing at the National Oceanography Centre Genomics Facility. This work was supported by grants of the David and Lucile Packard Foundation (to MBARI), the UK Natural Environment Research Council (grant number NE/N006496/1 to C.R.Y.), the US National Science Foundation (grant number OCE-1736932 to R.A.B.), the German Research Foundation (BR 5488/1-1 to C.B.), the Natural Science and Engineering Research Council of Canada (grant number 238600 to B.A., Alexander Graham Bell graduate scholarship and Michael Smith Foreign Study Supplements to

M.P.) and National Capability funding to the National Oceanography Centre (grant number NE/R015953/1). Sequencing of the *Bathymodiolus thermophilus* symbiont genome was funded by the Korean Ministry of Oceans and Fisheries (grant number 20170411 to Y.J.W.). Bioinformatic analyses were in part performed on ComputeCanada HPC clusters.

Chapter 3 – Shining light on a deep-sea bacterial symbiont population structure with CRISPR

Maëva Perez[1*], Bernard Angers[1], C. Robert Young[2], Kim Juniper[3]

[*]Corresponding author
[1]Université de Montréal, Quebec , Canada
[2]National Oceanography Center, Southampton, UK
[3]University of Victoria, British Columbia , Canada

Status:

Published in Microbial Genomics in 2021. doi: 10.1099/mgen.0.000625

Vulgarized presentation of the article (in French) available at :
https://biologiecsudem.weebly.com/maeva-perez.html

Contributions:

This article follows the finding made in Perez and Juniper (2016) that the symbiont of *R. piscesae* possess the CRISPR-cas immunity. In this study, I wanted to explore the possibility of using one of the two discovered CRISPR arrays as a genetic marker for discriminating between different symbiont strains. I (1) conceived the genetic sampling design (the decontamination and symbiont fraction enrichment protocol, the use of a combination of CRISPR and four loci amplicons specifically selected for their polymorphism), (2) designed and tested the gene primers, (3) conducted the worms' dissections, DNA extraction, loci amplification, and the initial steps of the multiplexed sequencing library preparation (barcoding attachment, DNA dosage in the multiplexed pools), (4) performed all upstream bioinformatic analyses (quality control, demultiplexing, detection of single-nucleotide variants, characterisation of CRISPR arrays, and strain assignment), (5) wrote the multithreaded python implementation of the Kupczok and Bollback (2013) algorithm, (6) conducted the downstream analyses and statistical testing (phylogenetic tree reconstruction, AMOVAs), (7) created the figures and (8) wrote the manuscript with input and edits from my co-authors.

Kim Juniper provided funding through the Canadian Healthy Oceans Network (CHONe) and the Canadian Natural Science and Engineering Research council (NSERC), designed the geographical sampling structure, and advised on the manuscript. C. Robert Young advised on the manuscript and suggested we apply the Kupczok and Bollback (2013) method to our data. Bernard Angers advised on the manuscript, contributed in some of the preliminary analyses of the data and advised on laboratory work prior sequencing.

Résumé

La majorité des espèces clé de voûte (ou fondatrices) des écosystèmes marins profonds sont dépendantes de bactéries symbiotiques chimiosynthétiques acquises dans leur environnement local. Ainsi, il est nécessaire de bien comprendre la distribution biogéographique locale et régionale de ces symbiotes pour mieux prédire le niveau d'isolement et de résilience de leurs hôtes (et donc de communautés entières). Cependant, de telles évaluations sont difficiles car elles nécessitent de mesurer la diversité génétique des bactéries à de très fines résolutions. Le gène CRISPR (pour *Clustered Regularly Interspaced Short Palindromic Repeats*; courtes répétitions palindromiques groupées et régulièrement espacées) qui a été découvert récemment, semble être un marqueur idéal pour aborder ce problème. Ces séquences présentes chez près de la moitié des espèces bactériennes contiennent leur mémoire immunitaire et pourraient ainsi permettre de discriminer les souches ayant eu différents nombre d'infections par les virus. Dans cette étude, nous avons évalué le potentiel du CRISPR comme un marqueur évolutif hypervariable appliqué à une population de bactéries naturelle non-cultivable. Nous avons caractérisé la structure régionale des symbiotes de l'espèce *Candidatus* Endoriftia persephone échantillonnée sur la dorsale Juan de Fuca, à l'aide de séquençage à haut débit du gène CRISPR et d'autres marqueurs typiquement utilisés pour génotyper les bactéries. Pour faire la part entre les facteurs dus à la connectivité et ceux dus aux conditions physico-chimiques de l'environnement, des populations mixtes de symbiotes ont été échantillonnées dans des hôtes provenant d'habitats variés et de différents sites hydrothermaux. Nos résultats ont montré que la diversité génétique révélée par l'utilisation du marqueur CRISPR est bien plus grande que celles des autres marqueurs génétiques. Plusieurs résultats concordants prouvent que cette diversité reflète l'hétérogénéité génétique réelle des populations de bactéries. Enfin, autant avec CRISPR que les autres marqueurs, nous avons montré que les populations de symbiotes sont partitionnées à l'échelle locale et fortement différenciées entre sites isolés par les courants océaniques de fond. Cette étude montre le fort potentiel du gène CRISPR pour résoudre la structure génétique de bactéries non-cultivables et présente un fort argument pour l'inclusion des microbes dans les plans de conservation des écosystèmes hydrothermaux face à leur exploitation minière imminente.

Abstract

Many foundation species in chemosynthesis-based ecosystems rely on environmentally-acquired symbiotic bacteria for their survival. Hence, understanding the biogeographic distributions of these symbionts at regional scales is key to understanding patterns of connectivity and predicting resilience of their host populations (and thus whole communities). However, such assessments are challenging because they necessitate measuring bacterial genetic diversity at fine resolutions. For this purpose, the recently discovered Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) constitutes a promising new genetic marker. These DNA sequences harbored by about half of bacteria hold their viral immune memory, and as such, might allow discrimination of different lineages or strains of otherwise indistinguishable bacteria. In this study, we assessed the potential of CRISPR as a hypervariable phylogenetic marker in the context of a population genetic study of an uncultured bacterial species. We used high-throughput CRISPR-based typing along with Multi-Locus Sequence Analysis (MLSA) to characterize the regional population structure of the obligate but environmentally acquired symbiont species *Candidatus* Endoriftia persephone on the Juan de Fuca Ridge. Mixed symbiont populations of *Ca.* Endoriftia persephone were sampled across individual *Ridgeia piscesae* hosts from contrasting habitats in order to determine if environmental conditions rather than barriers to connectivity are more important drivers of symbiont diversity. We showed that CRISPR revealed a much higher symbiont genetic diversity than the other housekeeping genes. Several lines of evidence imply this diversity is indicative of environmental strains. Finally, we found with both CRISPR and gene markers that local symbiont populations are strongly differentiated across sites known to be isolated by deep-sea circulation patterns. This research showed the high power of CRISPR to resolve the genetic structure of uncultured bacterial populations and represents a step towards making keystone microbial species an integral part of conservation policies for upcoming mining operations on the seafloor.

Introduction

Marine bacteria and archaea perform vital marine ecosystem functions including primary productivity at the sunlit surface, remineralization and storage of carbon in the water column and the ocean's interior through the biological carbon pump. They are also primary producers within the ocean's dark interior, inhabiting environments such as hydrothermal vents and hydrocarbon

seeps, where they utilize geochemical energy rather than sunlight to fix carbon. Given their fundamental roles in marine ecosystems, understanding the processes that govern microbial biogeographic distributions, community assembly, and ecosystem function is a primary pursuit of marine microbial ecology. To achieve this goal, we need to understand how microbial distributions are determined by the interaction of physical and biological factors. Indeed, the paradigm formulated by Baas-Becking (1934) that "everything is everywhere, but the environment selects" is increasingly being challenged in marine systems (Ward *et al.* 2020). Together with the collection of environmental data, the use of multiple hypervariable gene markers has provided a growing body of evidence suggesting that dispersal of microbes in the oceans is limited (Ho *et al.* 2017; Sul *et al.* 2013), and that geographical isolation even affects bacteria at the local scale (Cho and Tiedje 2000; Papke and Ward 2004; Whitaker *et al.* 2003).

Assessments of the structure of bacterial populations are limited by the resolution of genetic markers. Gene markers with high conservation and low diversity lack the resolution to reveal fine scale population structure. Previous studies have shown the very conserved 16S rRNA gene typically used to assess bacterial diversity is not variable enough to detect genetic diversity at the population level (Cho and Tiedje 2000). The Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR), on the other hand, might provide the high-definition needed. The CRISPR locus is the adaptive immune system of prokaryotes (Westra *et al.* 2016). It is composed of the Cas operon and the CRISPR array. The Cas operon contains genes responsible for editing the CRISPR array as well as genes with anti-viral functions (Sorek *et al.* 2008). The CRISPR array consists of short sequences (CRISPR spacers, ~40 bp) that are complements to sequences in phage nucleic acids. These spacers are separated by short sequences of palindromic repeats (CRISPR repeats) that are species specific. The spacers constitute a historical record of the viral encounters of a particular lineage, because they always accumulate at the 5' end of the CRISPR array. Furthermore, because each protospacer is randomly sampled from the virus genome, independent infections by the same virus phylotype would practically never result in the insertion of the exact same spacer sequence in two bacterial lineages (Barrangou *et al.* 2007; Held *et al.* 2010). Various CRISPR loci are already in use for tracking the micro-evolution of pathogenic bacteria (Bachmann *et al.* 2014; Beauruelle *et al.* 2017; Fabre *et al.* 2012; Kovanen *et al.* 2014; Kupczok and Bollback 2013; Shariat and Dudley 2014; Yin *et al.* 2013) but the temporal dynamics of spacer acquisition and deletion are dependent on the viral context (Kuno *et al.* 2014; Tyson and Banfield 2008) and are highly

variable between species, casting doubt on the suitability of CRISPR as a universal hypervariable marker. For instance, Beauruelle *et al.* (2017) discovered that new spacers were acquired by Group B *Streptococcus* in the span of a few years, whilst Savitskaya *et al.* (2017) observed nearly identical CRISPR arrays in a present-day strain of *E. coli* compared with one recovered from the guts of a 42 000 year-old frozen woolly mammoth. Yet, if CRISPR can be used for fine-scale strain typing, characterising the local structure of bacterial populations may become possible even for uncultivated bacterial species and mixed populations.

Fine-scale strain typing would be particularly useful in conservation applications requiring knowledge of the structure and connectivity of populations of symbiotic bacteria. For example, in deep-sea chemosynthetic ecosystems such as hydrothermal vents, estimating symbiont population connectivity could inform the development of conservation strategies to mitigate the impacts of future deep-sea mining of polymetallic sulfide deposits. Estimations of macrofaunal connectivity are already an integral part of several frameworks aimed at assessing the resilience of proposed mining sites and developing preservation areas (Boschen *et al.* 2016; Ellis *et al.* 2017; Decision of the Assembly of the International Seabed Authority Relating to the Regulations on Prospecting and Exploration for Polymetallic Sulphides in the Area 2010). Assessments of microbial population structure and connectivity would be prudent in areas populated by keystone taxa that rely on obligate symbionts for their survival.

One such environment exists in the eastern Pacific Ocean, where hydrothermal vent communities are dominated by various species of gutless siboglinid polychaetes whose dense aggregations create niches for other faunal species. These tubeworms all rely on a single species of uncultured chemolithoautotrophic Gammaproteobacteria coined *Candidatus* Endoriftia persephone (Robidart *et al.* 2008) for their nutrition. These bacterial symbionts are acquired *de novo* from the surrounding environment at each generation (Harmer *et al.* 2008) by young tubeworm larvae during a short infection-phase and proliferate within the cells of a special hosting organ known as the trophosome. Despite the essential nature of the symbionts for these habitat-forming worms and therefore entire vent communities, little is known about the organization of their populations and their connectivity, particularly at the regional scale which is the relevant scale for conservation purposes.

Genetic studies investigating the phylogeography of *Ca.* E. persephone showed the symbionts associated with the species of the East Pacific Rise (*Riftia pachyptila* and *Tevnia jerichonana*,

Oasisia alvinae) at tropical latitudes, and those associated with worms of the Juan de Fuca Ridge (*R. piscesae*, *Lamellibrachia* sp.) in the northeast Pacific, belong to two vicariant populations (Di Meo *et al.* 2000; McMullin *et al.* 2003; Perez and Juniper 2016). At different scales, previous studies of intra-host symbiont diversity in tubeworms inhabiting hydrothermal vents or hydrocarbon seeps (Breusing *et al.* 2020; Duperron *et al.* 2009; Forget *et al.* 2014; Patra *et al.* 2016; Perez and Juniper 2018; Polzin *et al.* 2019; Reveillaud *et al.* 2018; Zimmermann *et al.* 2014), have consistently found low genetic diversity at the species-level but evidence for multiple strains. Such results also highlight the fact that conventional genetic markers do not provide a high enough resolution to uncover the true strain-level diversity of the symbionts. Also, it has been proposed the symbionts can escape the tissues of dead hosts and return to a free-living stage (Klose *et al.* 2015), potentially linking host-associated and free-living symbiont pools by strong gene flow. Furthermore, metagenomic sampling of environmental biofilms (Polzin *et al.* 2019) and fluorescently-labeled in situ hybridisation of colonisation blocks (deployed for one year) (Harmer *et al.* 2008) revealed that free-living *Ca.* E. persephone are most abundant in close proximity to host aggregations and almost undetectable away from zones of hydrothermal activity. Taken together, these observations suggest the genetic diversity of the symbionts is spatially structured and can be uncovered from host-associated populations with a suitable hypervariable genetic marker.

The goal of this study was therefore to evaluate the CRISPR sequence as an appropriate genetic marker for distinguishing multiple environmental strains of the uncultured vestimentiferan symbiont *Ca.* E. persephone. To address this objective, we used CRISPR along with four other gene markers to characterize the structure of *Ca.* E. persephone populations along the Juan de Fuca Ridge, where the symbiont species is associated with the host tubeworm species *Ridgeia piscesae.* Doing so, we assessed if the physicochemical conditions of the worm habitat contributed to symbiont population structure, and the extent to which the populations along the Juan de Fuca Ridge were connected.

Material and methods

Sampling design

Populations of environmentally acquired *Ca.* E. persephone were sampled from their *R. piscesae* hosts in three active hydrothermal venting regions separated by increasing N-S distances along the Juan de Fuca Ridge: Main Endeavour Field (MEF), Clam-Bed (CB) and Middle Valley (MV). Within each region we sampled contrasting habitats which were identified from the morphotypic appearance of the individual *R. piscesae* hosts which exhibit environmentally-driven phenotypic plasticity (Southward *et al.* 1995) (Figure 3.1 and Table S3.1 in the Supplementary material, Annex II). The first habitat called 'High-flow' is typically located on sulfide edifices, close to points of vigorous discharge of hydrothermal fluids (Figure 3.1B). The average fluid temperature in the tubeworm aggregations growing in this environment was 10°C at level of the gills and 37°C at the base of the tubes. The second environment ('Low-flow') is also located on sulfide chimneys but away from discharge zones (Figure 3.1C). Temperatures in the 'Low-flow' tubeworm bushes ranged from 4°C (base) to 16°C (gill level). Finally, we referred to the third habitat type as 'Basalt-hosted'. This environment was located in the vicinity of the hydrothermal edifices, where the venting fluids emerged from basalts rather than sulfide accretions (Figure 3.1D). Temperature recorded at both the plume and base level of the tubeworm bushes in these peripheral habitats was around 2°C, slightly above the ambient seawater temperature of 1.8˚C. Temperature has been shown to be a reliable proxy for sulfide concentrations in Juan de Fuca Ridge hydrothermal vent fluids (Butterfield *et al.* 1990).

Using the intracellular symbionts to assess the diversity of the *Ca.* E persephone population as a whole is problematic because we do not know if the host-associated symbionts are representative of the free-living pool. For example, the infection process likely represents a bottleneck event and it is not known if discriminatory selection occurs during the development of the trophosome. To overcome this limitation, we sampled multiple worms from each aggregation/sampling site, and intra-host variation was assessed by partitioning the trophosomes of several worms from the MEF sites into three to four transverse sections (Figure 3.1A). Because of budget constraints, only three replicate symbiont populations (*i.e.* three individual host worms) were examined per site. For the same reason, intra-host variation could not be assessed for all worms so we restricted this analysis

to a few host worms from MEF in order to compare intra-host variation across all three habitats (Figure 3.1).

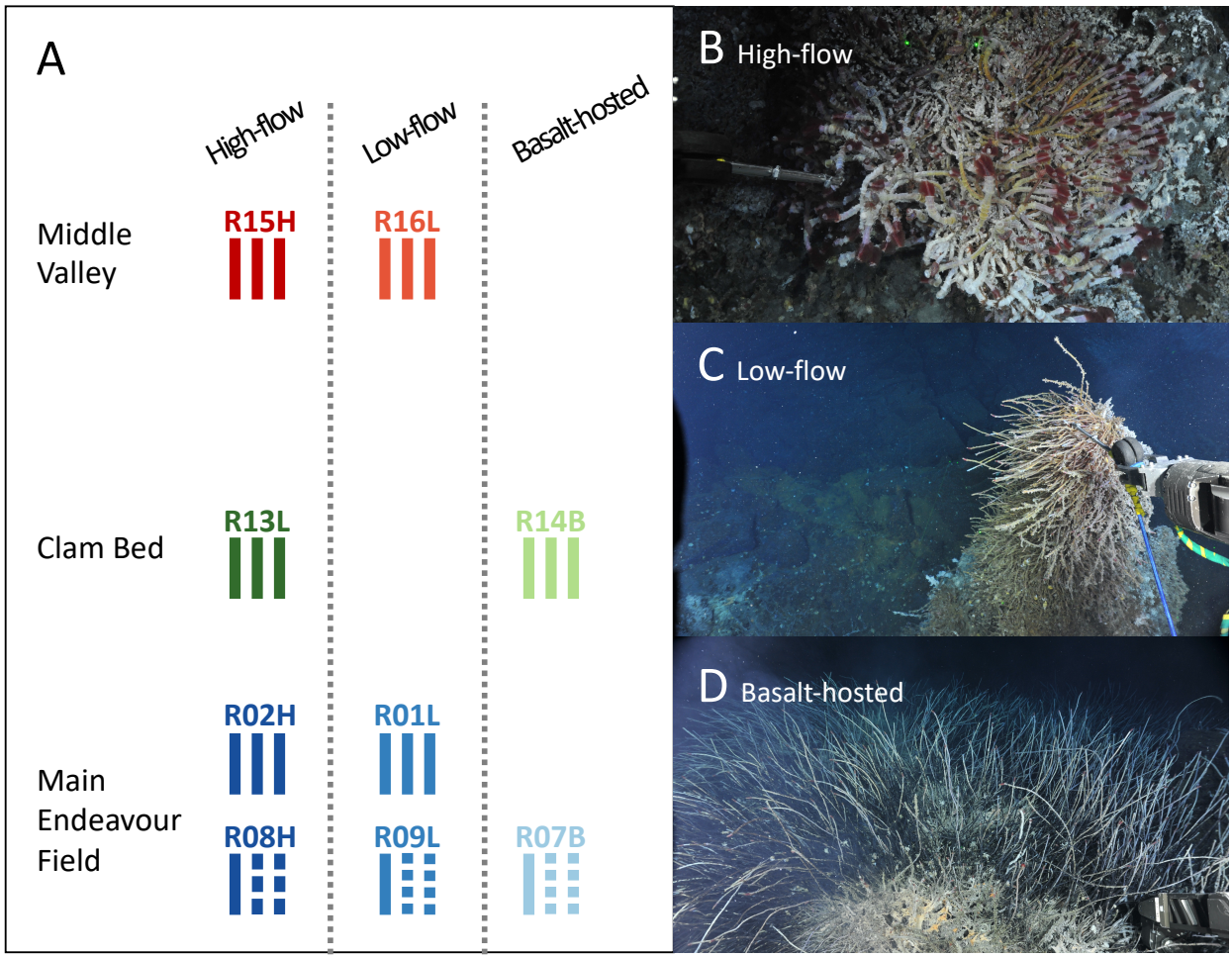

Figure 3.1 Environmental sampling design. A) Schematic representation of the sampling design. Each bar represents an individual worm from which a symbiont population was sampled. Segmented bars represent worms that were sectioned. Sampling sites within individual vent fields were separated by ~10 m. Clam Bed is located ~2 km north of the Main Endeavour Field and Middle Valley is ~60 km further north. **B)** Tubeworm aggregation at site R08H, a typical a 'high-flow' habitat. **C)** Tubeworm aggregation at the 'low-flow' site R09L. **D)** Tubeworms of a 'basalt-hosted' habitat (site R07B).

Sample collection and DNA extraction

Tubeworm assemblages typical of the three contrasting environmental conditions, were sampled in June 2013 and June 2016 during two deep-sea expeditions that deployed remotely-operated vehicles (ROVs) from research vessels. The first, on board the R/V Thomas G. Thompson, used the ROV Oceaneering Millennium, whilst the second, on the CCGS John P. Tully, used the ROV ROPOS (Table S3.1 in the Supplementary material, Annex II). Worms collected in 2013 were

processed upon recovery to the surface vessel, whereas those collected in 2016 were individually packed, frozen at -80°C and later dissected in the laboratory. The individual worms were carefully removed from their tubes and treated with lysozyme and DNAse according to Elsaied and Nagamura (2001) to remove epibiotic contamination. Subsequently, the trunks of the worms were separated, and for some individuals split into 3 to 4 segments (see sampling design), before being placed in 95% ETOH pending DNA extraction. Trunk sections were later finely chopped with scissors and homogenized by strong vortexing to release symbionts cells from the trophosome tissues. We then collected and precipitated a subsample of each of the symbiont-enriched suspensions. DNA was extracted using the phenol-chloroform method followed by ethanol precipitation (Sambrook *et al.* 1989).

Genetic sampling and sequencing

To confirm that *Ca.* E. persephone was the only symbiont species within *R. piscesae*, we amplified and sequenced the hypervariable region V4 of the bacterial 16s rRNA gene using universal primers. To assess the intra-specific genetic diversity of the symbionts we amplified and sequenced a complete CRISPR array previously found on the scaffold KQ557120 (48218..48978) (start-end positions) in the assembly '*Ridgeia* 1 symbionts' (GenBank accession LDXT01). Additionally, we performed a Multi-Locus Sequence Analysis (MLSA) by sequencing three additional protein-coding genomic regions. Rather than using genes typically employed in MLSA analyzes (*e.g. recA*, *gyrB, rpoB, rpoD, groEL, atpD* [Glaeser and Kämpfer 2015]), we selected the protein-coding sequences that had the highest potential for displaying polymorphism based on Single Nucleotide Polymorphisms (SNPs) previously detected in the metagenomics sequences from the trophosomes of one, and a pool of five individual tubeworms, respectively (Perez and Juniper 2018). Candidate genes had to be uniquely represented in the *Ca.* E. persephone genome, belong to a well-defined COG (Cluster of Orthologous Genes) category, have multiple SNPs within 600bp of each other but no indels. Six genes fitted these criteria but only three were successfully amplified: *lpxA*, *pleD,* and *tufB*.

Libraries were prepared according to Génome Québec guidelines. A first PCR was performed to amplify the genomic regions of interest. We used gene-specific primers that carried the CS1 and CS2 universal overhangs. These extra 22bp sequences allowed for the attachment of sample-

specific barcodes during a second PCR round. Primer sequences and PCR conditions are presented in Table S3.2 in the supplementary material (Annex II). Ultimately, 35 libraries of pooled barcoded amplicons (one barcode per sample) for the polymorphic gene fragments (*i.e.*, 16S rRNA V4 region, *lpxA*, *pleD,* and *tufB*) were then sent to Génome Québec for sequencing on the Illumina MiSeq 2500 platform (1% of a lane), and 41 libraries for the CRISPR PCR products were sequenced on one PacBio SMRT cell.

In silico haplotype detection

Gene amplicons

Sequencing of the four pooled gene amplicons (16S rRNA gene- V4 region, *lpxA*, *pleD,* and *tufB* gene fragments) yielded between 1385 and 3199 reads per sample. Of these, 70% were concordantly mapped to the reference genome. In order to isolate the amplicons of different loci, the paired-end reads were mapped onto the reference genome using bowtie2 v2.3.2 (Langmead and Salzberg 2012) with the following parameters: -D 15 -R 2 –N -L 20 –I S,1,0.75 –dovetail –qseq –X 600. For each gene, the mapped reads were then extracted with samtools v1.9 (Li *et al.* 2009) and the bamtofastq program from bedtools v2.27.1 (Quinlan and Hall 2010).

For the hypervariable region V4 of the bacterial 16s rRNA gene, we extracted both the reads that mapped to the 16S rRNA gene and those that did not map to the reference at all. These sequences were then together processed with the software package Divisive Amplicon Denoising Algorithm 2 (DADA2 v1.17.0) (Callahan *et al.* 2016, p. 2) in R, according to the pipeline tutorial version 1.6 (https://benjjneb.github.io/dada2/tutorial.html). For the other housekeeping genes (*lpxA*, *pleD,* and *tufB*), polymorphic positions across all mapped reads were initially detected with VarScan v2.3.9 using the following parameters: --min-coverage 100 --min-reads2 10 --min-avg-qual 25 --min-var-freq 0.01 (Koboldt *et al.* 2009). All SNPs identified matched known SNPs from the reference (Table S3.2 in the Supplementary material, Annex II) and no additional variable sites were found. The putative ancestral haplotypes for these genes was determined from their respective nucleotide sequences in the genome of *Ca.* E. persephone associated with tubeworms of the East Pacific Rise (Gardebrecht *et al.* 2012). Then, extracted reads were merged with bbmerge (BBmap v38.70) (Bushnell *et al.* 2017) using 3’ quality trimming, transformed to fasta format conservatively changing low quality (<28) nucleotides to Ns, and aligned with Muscle v3.8.31 using default

parameters (Edgar 2004). Finally, the alignments were truncated to the two SNP positions and the haplotype frequencies were counted with a custom python script. Except for two samples for the *tufB* amplicon which failed to amplify, the final minimum coverage on the gene markers reached 94X (average 286X, 579X, and 378X for *lpxA*, *pleD*, and *tufB*, respectively)

CRISPR array

A total of 35840 high quality PacBio reads from CRISPR amplicons were generated through circular consensus sequencing (CCS). Initial attempts at direct spacer detection in the CCS reads using two available methods designed for Illumina reads, Crass v0.3.12 (Skennerton *et al.* 2013) and MetaCRAST (Moller and Liang 2017), yielded too many artifactual spacers because of the higher error rate of the CCS reads compared to Illumina. We thus adopted a different approach shown in (Figure 3.2).

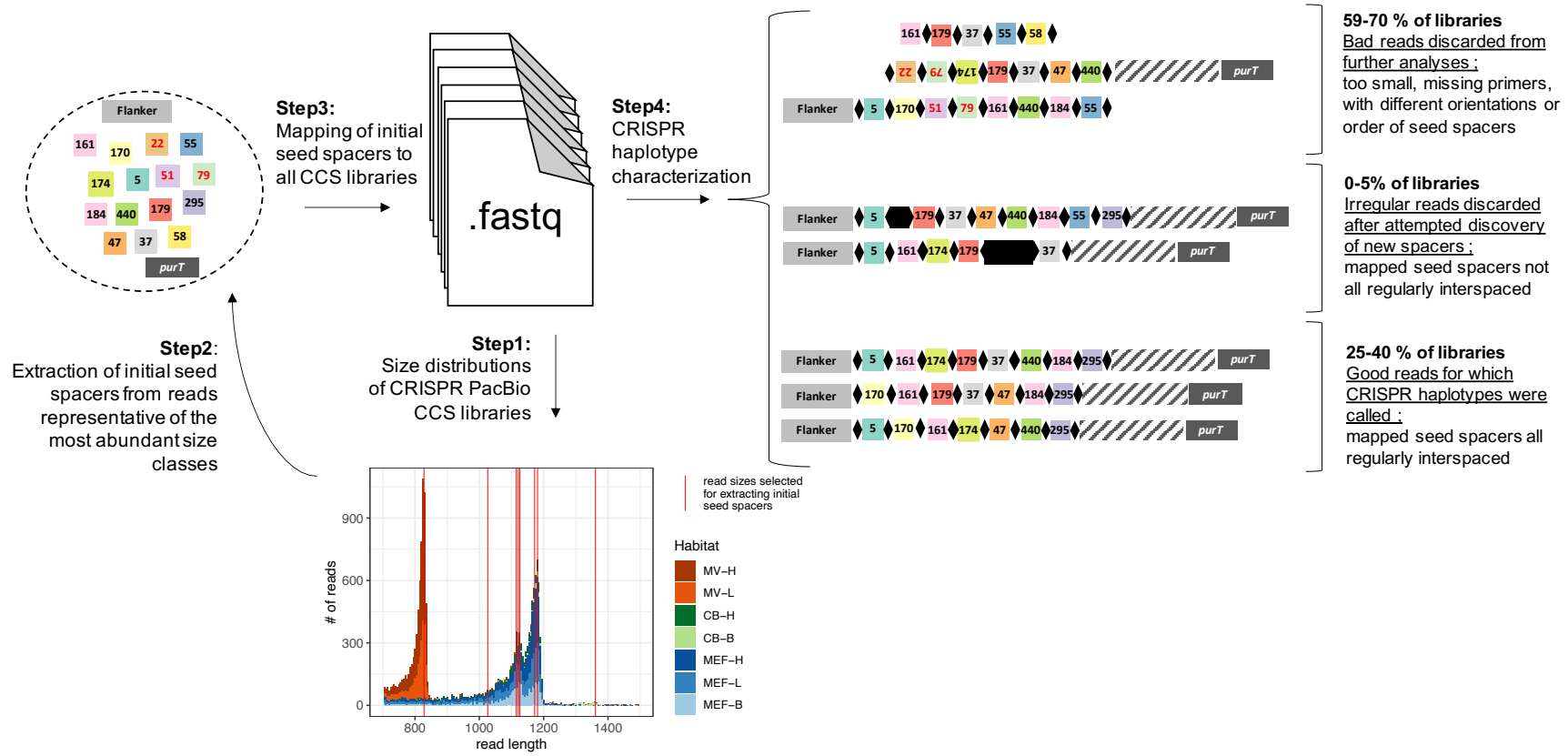

Figure 3.2 Schematic representation of the workflow for CRISPR haplotype detection.

First, we established read length distributions for each sample. A small subset of reads representative of the most abundant size classes was then extracted and a set of seed spacers were identified with Crass v0.3.12 (Skennerton *et al.* 2013) using all default parameters except kmer length which we picked to match the length of the repeat sequence (-K30). Of the resulting 15 seed spacers identified, 11 had been previously described for this array (Perez and Juniper 2018) and three were new. Next, the seed spacers along with subsequences of the two primers used for PCR amplification of the whole CRISPR array were mapped back to all CCS reads using an approximate string-matching algorithm implemented in the python package fuzzysearch 0.7.2. Near-matches with a minimum of 85% identity (*i.e.,* a maximum of 5 nucleotide mismatches) between the spacer sequence and the read were tolerated to account for sequencing errors and mutations. We discarded from further analysis dubious reads which were missing at least one of the primer sequences, were less than 33 bp (the approximate size of a single spacer), or whose mapped seed spacers did not all have the same orientation or were properly ordered. Discarded reads represented 59 to 70% (22738 reads in total) of the CCS libraries and most (>90%) were excluded because of missing primer sequences. To assess whether additional spacers (not detected from the initial read subset) were present in the samples, we flagged all reads for which mapped seed spacers were not regularly interspaced (0-5% of the libraries; 439 reads in total), as putatively containing new spacers. The sequences of these putative new spacers including their bordering repeats were extracted to a new fastq file and processed with Crass v0.3.12 using the same parameters as for the initial seed spacer search (Skennerton *et al.* 2013). Only three distinct spacers were found and all had matches in the initial seed spacer set but with a score slightly below the conservative threshold of 85% identity we used for initial mapping. Because they represented only a small proportion of the libraries and had lower quality, these reads were also excluded from further analyses. The remaining 25-40% of the libraries contained the highest quality reads for which mapped seed spacers were all regularly interspaced. It is in this read set (12663 reads) that unique arrays of CRISPR spacers (also referred to as CRISPR haplotypes) were called.

CRISPR phylogeny

We estimated the genetic distance between pairs of CRISPR haplotypes by implementing the probabilistic algorithm described by Kupczok and Bollback (2013) for estimating the parameters underlying the ordered independent spacer loss model. This model assumes spacers are

independently added at the leader end of the array and independently lost one at the time throughout the array. The parameters estimated are the insertion to deletion rate ratio and the divergence time between each haplotype pair and its most recent common ancestor. The resulting distance matrix between the haplotypes and their most recent common ancestors was used to reconstruct the phylogeny of the CRISPR arrays. To do so, we used a modified version of the rooted neighbor joining method presented in Kupczok and Bollback (Kupczok and Bollback 2013) which does not allow for negative branch lengths. As in Kuhner and Felsenstein ("A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates.," 1994), each negative branch was corrected to zero during tree construction and the corresponding difference was added to the adjacent branch length in order to preserve the total distance between adjacent pairs of terminal nodes. The genetic distances between haplotype pairs were then computed from the distances between pairs of terminal nodes in the tree.

Population structure

Analyses of population structure were performed for the CRISPR array and each gene amplicon independently in R using the package "Proppr" v2.8.6 (Kamvar *et al.* 2014). Each read was considered as an individual (*i.e.* a unique symbiont cell; ignoring PCR amplification biases) and each individual worm host represented a discrete bacterial population.

Minimum-spanning trees based on the previously estimated genetic distances between CRISPR haplotypes were constructed using the function *poppr.msn*. The function first computes a minimum spanning tree from a graph representation of an adjacency matrix (here, that of the pairwise CRISPR distances) and then add population parameters as attributes to this tree.

Hierarchical AMOVAs (Excoffier *et al.* 1992) were performed with the wrapper *poppr.amova* which uses the *amova* function from the "ade4" package (Dray and Dufour 2007). The following levels were tested: regions, habitats, sampling sites, individual hosts, trophosome sections within hosts, technical replicates. A permutation test with 1000 permutations (function *ade4:randtest*) was used to assess the statistical significance of the various covariance components (*i.e.*, the hierarchical levels). To assess the concordance of the symbiont population structures according to the CRISPR array and each of the other gene fragments, we performed Mantel tests on their respective matrices of pairwise population differentiation. For the gene amplicons, the F_{ST} index

computed in Arlequin v3.5 (Excoffier and Lischer 2010) was used as a measure of distance between pairs of symbiont populations and non-significant F_{ST} values were treated as 0 (function *mantel.test* from the "ape" R package) or removed from the correlation coefficient computation (function *mantel* from the "vegan" R package).

Data availability

All supporting data have been provided within the article or through supplementary data files (Annex II). Raw sequences used in this paper were deposited on GenBank under the BioProject PRJNA641184. Python scripts, R scripts, and the implementation of Kupczok and Bollback (2013) method are publicly available at http://github.com/maepz/CRISPR_distance.

Results and discussion

Unlike DNA barcoding using 16S rRNA gene, CRISPR-typing uncovers the high diversity of environmental symbiont strains.

The sequencing of the 16S rRNA gene- V4 amplicon yielded 8571 paired-end reads in total. The average per-sample coverage was 226X but there were important disparities across samples (Figure 3.3). Using these reads together with all unmapped amplicons, we used dada2 to determine the symbiont genetic diversity.

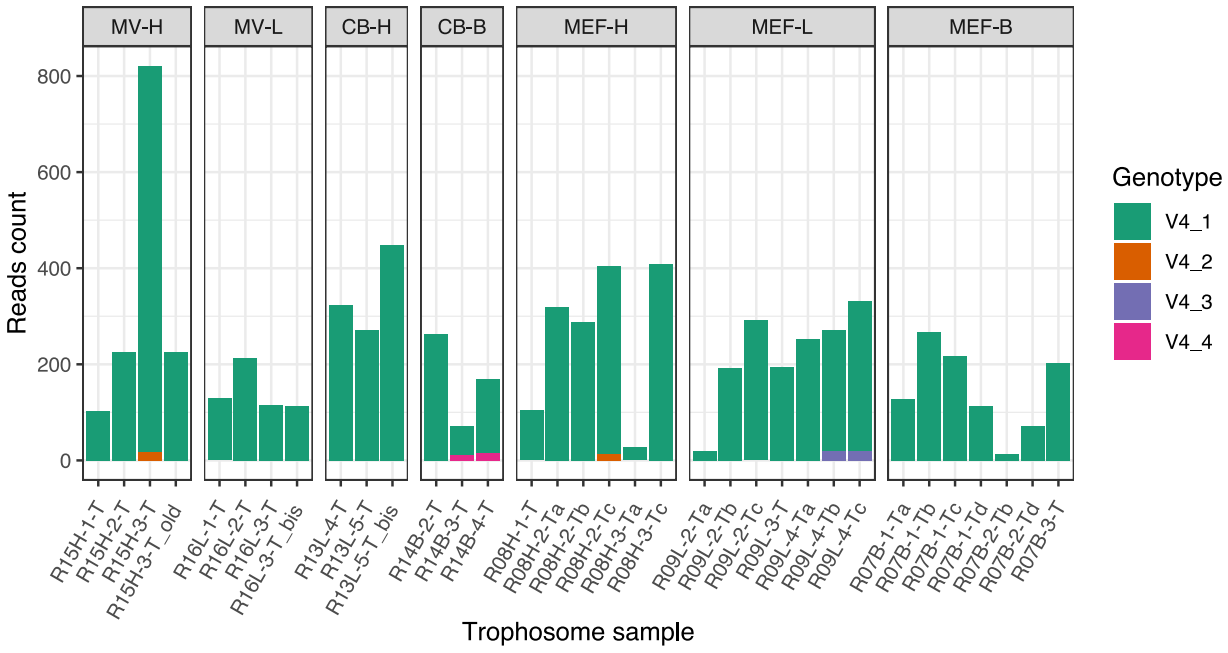

Figure 3.3 Symbiont genetic diversity according to the 16S rRNA gene16S rRNA gene hypervariable V4 region. Samples suffixed 'old' and 'bis' are technical duplicates. Only high confidence genotypes, which accounted for ten reads in at least one of the samples are represented. The minimum nucleotide identity between genotypes was 81% between V4_2 and V4_3. All genotypes were identified as *R. piscesae* symbionts by nucleotide blast against the NCBI's nucleotide collection. MV: Middle Valley; CB: Clam-Bed; MEF: Main Endeavour Field; H: high-flow, L: low-flow, B: basalt-hosted.

We identified four bacterial phylotypes with a minimum nucleotide identity of 81% between V4_1 and V4_4 (Figure 3.3). The phylotypes V4_1 was identical to that of the reference symbiont genome (Perez and Juniper 2016) and overwhelmingly dominated the trophosome assemblages. A blast search of the other phylotypes against NCBI's database identified them all as known endosymbionts of *R. piscesae*. Although great precaution was taken to reduce contamination, it is unclear given their low abundance and considerable divergence to the dominant endosymbiont phylotype if these alternative rare taxa are truly from host associated bacteria. Regardless, these observations support the hypothesis that host-symbiont molecular interactions and microbial competition prevent the infection and intra-host proliferation of excessively divergent phylotypes (Polzin *et al.* 2019).

With 123 distinct CRISPR arrays detected, the hypervariable CRISPR region revealed a much higher symbiont genetic diversity than the 16S marker. Between 3 and 32 (median = 11) distinct CRISPR haplotypes were found in each individual host. The majority of these haplotypes were in very low abundance; two thirds or more of the haplotypes were represented in fewer than 5% of

the reads. Nonetheless, rare haplotypes could not be identified as somatic variants (*i.e.*, strains resulting from within-host mutations). Indeed, examination of a minimum spanning tree showed most CRISPR haplotypes were shared amongst host worms (Figure 3.4). Furthermore, haplotypes present in a single host did not form phylogenetic clusters as would be expected from clonal populations (Figure 3.4). Hence, we conclude that these haplotypes probably reflect the diversity of the environmental strains of *Ca.* E. persephone.

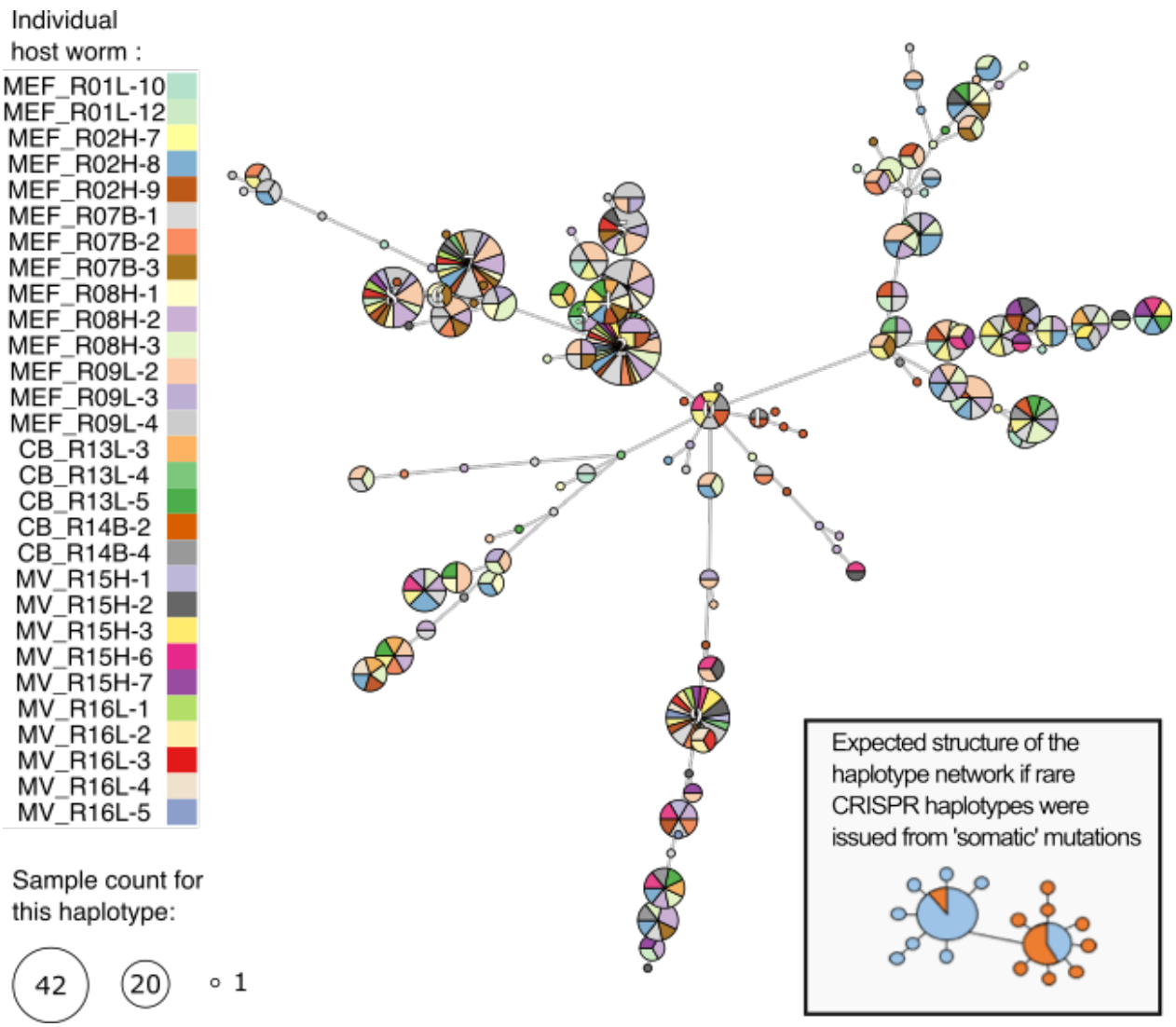

Figure 3.4 Minimum spanning tree for the CRISPR haplotypes coloured according to individual hosts. The sizes of the circles represent the number of trophosome samples within which a particular haplotype was found. The 10 most abundant haplotypes are labelled.

The architecture of the CRISPR array retraces the known symbiont phylogeny.

Examining the structure of all the CRISPR arrays recovered from the Juan de Fuca Ridge tubeworms, we found that the genetic diversity amongst the symbiont haplotypes is defined by various spacer deletions from the longest array. The longest array thus represents the most ancestral state amongst the CRISPR arrays we sampled (Figure 3.5).

Surprisingly, we observed no new spacers at the array's leading end where insertions would be indicative of recent viral infections. Other known examples of deletion-driven CRISPR polymorphism are found in *E. coli, Salmonella* and *Klebsiella* species (Shariat *et al.* 2015; Shen *et al.* 2017; Touchon *et al.* 2011). It is possible the immune function of CRISPR has been lost in this species. At evolutionary time scales, metabolic functions that are associated with a free-living lifestyle such as these providing immunity against phages are the first to be lost when symbionts are transitioning from being environmentally-acquired facultative partners to becoming obligate and vertically transmitted organelle-like organisms (Burstein *et al.* 2016; Moran *et al.* 2008; Newton and Bordenstein 2010). Furthermore, in *R. pachyptila* the CRISPR-cas operon of *Ca.* E. persephone does not seem to be expressed within the trophosome (Gardebrecht *et al.* 2012; Hinzke *et al.* 2019; Markert *et al.* 2007). Yet, in *Ca.* E. persephone the maintenance of the structural integrity of the arrays suggests the CRISPR/cas system may not be completely defunct and could hold instead an alternative function, notably during the free-living stage of the symbionts. Indeed, CRISPR/cas systems in free-living *Ca.* E. persephone may be involved in a number of physiological responses to environmental stress (Ratner *et al.* 2015) and promote host colonization (Gunderson and Cianciotto 2013; Sampson *et al.* 2013; Veesenmeyer *et al.* 2014).

We then assessed if any of the 15 identified spacers had previously been sequenced by blasting them against JGI's IMG-MER database. Three spacers found match in contigs from other bacterial genomes and metagenomes. Because these contigs possessed the conserved end of the CRISPR array up to the *purT* gene, we are confident they are DNA fragments of *Ca.* E. persephone. The oldest identified spacer (spacer 295) in our dataset was also present in contigs from three metagenomes of bacterial communities sampled from the surface of three species of polychaete worms from the East Pacific Rise: *T. jerichonana, R. pachyptila* and *Alvinella pompejana*. Hence, spacer 295, the most ancient spacer in the array was most likely acquired before the vicariance of

the Juan de Fuca Ridge and East Pacific Rise symbiont populations, following separation of the two ridges by the fragmentation of the Farallon plate about 30 million years ago (Atwater and Stock 1998; Perez and Juniper 2016).

Given the presence of spacer 295 in many samples from the East Pacific Rise, it is surprising this spacer was not found in the reference genomes of *Tevnia*- and *Riftia*-associated symbionts (Gardebrecht *et al.* 2012). We suspect this is due to the fragmented and incomplete nature of these assemblies; the CRISPR locus was consistently found at contig ends.

The next two most ancient spacers (spacers 58 and 55) were both detected in a contig from a metagenome of diffuse hydrothermal fluids at Axial Seamount, a shallower site on the Juan de Fuca Ridge located about 200 km south of the Main Endeavour field. None of the other spacers inserted at the leader end of the arrays were shared between Axial Seamount and the northern Juan de Fuca Ridge populations, suggesting a lack of connectivity between these populations which is supported by genetic analyses of their *R. piscesae* host (Puetz 2014; Young *et al.* 2008). Together, these results show the CRISPR array retraces the known phylogeny of the symbiont over millions of years.

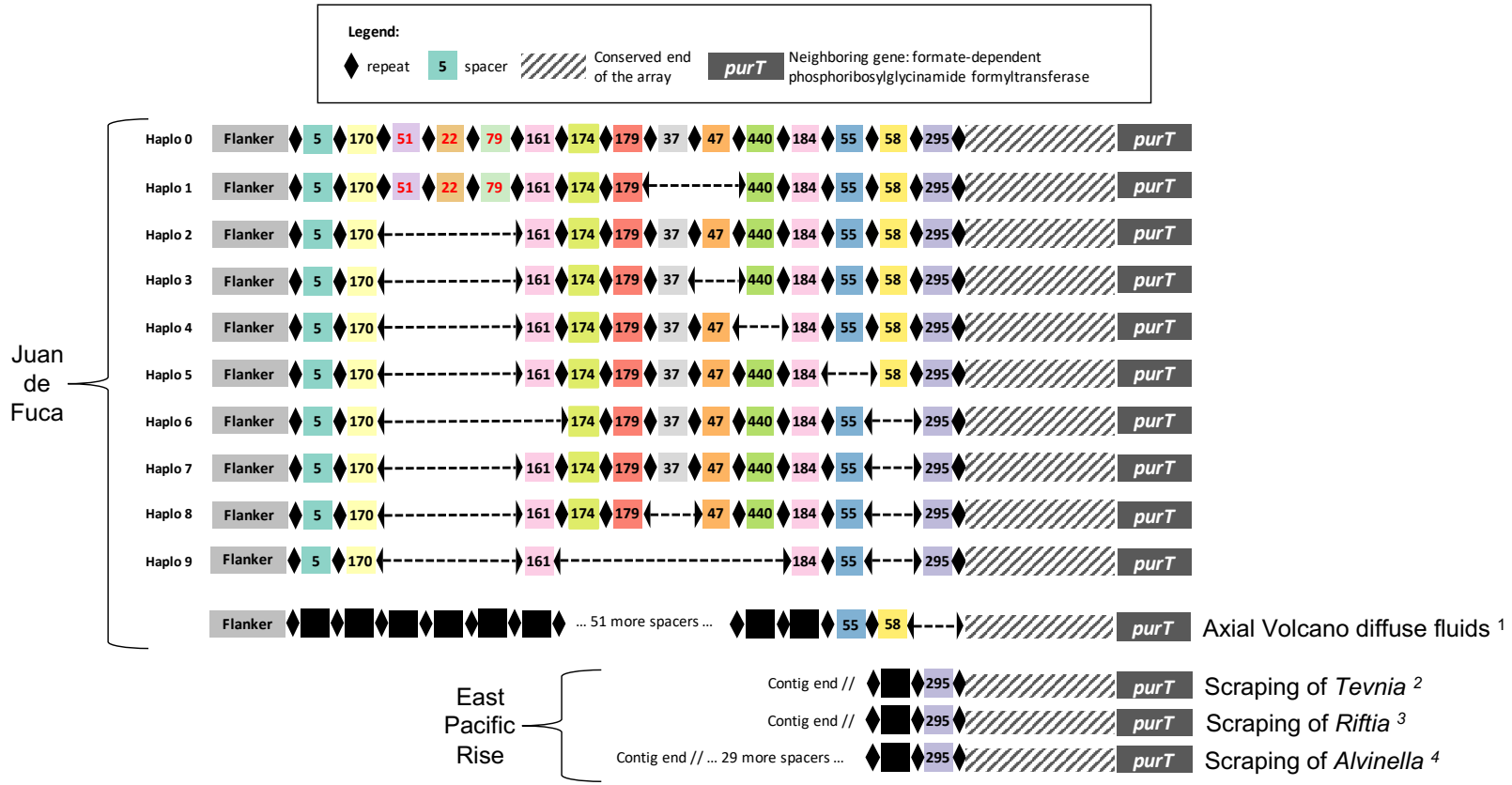

Figure 3.5 Schematic representation of the main CRISPR arrays of symbionts on the Juan de Fuca Ridge and East Pacific Rise. These arrays were present in at least 5% of the CCS reads in at least one trophosome sample and represented more than 95% of the reads overall. Repeats are represented by lozenges and each unique spacer is represented by a coloured square and ID number. Spacer ID numbers were randomly assigned and refer to those used in Perez and Juniper (2016). Spacers newly detected in this study are identified by red ID numbers. Dashed line between repeats represent missing spacers in the arrays. Black squares represent unique spacers in the metagenomic datasets. Note that while the most ancient spacer identified was spacer 295, a region of about 360 bp (hatched segment) extended from this spacer to the start of the following gene (purT). We suspect spacers in this region were not detected because accumulated mutations in the repeats rendered them unrecognizable by the spacer detection software. IMG scaffold references: [1] Ga0105700_1013652 ; [2] Ga0256846_1000282, Ga0256846_1005979, Ga0256846_1868271, Ga0256846_1117288; [3] Ga0256845_1170191; [4] Ga0256843_1000224.

The local symbiont population structure according to CRISPR is corroborated by other hypervariable gene markers.

Three additional housekeeping gene fragments were amplified from the same individual hosts examined for CRISPR; *lpxA*, *pleD* and *tufB* (see Supplementary material in Annex II). These genes were specifically chosen because they exhibited polymorphism in metagenomic data (Perez and Juniper 2018) but only two of the three were informative. The genetic diversity of *tufB* was characterized by two pairs of haplotypes in similar proportions which we suspect results from sequencing two recombining paralogous sequences (Hughes 2000; Lathe and Bork 2001; Santoyo and Romero 2005); the second copy may have been missing from our incomplete reference assembly causing us to mistakenly consider this gene as single copy gene. Haplotype frequencies for the two informative gene markers (*lpxA* and *pleD*) and the CRISPR array were used to compute matrices of population differentiation based on pairwise F-statistics. Mantel tests (see Supplementary material in Annex II) confirmed the haplotypes across *lpxA* and *pleD* markers exhibit a significant degree of covariation and revealed strong concordance to the symbiont population structure uncovered with CRISPR.

Barriers to connectivity rather than local environmental conditions seem to be responsible for partitioning the symbiont populations.

CRISPR-based inferences are presented here and corroborating AMOVA results for the other gene amplicons are provided in the Supplementary material (Annex II). Within the Main Endeavour Field (MEF) region, two independent tubeworm aggregations for each of the High- and Low-flow environmental conditions were sampled to discriminate between habitat and aggregation-specific variation (Table 3.1). Furthermore, the symbiont housing organs of several individuals were partitioned into three to four sections to assess intra-host variation.

Within and between host variation

Significant variation in the composition of the symbiont strain assemblages was found across the length of the trophosome with the gene markers but not with CRISPR, the marker displaying a greater allelic diversity. This indicates that small contrasts in the symbiont composition along the length of the trophosome are likely exacerbated when genetic resolution is low and probably result

from the random distribution of the different symbiont strains in the trophosome. In other words, even though the different symbiont strains may not be homogeneously distributed within the host housing organ, they are not partitioned in a specific way along its antero-posterior axis.

AMOVA analyses within MEF revealed that individual worms from the same aggregation could host markedly different strains; between-hosts variance accounted for nearly 25% of the total variation (Table 3.1). This differentiation supports the hypothesis that the infection is not a continuous process but occurs during a small window of time (Nussbaumer *et al.* 2006).

Table 3.1 Phylogenetically-informed hierarchical AMOVA for symbiont populations in the Main Endeavour Field (MEF). The Lingoes transformation was applied to the haplotype distance matrix to satisfy the Euclidian criterion.

Hierarchical level of variation	Df	Sum Sq	Mean Sq	Sigma	%	P.value	F-statistic
Between habitat	2	18148	9074	3.70	11.44	N.S.	F Flow-Total: 0.11
Between sites within habitat	2	2173	1086	-2.90	-8.96	N.S.	F Site-Flow: -0.10
Between hosts within sites	9	33714	3746	7.80	24.11	0.009	F Ind-Site: 0.25
Between sections within hosts	12	9596	800	2.08	6.42	N.S.	F Section-Ind: 0.09
Between duplicates within samples	1	154	154	1.08	3.33	0.001	F Samples-Section: 0.05
Within samples	6159	126799	21	20.59	63.66	0.001	F Samples-Total: 0.36
Total	6185	190584	31	32.34	100.00		

Table 3.2 Phylogenetically-informed hierarchical AMOVA for symbiont populations in the Main Endeavour Field (MEF), Clam-Bed (CB) and Middle Valley (MV). The Lingoes transformation was applied to the haplotype distance matrix to satisfy the Euclidian criterion.

		Df	Sum Sq	Mean Sq	Sigma	%	P.value	F-statistic
All regions	Between regions	2	446357	223178	59.39	66.87	0.001	$F_{Region-Total}$: 0.67
	Between habitats within regions	4	31696	7924	1.41	1.58	N.S.	$F_{Flow-Region}$: 0.05
	Between hosts within habitats	23	91384	3973	10.17	11.45	0.001	$F_{Ind-Flow}$: 036
	Within hosts	12633	225439	18	17.85	20.09	0.001	$F_{Ind-Total}$: 0.80
	Total	12662	794875	63	88.81	100.00		
	Between habitats	2	63857	31928	0.76	1.17	N.S.	$F_{Flow-Total}$: 0.01
	Between hosts within habitats	27	505579	18725	46.81	71.55	0.001	$F_{Ind-Flow}$: 0.72
	Within hosts	12633	225439	18	17.85	27.28	0.001	$F_{Samples-Total}$: 0.73
	Total	12662	794875	63	65.42	100.00		
Without MV	Between regions	1	9309	9309	2.25	5.33	N.S.	$F_{Region-Total}$: 0.05
	Between habitats within regions	3	25887	8629	0.34	0.80	N.S.	$F_{Flow-Region}$: 0.01
	Between hosts within habitats	15	75356	5024	15.52	36.74	0.001	$F_{Ind-Flow}$: 0.39
	Within hosts	7284	175774	24	24.13	57.13	0.001	$F_{Ind-Total}$: 0.43
	Total	7303	286325	39	42.24	100.00		
	Between habitats	2	18183	9091	-0.57	-1.41	N.S.	$F_{Flow-Total}$: -0.01
	Between hosts within habitats	17	92368	5433	16.82	41.65	0.001	$F_{Ind-Flow}$: 0.41
	Within hosts	7284	175774	24	24.13	59.75	0.001	$F_{Samples-Total}$: 0.40
	Total	7303	286325	39	40.38	100.00		

Between habitat variation

At Clam Bed (CB), the allelic composition of symbionts from the High-flow and Basalt-hosted worm populations was markedly different (see Supplementary Material in Annex II) and could be the result of larger differences in age between the two tubeworm populations. The Basalt-hosted worms at this site are known to be at least several decades old (Urcuyo *et al.* 2007), and closely-related species living in similar environmental conditions may live for centuries (Durkin *et al.* 2017). In contrast, the High-flow worms have likely colonized the CB chimney much more recently (Sarrazin *et al.* 1997; Tunnicliffe *et al.* 2014). Supporting this hypothesis, we found the ancestral haplotypes for all three symbiont genes (CRISPR, lpxA and pleD) were predominant amongst CB's Basalt-hosted populations (see Supplementary Material in Annex II) suggesting the symbionts were established in the Basalt-hosted tubeworms before hosts from the High-flow environment acquired theirs. It is also possible that the ancestral haplotypes of *Ca*. E. persephone were uniquely sustained in high abundance amongst the free-living population at this site. However, our fine-scale genetic survey revealed that while the symbiont populations were structured at the scale of a vent field, this structure was not driven by differences between habitats.

Indeed, broad environmental conditions associated with the concentration of hydrothermal discharge in the worms' habitat generally did not significantly explain the variation observed in the data even when controlling for regional variation or excluding the highly homogeneous Middle Valley sites (Table 3.1 and Table 3.2).

Between region variation

Variance in the symbiont strain diversity appears to reflect general patterns of connectivity along the Juan de Fuca Ridge rather than environmental selection. Regional differences between Middle Valley and the two Endeavour sites (CB and MEF) accounted for most of the regional variance (67%, Table 3.2, A) whereas the symbiont meta-populations were not significantly differentiated between CB and MEF (Table 3.2, B).

Our results interpreted alongside those of the host populations suggest that both host larvae and symbiont cell dispersal depend on patterns of deep-sea circulation that restrict connectivity across disjointed axial rift valleys but maintain it within them. Young *et al.* (2008) and Puetz (2014) found

a similar structure for the host populations. In both studies, tubeworm populations from Middle Valley at the northern extremity of the Juan de Fuca Ridge, which is a topologically isolated basin (McManus *et al.* 1972), were distinct from those of the Endeavour Segment to the south. Furthermore, Puetz (2014) showed high gene flow between host populations inhabiting High- and Low-flow habitat types.

It is noteworthy that in addition to their apparent isolation, symbiont populations from Middle Valley exhibited a surprisingly low diversity. All host individuals in this region were associated with a single symbiont strain identified by all three of the suitable hypervariable markers (CRISPR, *lpxA*, and *pleD*) (Figure 3.6 and Supplementary material in Annex II). This homogeneity likely reflects that of the environmental infection pool in this region. Little is known about the dependence of tubeworm recruitment on the resident environmental symbiont population or how important resident hosts are in maintaining this pool. If *R. piscesae* recruitment in Middle Valley is dependent on this symbiont strain or if robust host populations must be present to seed and maintain the symbiont populations, these worms and the associated communities that depend on them could be extremely vulnerable to disturbance from mining activities. Hence, our results highlight the fundamental importance of better understanding the diversity and connectivity of natural populations of obligate microbial symbionts. As the International Seabed Authority is drafting the first regulations for hydrothermal vent mining, we argue that it is imperative for such keystone bacterial species to be taken into account within conservation schemes.

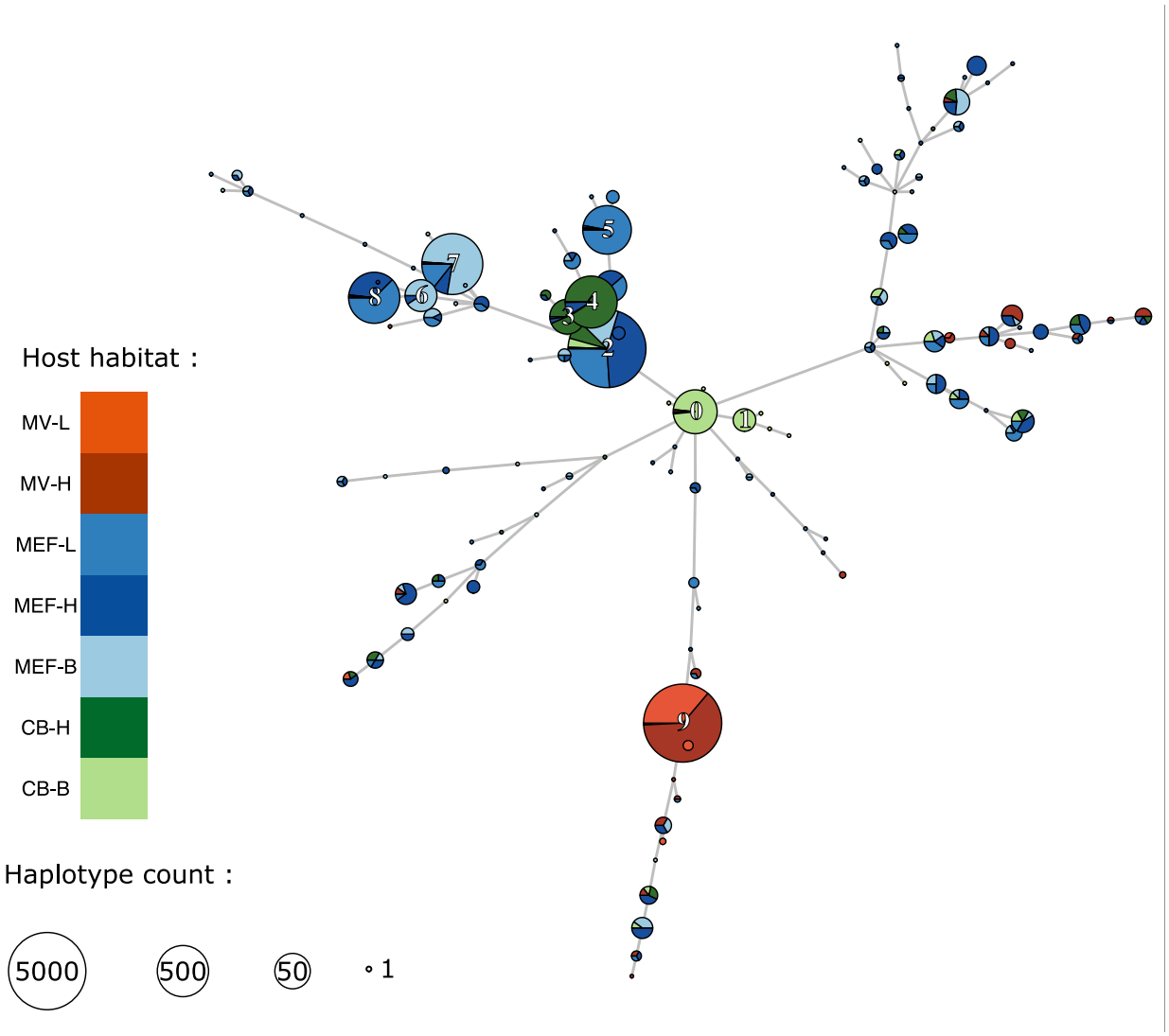

Figure 3.6 Minimum spanning tree for the CRISPR haplotypes coloured according to habitats. The sizes of the circles represent the number of sequenced reads. The 10 most abundant haplotypes are labelled.

Conclusions

Characterising and comparing natural heterogeneous bacterial populations at the strain level is challenging. We have demonstrated that the CRISPR array is a genetic marker fit-for-purpose for the uncultured chemoautotrophic symbiont species *Ca.* E. persephone. In our study, the CRISPR array retraced the known symbiont phylogeny over millions of years but also allowed for discrimination amongst very closely-related lineages. The CRISPR haplotypes we identified only varied through spacer deletions and yet they revealed 30 times more diversity than any of the other gene markers specifically selected for their polymorphism. Furthermore, unlike MLSA methods

which require multiple gene primers and may be biased by paralogous sequences and homologous recombination if the gene markers are not carefully chosen, working with CRISPR requires a single set of primers, the orthology of the marker can be guaranteed through genomic context, and while homologous recombination within CRISPR has been observed, it appears to be extremely rare (Kupczok *et al.* 2015). As an alternative to MLSA, whole genome shotgun sequencing is often preferred for assessing the genetic diversity of heterogeneous bacterial populations. Such an approach has the advantage of revealing the genetic diversity across whole genomes but, in addition to its higher cost (for endosymbionts the sequencing yield from metagenomes is largely reduced by host DNA contamination), this method cannot discriminate between strains at the level of individual bacteria. Thus, the smallest sampling unit is that of the metagenomic population; individual hosts in the case of symbiont studies. Hence, whole genome shotgun sequencing would require extensive field sampling in order to resolve strain-level beta diversity. In contrast, with the CRISPR marker, one sequencing read represents one bacterial cell which can be identified at the strain level. Therefore, this method better harnesses the power of high throughput sequencing for the purpose of strain-level population genetic studies particularly when dealing with unculturable bacterial species.

Nonetheless, there are several limitations to the use of CRISPR for DNA barcoding. First, not all prokaryote species possess the marker. The CRISPR-cas immunity is only present in about half of bacteria (Burstein *et al.* 2016) and the system is rapidly lost in species undergoing reductive genome evolution such as vertically transmitted symbionts (Burstein *et al.* 2016; Moran *et al.* 2008). Second, this marker is not appropriate for characterising whole communities. Because of the great diversity of CRISPR-cas systems and CRISPR arrays (neither the flanker nor the repeat sequences are conserved across species), a universal primer may never be developed. Third, primer development for a single species necessitates a reference genome that includes the CRISPR genomic context. This is because whole CRISPR systems can be horizontally transferred across species (Godde and Bickerton 2006; Shen *et al.* 2017). Hence, to insure orthology of the amplified sequences, one of the primers must target a region next to the array and outside of the operon. Finally, the cost of long-read high-throughput sequencing is still prohibitive. In this study, for the same effective depth of coverage, sequencing the CRISPR array cost roughly three times the amount needed for sequencing the smaller gene amplicons. However, third generation sequencing technology costs are steadily falling and for well-characterized CRISPR arrays other, less onerous,

genotyping methods exist (c.f. CRISPR-typing [Shariat and Dudley 2014]). We therefore conclude that despite these limitations, CRISPR represents a promising tool for strain-tracking in a wide variety of uncultured bacteria.

Acknowledgments

The authors thank the crew of the CCGS John P. Tully, the pilots of the ROV ROPOS, Sheryl Murdock and Catherine Stevens for the sample collection. MP also thanks Connor Bottrell for his help in parallelizing the CRISPR_distance algorithm. We are grateful to the two anonymous reviewers whose comments allowed us to improve the manuscript. This research would not have been possible without the amazing computing resources and excellent user support of ComputeCanada.

Chapter 4 – Third-generation sequencing reveals the adaptive role of the epigenome in three deep-sea polychaetes

Maeva Perez[1,2,3*], Oluchi Aroh[4], Yanan Sun[5], Yi Lan[1,2], Kim Juniper[6], C. Robert Young[7], Bernard Angers[3] Pei-Yuan Qian[1,2*]

[*]Corresponding authors
[1] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China
[2] Department of Ocean Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
[3] Department of Biological Sciences, Université de Montreal, Montréal, Canada
[4] Department of Biological Sciences, Auburn University, Auburn, AL., USA
[5] Laboratory of Marine Organism Taxonomy and Phylogeny, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China
[6] School of Earth and Ocean Sciences, University of Victoria, Victoria, Canada
[7] National Oceanography Center, Southampton, UK

Status:

Published in Molecular Biology and Evolution in 2023. doi: 10.1093/molbev/msad172

Contributions:

I conceived the study and conducted most of the laboratory work and data analyses under the mentorship of Yi Lan, Bernard Angers and Pei-Yuan Qian. Oluchi Aroh detected LTR-retrotransposons in the genome of the three polychaete species intact using a pipeline she has previously developed for the tubeworm *Lamellibrachia luymesi* (Aroh and Halanych 2021). Yanan Sun provided the genome of *Paraescarpia echinospica* (Sun *et al.* 2021), the corresponding raw Nanopore data and additional information (e.g. species-specific repeat databases, annotations files etc.). Kim Juniper supplied the Endeavour samples and together with Robert Young, Bernard Angers and Pei-Yuan Qian provided support for funding acquisition. I wrote the manuscript with advice and edition from all co-authors.

Résumé

Les rôles fonctionnels de la méthylation de l'ADN sont peu compris notamment pour l'embranchement des Annélides pour lequel on n'a que très peu de données. Nous comblons ce manque de connaissances en menant le premier sondage du méthylome complet chez trois espèces de polychètes des profondeurs qui dominent les communautés des sources hydrothermales et suintements d'hydrocarbures du Pacifique: *Paraescarpia echinospica*, *Ridgeia piscesae* et *Paralvinella palmiformis*. Après le séquençage et l'assemblage de génomes de haute qualité pour ces espèces, nous observons que ceux-ci codent pour toutes les enzymes clés du métabolisme de la méthylation de l'ADN. Ces vers possèdent également un méthylome en mosaïque similaire à celui des autres invertébrés. Des données de transcriptomiques additionnelles soutiennent les hypothèses selon lesquelles la méthylation des corps de gènes renforce l'expression des gènes essentiels, et la méthylation des promoteurs inhibe l'expression génique. En revanche, nos résultats ne supportent pas l'hypothèse qui suggère que l'expression des éléments transposables est inhibée par la méthylation. Les profils épigénétiques conservés des gènes responsables du maintien de l'homéostasie sous des conditions de pression hydrostatique extrêmes suggèrent néanmoins que la méthylation de l'ADN joue un rôle adaptatif important chez ces vers.

Abstract

The roles of DNA methylation in invertebrates are poorly characterised, and critical data are missing for the phylum Annelida. We fill this knowledge gap by conducting the first genome-wide survey of DNA methylation in the deep-sea polychaetes dominant in deep-sea vents and seeps: *Paraescarpia echinospica*, *Ridgeia piscesae* and *Paralvinella palmiformis*. DNA methylation calls were inferred from Oxford Nanopore sequencing after assembling high-quality genomes of these animals. The genomes of these worms encode all the key enzymes of the DNA methylation metabolism and possess a mosaic methylome similar to that of other invertebrates. Transcriptomic data of these polychaetes support the hypotheses that gene body methylation strengthens the expression of housekeeping genes and that promoter methylation acts as a silencing mechanism but not the hypothesis that DNA methylation suppress the activity of transposable elements. The conserved epigenetic profiles of genes responsible for maintaining homeostasis under extreme hydrostatic pressure suggest DNA methylation plays an important adaptive role in these worms.

Introduction

DNA methylation is a form of epigenetic control, which integrates genetic, environmental, and stochastic cues (Angers *et al.* 2020). At the cellular and organismal level, the variation in DNA methylation patterns upon a given genome facilitates cell differentiation and acclimation. At the individual and population levels, differences in epigenetic makeup have ecological and evolutionary repercussions. Among the known forms of DNA methylation, the addition of a methyl group to the 5th cytosine's carbon (5mC) is common in animals and the best studied, particularly in vertebrate models. In vertebrates, cytosine methylation in the CpG context (*i.e.* adjacent cytosine and guanine in the 5'-3' direction) leads to gene silencing by preventing the binding of the transcriptional machinery to the DNA (Lyko 2018), whereas the roles of 5mCpG methylation in invertebrates remain poorly understood. The possible roles for DNA methylation in invertebrates include gene expression regulation, alternative splicing, or silencing of transposable elements (TEs) (Flores *et al.* 2012; Gavery and Roberts 2014). Furthermore, the genomes of vertebrates are fully methylated with only a few hypo-methylated regions generally located in gene promoters (Suzuki and Bird 2008), whereas the genomes of invertebrates display wide variations among species in terms of how much and where DNA methylation occur. Such contrasts in methylation patterns between vertebrates and invertebrates suggest different functions and evolutionary histories (de Mendoza *et al.* 2020).

The functions and origin of DNA methylation can be better understood by examining more species from different invertebrate phyla. Comparative epigenomic analyses at the species and population level can provide valuable insights into their adaptative strategies and physiological resilience (McCaw *et al.* 2020). Phylum Annelida is an invertebrate taxon for which epigenetics data are remarkably scarce. These segmented worms are amongst the most diverse and abundant animals in marine ecosystems, occupy a wide variety of niches from soil and marine sediments and from deposit feeders to pelagic and parasitic lifestyles, and provide important ecological services such as bioturbation (Kvist and Oceguera-Figueroa 2021). So far there has been no complete genome-wide survey on annelids methylomes. DNA methylation that has been detected in a few annelid species (Newbold *et al.* 2019; Bicho *et al.* 2020; Ogunlaja *et al.* 2020; Planques *et al.* 2021) responded to environmental shifts, indicating its possible role in acclimation and resilience (Marsh and Pasqualone 2014). Furthermore, DNA methylation of their distant cousins in the phyla

Nematoda and Arthropoda presents striking inter-specific variation. For instance, insect genomes display much lower levels of DNA methylation (< 10%) than those of Crustaceans and Arachnids (~30%) (Provataris *et al.* 2018). Amongst nematodes, some taxa have completely lost DNA methylation and its associated machinery (e.g. *Caenorhabditis elegans*) (de Mendoza *et al.* 2020). An epigenetic diversity similar to that of Arthropods and Nematodes may exist amongst annelid species.

Deep-sea chemosynthetic ecosystems (e.g., hydrothermal vents, hydrocarbon seeps, and organic falls) are characteristically unstable and typified by extreme gradients of the temperature, concentration of chemicals, and food resources. In hydrothermal vents, the environmental conditions can drastically change in a span of a few centimetres and within minutes (Cuvelier *et al.* 2011; Lee *et al.* 2015) whereas in hydrocarbon seeps and organic falls, the conditions are relatively stable but are characterised by strong spatial zonation (Zhao *et al.* 2020). Endemic polychaetes that inhabit these ecosystems belong to three main families, namely, *Siboglinidae*, *Alvinellidae* and *Polynoidae*. Siboglinids are sessile gutless tubeworms that rely on intracellular chemolithotrophic bacteria for their nutrition, whereas alvinellids and polynoids are mobile and occupy higher trophic levels. The diversity of their phylogenetic and life history traits as well as the acute variability of conditions in their habitats make these segmented worms exceptional models for the investigation of how species use different epigenetic traits to cope with variable and unpredictable environments. Accordingly, solid baselines for the epigenomes of these animals need to be established.

Breaking ground on the study of deep-sea adaptive epigenomics, the present study provides the first genome-wide methylome survey of three deep-sea polychaete worms, namely, the siboglinids *Paraescarpia echinospica* and *Ridgeia piscesae* and the alvinellid *Paralvinella palmiformis* (Figure 4.1). New genomes for *R. piscesae* and *P. palmiformis* were assembled. These two species inhabit hydrothermal vents in the north-eastern Pacific Ocean and occupy a wide range of temperature conditions (2–60 °C) (Dilly *et al.* 2012; Tunnicliffe *et al.* 2014). While the former is mobile, the latter displays a great degree of phenotypic plasticity associated with the different environments it settles, suggesting a unique epigenetically-driven polyphenism in this species (Tunnicliffe *et al.* 2014). By contrast, *P. echinospica*, which is closely related to *R. piscesae*, lives

in hydrocarbon seeps in the western Pacific Ocean, does not display much morphotypic variation, and occupies a relatively narrower range of environmental conditions (Sun *et al.* 2021).

Our epigenomic surveys were conducted using data derived from Oxford Nanopore Technologies (ONT). The putative DNA methylation metabolism and genome-wide somatic 5mCpG methylation landscapes of *P. echinospica*, *R. piscesae* and *P. palmiformis* were characterised within the context of invertebrate methylomics. The three following hypothesis on the putative roles of DNA methylation were tested: 1) the methylation of TEs inhibit their activity, 2) DNA methylation located within gene bodies (i.e. the transcriptional region which includes both introns and exons) positively affects their expression and 3) promoter methylation acts as a gene silencing mechanism. Lastly, we compared the epigenetic profiles of orthologous genes across the three species to identify putative epigenetic adaptation their deep-sea environment. Our results suggest that acclimation and adaptation to the deep-sea environment could arise or take effect at the epigenetic level.

Figure 4.1 Species investigated in this study. Left panel: Colony of *P. echinospica* at the Haima colds seep in the South China Sea (adopted from Sun *et al.* [2021]), **Right panel**: Co-localising colonies of *R. piscesae* and *P. palmiformis* at Endeavour hydrothermal vents in the north-eastern Pacific Ocean (photo courtesy of Canadian Scientific Submersible Facility/Remotely Operated Platform for Ocean Sciences).

Material and Methods

Sample collection, nucleic acid extraction and sequencing

R. piscesae and *P. palmiformis* specimens were collected together from the same vent field on the Juan de Fuca ridge (47°57.00' N, 129°5.82' W, 2196m) with the ROV ROPOS in June 2016 during a cruise of the Canadian Healthy Oceans Network on board the CCGS John P. Tully. *P. echinospica* was collected from the Haima cold seep in the South China Sea (16°43.80′ N, 110°28.50′ E, 1390 m) (Sun *et al.* 2021). *R. piscesae* and *P. echinospica* were frozen at −80 °C immediately after reaching the vessel deck, while *P. palmiformis* was kept in 95% ETOH.

DNA from the vestimentum of *R. piscesae* and the body wall of *P. palmiformis* was extracted via phenol-chloroform and ethanol precipitation (Sambrook *et al.* 1989) and purified using the Genomic DNA Clean and Concentrator-10 kit (ZYMO Research, Ca, USA) according to the manufacturer's protocol. DNA from *P. echinospica* was extracted using the MagAttract HMV DNA kit (Qiagen, Hilden, Germany). RNA was extracted with TRIzol reagent from four tissues of *R. piscesae* (gills, vestimentum, epiderm, trophosome) and five tissues of *P. echinospica* (gills, collar, vestimentum, trophosome, opistosome). For *P. palmiformis*, the publicly available whole-body transcriptome of another individual (NCBI BioSample: SAMN14351933) was used.

DNA and RNA samples were sequenced by Novogene (Beijing, China) on the Illumina NovaSeq 6000 platform. Paired-end libraries with insert size of 350 bp were constructed using the NEBNext® DNA Library Prep Kit. DNA samples were also sequenced on the ONT PromethION platform. All genomic libraries reached a nuclear genome coverage of more than 85X, except for the Nanopore library of *P. palmiformis*, which reached approximately 44X (Figure S4.1 and Table S4.1).

CpG methylation calling

Nanopolish v0.13.3 (Simpson *et al.* 2017) pipeline was used under default parameters to detect CpG methylation in Oxford ONT data, because it offers the best compromise between accuracy and sensitivity of methylation detection and computational efficiency in several benchmarking studies (Liu *et al.* 2021; Yuen *et al.* 2021). Nanopolish uses a pre-trained hidden Markov model to assign methylation log-likelihood ratios (LLRs) to all CpGs within a 10 bp window. Briefly,

nanopore reads were indexed and mapped into the reference genomes. Then, the raw electric signatures (or squiggles) were fetched and aligned back onto their respective mapped reads (Figure S4.2). The likelihoods of each CpG-containing 6-mer given a five-base alphabet (with A, T, G, C and M for methylated cytosines) were then calculated, and CpG call groups (sequences containing one or multiple CpG motifs within 10bp of each other) were considered methylated if their log-likelihood of being methylated was twice that of being unmethylated (LLR>2). CpG call groups containing more than one CpG motif represented 26%–32% of all calls. The mean group coverage and methylation frequency were attributed to all CpGs within these groups. Accordingly, a bed file of all CpGs for each genome was generated and mapped to the respective methylation annotations (call coverage and frequency) onto them by using the bedtools map function (Quinlan and Hall 2010).

The precision of DNA methylation calls derived from Oxford Nanopore sequencing was assessed by comparing the whole-genome DNA methylation estimates of the same DNA extraction of the vestimentum of one *R. piscesae* individual that was obtained through Nanopore sequencing (sequencing depth >100X) to those obtained via whole genome bisulfite sequencing (WGBS; sequencing depth >85X). The WGBS method description is available in the supplementary material. The effect of sequencing effort was tested by artificially reducing the depth of coverage of the Nanopore dataset to 20X and 10X (Figure S4.3). Nanopore-derived methylation calls closely matched those of WGBS even at a low sequencing depth (Figure S4.4-S4.6, Table S4.2 and Table S4.3). The increasing affordability of DNA sequencing has led to numerous genome sequencing projects, which make use of third-generation sequencing technology to achieve high assembly contiguity (Figure S4.7). These data can be reinvestigated for DNA methylation. The optimal accuracy and sensitivity was ensured by applying thresholds on the methylation call coverage (10X for *P. echinospica* and *R. piscesae* and 1X to *P. palmiformis* because of its low sequencing depth), and the window size used for averaging methylation calls over genomic regions was set to 1,000 bp.

Genome assemblies and annotations

The annotated genome assembly for *Paraescarpia echinospica* was based on the study of Sun *et al.* (2021). For *R. piscesae* and *P. palmiformis*, new genome and transcriptome assemblies were constructed and annotated following a similar protocol to that of Sun *et al.* (2021). The detailed

methods for these genome reconstructions are presented as supplementary material and summarised below.

Contig-level assemblies were produced for the three genomes by using a combination of short (Illumina) and long (ONT) reads. For each genome, multiple assembly pipelines were used, and the assembly with the best completeness and contiguity statistics was retained (Figure S4.8 and Table S4.4). Putative misassemblies were avoided by removing reads from extra-nuclear compartments from the libraries of *R. piscesae* and *P. palmiformis* by mapping them against the reference mitochondrial (KJ872501 [Jun *et al.* 2016], OL802212 [Perez *et al.* 2022]) and endosymbiont (LDXT01 [Perez and Juniper 2016]) genomes.

Prior annotations, repeat regions and putative TEs were masked using a custom repeat database composed of known repeat motifs and sequences issued from the universal databases RepBase27.02 (Bao *et al.* 2015) and dfam3.3 (Storer *et al.* 2021), and species-specific repeats were detected using RepeatModeler (v2.0.1) (Chen 2004). Long terminal repeat retrotransposons were detected following Aroh and Halanych (Aroh and Halanych 2021) (see Supplementary Methods). Gene model inferences were obtained using Maker2 (Holt and Yandell 2011) by integrating (1) ab-initio gene predictions, (2) species-specific transcriptomes and (3) additional proteomic evidence from closely-related species. The ab initio gene predictions were performed using Augustus, and species-specific sets of modelling parameters were obtained by training the algorithm with a set of high-confidence gene models based on transcriptomic and proteomic evidence. The transcriptomic evidence included transcriptomes assembled from the same individuals that produced the reference genomes (*P. echinospica* and *R. piscesae*) and additional assemblies from publicly available databases (*R. piscesae* and *P. palmiformis*). These transcriptomes were assembled with Trinity (v2.11.0) (Grabherr *et al.* 2011) in de novo and genome-guided (when reference genome was available) modes (Figure S4.9). Proteomic evidence was obtained from the siboglinid species *Riftia pachyptila* (de Oliveira *et al.* 2022), *Lamellibrachia luymesi* (Li *et al.* 2017; Li *et al.* 2019) and *Escarpia spicata* (Li *et al.* 2017) for *R. piscesae* and the alvinellid species *Alvinella pompejana* (Gagnière *et al.* 2010), *Paralvinella hessleri* (Wang *et al.* unpublished) and *Paralvinella grasslei* (Stiller *et al.* 2020) for *P. palmiformis* (Table S4.5). The exon-level annotation edit distance (eAED) was superior to 0.5 for more than 85% of the final gene models (Figure S4.10) with complete and single-copy BUSCO (Simão *et al.* 2015) scores >85% against the metazoan database (Table 4.1 and Table S4.6). The final gene models were blast-searched against NCBI's nr database

and screened for functional annotation against the Gene Ontology (GO), Protein families (Pfam), EuKaryotic Orthologous Groups (KOG), Kyoto Encyclopedia of Genes and Genomes (KEGG), Enzyme Commission (EC) and The Institute for Genomic Research curated protein families (TIGRFAM) databases (Figure S4.11). Protein blast searches against the gene models of *P. echinospica* and *R. pachyptila* were further conducted for *R. piscesae* (Figure S4.12). Finally, genes associated with the methylation metabolism were further identified amongst the gene models and de novo transcriptomes through their specific functional domains and homology to protein sequences in UniProt's Swiss-Prot and TrEMBL databases (blastp searches).

Methylome annotations

The per-CpG methylation information was averaged over the genomic annotations (and vice versa) by using the bedtools map function (Quinlan and Hall 2010). A detailed description of these steps is available in the Supplementary Methods. For each feature, methylation coverage was defined as the fraction of called CpGs for which methylation was detected, and methylation depth represents the fraction of reads for which methylation was detected amongst methylated CpGs . The parameters are related by the following expression: mean methylation frequency = methylation coverage × methylation depth. Features were also annotated as highly methylated (or hypermethylated) at methylation coverage > 75% and methylation depth > 50% according to the observed density clustering (Figure S4.13).

Gene expression quantification and protein ortholog identification

The expression of gene model mRNA sequences and de novo transcriptome contigs were quantified with Salmon (Patro *et al.* 2017) by using their corresponding transcriptomic libraries. Protein orthologs were identified with Proteinortho (Lechner *et al.* 2014) by using DIAMOND (Buchfink *et al.* 2021) (parameters: -e = 1e-06 –sim = 1 –identity = 40 –cov = 50) and the `--synteny` option.

Phylogenetic and statistical analyses

Phylogenetic trees for the *Mat* gene were reconstructed using Bayesian inference (BI) and maximum likelihood (ML) methods from codon alignments of polychaetes using a the General

Time-Reversible (GTR) substitution model. These alignments included additional Siboglinidae (Moggioli *et al.* 2023), Sabellida (Tilic *et al.* 2020) and Terebellida (Stiller *et al.* 2020) species. Relaxed selection in *Mat*-b compared to *Mat*-a paralogs was tested with RELAX (Wertheim *et al.* 2015) while episodic diversifying selection was detected with aBSRel (Smith *et al.* 2015). Detailed methods for the phylogenetic and selection analyses are presented in the supplementary materials. The multiple sequence alignment files are available on our Github page.

Statistical analyses were performed in R (Ihaka and Gentleman 1996). The significance of gene expression differences between methylation group pairs was assessed using Kolmogorov-Smirnov (KS) tests (function ks.test from the stats package v4.1.1) at a p-value threshold of 0.05. Hypergeometric tests against 100 random subsamples were used to test whether the distribution of functional categories in a specific gene subset was representative of the whole. Fisher tests followed by Holms correction were used to detect the overrepresentation of certain functional categories within gene subsets. Partial correlations (function pcor.test from the package ppcor v1.1) were used to control for the confounding factors of gene length and GC content for the comparison of promoter methylation and gene expression and epigenomic context for the comparison of LTR-retrotransposon methylation level and estimated insertion time. Pairwise Pearson correlations for LTR-retrotransposon variables were computed using the function rcorr from the package Hmisc (Harrell 2019). The strength of the epigenetic/genetic signature coupling was quantified by computing the goodness of fit of the logistic regression of the methylation-driven mutational bias (CpG observed/expected) across two observed methylation categories (high and low-intermediate); confidence intervals were estimated through 1,000 bootstrap resampling.

Data availability

Raw sequence reads are available in NCBI's SRA database under the accession numbers SRR21707426–SRR21707428 (*P. palmiformis* Illumina), SRR217033–SRR217038 (Nanopore reads of *R. piscesae* and *P. palmiformis*) and SRR21707446–SRR21707448 (*R. piscesae* Illumina). The Whole Genome Shotgun projects for *P. echinospica*, *R. piscesae* and *P. palmiformis* were deposited in GenBank under accession numbers JAHLWY000000000, JAODUO000000000, and JAODUP000000000. The genome versions described in this paper are JAHLWY010000000, JAODUO010000000 and JAODUP010000000. The methylomes are available in NCBI's GEO

database under the SuperSeries accession GSE217309. Additional files can be accessed at https://github.com/maepz/deepsea_worms_epigenomes/data/.

Code availability

All scripts used for the upstream data processing steps and downstream analyses can be accessed at https://github.com/maepz/deepsea_worms_epigenomes.

Results

Methylation metabolism

The methylome of *P. echinospica*, *R. piscesae* and *P. palmiformis* was characterised by reconstructing high-quality genomes for these species. Based on the previously published genome of *P. echinospica* (Sun *et al.* 2021), the genomes of *R. piscesae* and *P. palmiformis* were sequenced and assembled using a combination of short (Illumina) and long (Nanopore) reads derived from the same DNA extractions (see detailed genome information in Table 4.1). The contig-level assembly of *R. piscesae* from this study is the most contiguous to date (Wang *et al.* 2023) and that of *P. palmiformis* is the first representative genome for the family Alvinellidae. Our conservative genome annotation protocol, in which only transcriptomic data from their respective species and that of species from the same family were used, resulted in the identification of 22,642 and 31,703 high-quality protein-coding gene models (complete and single copy BUSCO score > 85%) for *R. piscesae* and *P. palmiformis*, respectively (Table 4.1), while 24,682 protein-coding genes were obtained for *P. echinospica*.

Enzymes essential to the DNA methylation metabolism were detected in all three species (Figure 4.2, Table S4.7-Table S4.14). The transcripts for almost all these enzymes were found in the de novo transcriptomes of worms, with some exceptions in *R. piscesae* whose transcriptome was of the lowest quality among the three worm species (see supplementary methods). Sequences were in some cases fragmented or truncated in either the transcriptomic assemblies (Figure S4.14) or the gene models (Figure S4.15) but not in both, which suggests these cases were resulted from assembly and annotation issues. We were not able to resolve these issues with our current data.

The genomes of *P. echinospica*, *R. piscesae* and *P. palmiformis* encode genes for DNMT1 (domain architecture: DMAP-RFD-zfCXXZ-BAH-DNA methylase), which is the DNA methyltransferase enzyme that is principally responsible for propagating cytosine methylation marks in daughter cells after mitosis, DNMT3 (domain architecture: PWWP-ADD-DNA methylase), which is the de novo cytosine DNA methyltransferase gene, and the methylcytosine dioxygenase TET (domain architecture: CXXC-cysteine rich region-catalytic domain), which is responsible for DNA demethylation. Partial sequences for these enzymes were found in the species de novo transcriptomes (Table S4.7-Table S4.9), but their expression levels were low. DNMT1, DNMT3, and TET transcripts were also detected in the RNA-Seq data of closely related species (tubeworms *Riftia pachyptila* and *Lamellibrachia luymesi* and alvinellid *Paralvinella grasslei*).

Our gene survey of the worm genomes further highlighted putative differences in the DNA methylation metabolism between the siboglinid and alvinellid worm taxa. Tubeworms possess two enzymes for the remethylation of homocysteine to methionine, the B-12 dependent methionine synthase (MTR) and the betaine homocysteine methyltransferase (BHMT), whereas only MTR was detected in *P. palmiformis* (genome and transcriptome) and *P. grasslei* (transcriptomes).

Lastly, two genes for the methionine adenosyltransferases (MATa and MATb) were found in both the tubeworms and alvinellid species (Table S4.4). Recent genetic and transcriptomic data revealed two versions of MAT also exist in the Alvinellidae's sister group Amphraetidae (Stiller *et al.* 2020), Osedax (Moggioli *et al.* 2023) and some species of the Terebellidae family (Figure S4.16). *Mat* paralogs were found in tandem in the vestimentiferans but not in *Osedax* or *P. palmiformis*'s genomes. Evidence of relaxed selection was only found in the *Mat*-b paralogs of vestimentiferans (Table S4.15). No evident episodic diversifying selection was detected across the worm's phylogeny. Taken together, these results suggest the *Mat* gene may have duplicated multiple times in polychaetes, including one recent duplication in vestimentiferans, but in absence of a fully resolved phylogeny, the evolutive history of this gene remains uncertain.

Table 4.1 Summary of contig-level assembly statistics and gene model annotations for all available deep-sea polychaete genomes. BUSCO scores are based on gene models protein searches against the metazoan database (odb10). *This study; [a]Sun *et al.* (2021), [b]Li *et al.* (2019), [c]de Oliviera *et al.* (2022), [d]Wang *et al.* (2023), [e]Patra *et al.* (in prep), [f]Moggioli *et al.* (2023),[g]data available at https://phaidra.univie.ac.at/detail/o:1220865, [h]data available at https://github.com/ChemaMD/OsedaxGenome ; n.a : not available.

	***P. echinospica**[a]*	***R. piscesae****	***P. palmiformis****	***L. luymesi**[b]*	***R. pachyptila**[c]*	***R. piscesae**[d]*	***L. satsuma**[e]*	***R. pachyptila**[f]*	***O. alvinae**[f]*	***O. frankpressi**[f]*
NCBI's WGS accession	JAHLWY01	JAODUO01	JAODUP01	SDWI01	n.a.[g]	JALOCR01	JAHXPS01	n.a.[h]	n.a.[h]	n.a.[h]
Contig-level assembly summary										
Tot. length (Mbp)	1,095	658	601	602	561	530	665	554	808	285
# of contigs	12,702	9,128	3,622	55,092	447	113,548	6,320	926	785	1,194
Max. contig length (Kbp)	1,903	2,546	3,777	274	13,119	174	1,857	8,302	10,682	2,139
contig N50 (Kbp)	243	365	557	24	2,870	10	267	1.422	2,901	426
GC content	41%	41%	36%	40%	41%	41%	40%	41%	41%	29%
# of CpGs	35,178,402	20,589,468	14,611,299	17,810,943	17,346,573	16,435,603	19,045,649	16,783,278	28,935,539	4,591,061
Annotation summary										
Coding genome tot. length (Mbp)	26	33	32	42	59	n.a	39	55	59	29
# of protein-coding genes (without isoforms)	22,642	31,703	24,682	40,316	25,983	24,096	33,184	37,037	37,777	18,657
Tot. BUSCOs score	88%	94%	96%	98%	99%	93%	91%	96%	97%	91%
Complete and single BUSCOs score	79%	86%	87%	91%	43%	n.a	90%	91%	90%	87%

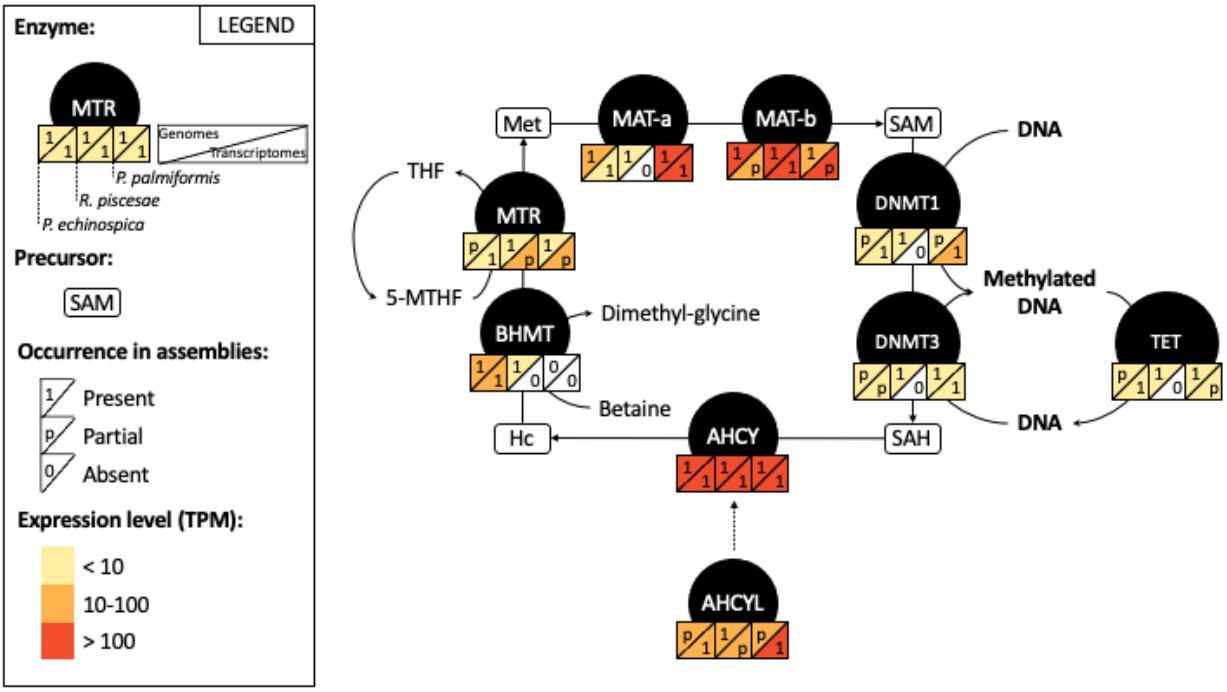

Figure 4.2 Simplified schematic representation of the DNA methylation metabolism that shows the occurrence and expression level of essential enzymes in the studied worms' genomes and de novo transcriptomes. 5-MTHF: 5-methyltetrahydrofolate, AHCY:S-adenosyl-homocysteine hydrolase, AHCYL: S-adenosyl-homocysteine hydrolase like, BHMT: betaine homocysteine methyltransferase, DNMT: DNA methyltransferase, Hcy: Homocysteine, MAT: methionine adenosyltransferase, Met: Methionine, MTR: B12-dependent methionine synthase, SAH: S-adenosyl-homocysteine, SAM: S-adenosyl methionine, TET: methylcytosine dioxygenase, THF: tetrahydrofolate.

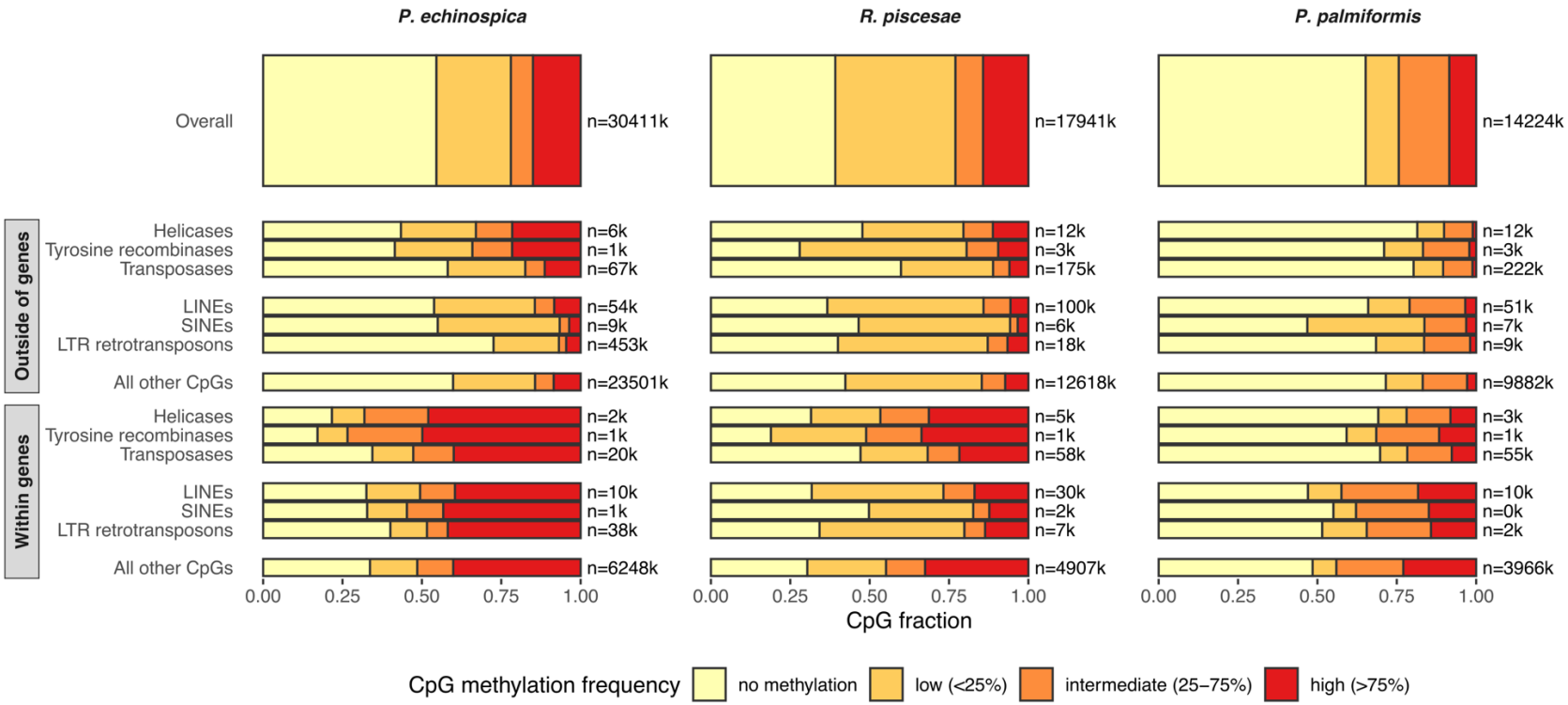

Figure 4.3 Genome-wide methylation frequency spectra in three polychaete species and CpG methylation on genetic sequences identified as mobile elements by RepeatMasker outside and within genes. The number of CpGs assessed is indicated on the right. CpGs with less than 10× coverage were only considered in P. palmiformis because of its low sequencing depth. DNA transposons (helicases, tyrosine recombinases and transposases) and retrotransposons (LINEs, SINEs and LTR-retrotransposons) were grouped separately.

Genome-wide methylation profile

Considering the high accuracy of nanopore sequencing data (see supplementary methods), they were used to detect 5mC methylation in CpG context (5mCpG) in *P. echinospica*, *R. piscesae* and *P. palmiformis*. Methylation calls on CpG motifs covered by less than 10 reads were only considered for *P. palmiformis* because of their low sequencing depth. After coverage-based filtering, the remaining CpGs called for methylation represented 86%, 87%, and 97% of all CpGs present in the genome assemblies of *P. echinospica* (35,178,402 CpGs), *R. piscesae* (20,589,468 CpGs) and *P. palmiformis (*14,611,299 CpGs), respectively.

The genomes of *R. piscesae*, *P. echinospica,* and *P. palmiformis* showed similar methylation profiles which are characterised by relatively low levels of CpG methylation (Figure 4.3). Global CpG methylation in these worms was higher than that of insects but well within the range reported for other arthropod and nematode species. Methylation was present on 35%–61% of CpGs. A greater proportion of highly methylated CpGs was found in both tubeworms (~15%) than in *P. palmiformis* (8%). In all three polychaetes, CpGs within genes were more methylated than those in intergenic regions (Figure 4.3). Within genes, methylation was more polarized on exons (Figure S4.17) than on introns with the first exon of genes being typically unmethylated (Figure S4.18).

Transposable elements and LTR-retrotransposons methylation

The analysis of the mobilome of *P. echinospica*, *R. piscesae* and *P. palmiformis* with RepeatMasker (Chen 2004) identified genetic sequences that belong to DNA transposons (helicases, tyrosine recombinases and transposases), SINE and LINE retrotransposon elements. A more stringent bioinformatic pipeline annotated intact and well-defined long terminal repeat (LTR) retrotransposon (Aroh and Halanych 2021). LTR-retrotransposons characteristically possess two identical sets of long terminal repeats, which independently accumulate mutations after their integration in the genome. Thus, insertion times were estimated using the sequence divergence between these paralogs.

CpG methylation levels on mobile elements matched those of their flanking regions (i.e. genomic context) with higher methylation on intragenic elements than intergenic ones (Figure 4.3). A total of 2,019, 146 and 84 intact LTR-retrotransposons were detected in *P. echinospica*, *R. piscesae* and

P. palmiformis, respectively (Figure S4.19). DNA methylation level (i.e. mean methylation frequency) and insertion time (in Mys) were estimated for most of these LTR-retrotransposons (1984, 141 and 84 in *P. echinospica*, *R. piscesae* and *P. palmiformis*, respectively). Contrary to our expectations, younger LTR-retrotransposons were not more methylated than older ones (Figure 4.4). In fact, when controlling for genomic context, insertion time and methylation load were positively correlated in *P. echinospica* (Table S4.16). However, a strong association was observed between insertion time and CpG depletion, indicating that in *P. echinospica*, the mutation rate on methylated LTRs was considerably higher than on unmethylated ones and may have biased our insertion time estimates (Figure S4.20). Still, highly methylated LTR-retrotransposons were hypermethylated compared to their flanking regions and inversely for lowly methylated LTR-retrotransposons (Figure 4.4,Figure S4.21 and Figure S4.22, Table S4.17). The epigenetic state of lowly methylated LTR-retrotransposons was less dependent on the surrounding epigenetic context than highly methylated ones (Figure S4.20-Figure S4.22) but could not be consistently explained by the other factors tested (LTR-retrotransposon size, insertion time and location).

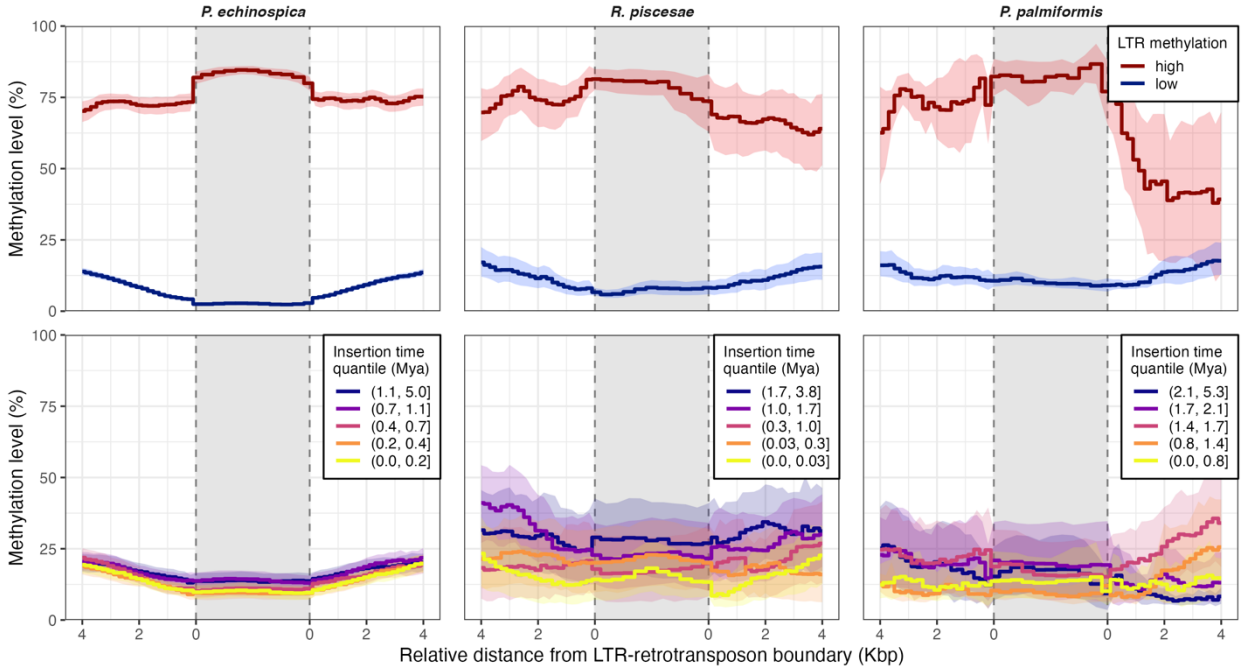

Figure 4.4 LTR-retrotransposons methylation and their genomic context. The average methylation frequency was computed for each LTR-retrotransposon and their 4Kbp upstream and downstream flanking regions (1 Kbp sliding window with a 200 bp step). The step-lines represent the dataset aggregated averages, and the ribbons represent the 95% confidence interval from a 1000 bootstrap resampling. LTR-retrotransposon sizes were normalised. High LTR methylation = mean coverage > 75% and mean depth > 50%, low methylation = mean coverage ≤ 75% and mean depth ≤ 50 %. Insertion times for each species are grouped in quantiles.

Gene body and promoter methylation

We found supporting evidence for the role of gene body methylation in the regulation of gene expression in the surveyed worms. The mean methylation frequency (methylation level) of genes was the highest in functional categories associated with housekeeping functions, such as translation and RNA processing, nuclear and ribosomal structure and cell cycle (Figure S4.23 Figure S4.24) and was positively correlated to expression (Figure 4.5 and Figure S4.25) even after controlling for the confounding effects of gene length and GC content (Figure S4.26 and Table S4.18 for results of partial Spearman correlation analyses).

Gene expression was also affected by promoter methylation. Genes characterised by low gene-body methylation had higher expression when their 4Kbp upstream region was heavily methylated (high U low G > low U G in Figure 4.5). The reverse relation was true amongst genes with high gene-body methylation (high U G < low U high G in Figure 4.5). These correlations held whether the methylomics and transcriptomic data were generated from the same individual and tissues (*P. echinospica*) or not (*R. piscesae* and *P. palmiformis*). Among genes with high gene-body methylation, the negative correlation between expression and upstream broke down with increasing distance to the TSS and mostly held when only the best genes models (i.e. the highly conserved metazoan genes identified by BUSCO [Simão *et al.* 2015] as complete in our genomes) were included in the analysis (Figure S4.27, Table S4.19 andTable S4.20). This indicates that our results were not overly confounded by misidentified promoters due to inaccurate gene models. Genes with unmethylated promoters were preferentially associated with protein folding, intracellular trafficking, and energy metabolism, whereas genes with methylated promoters were particularly enriched in signal transduction (in the tubeworms), replication, and repair (in *P. palmiformis*) functions (Figure S4.28 and Figure S4.29).

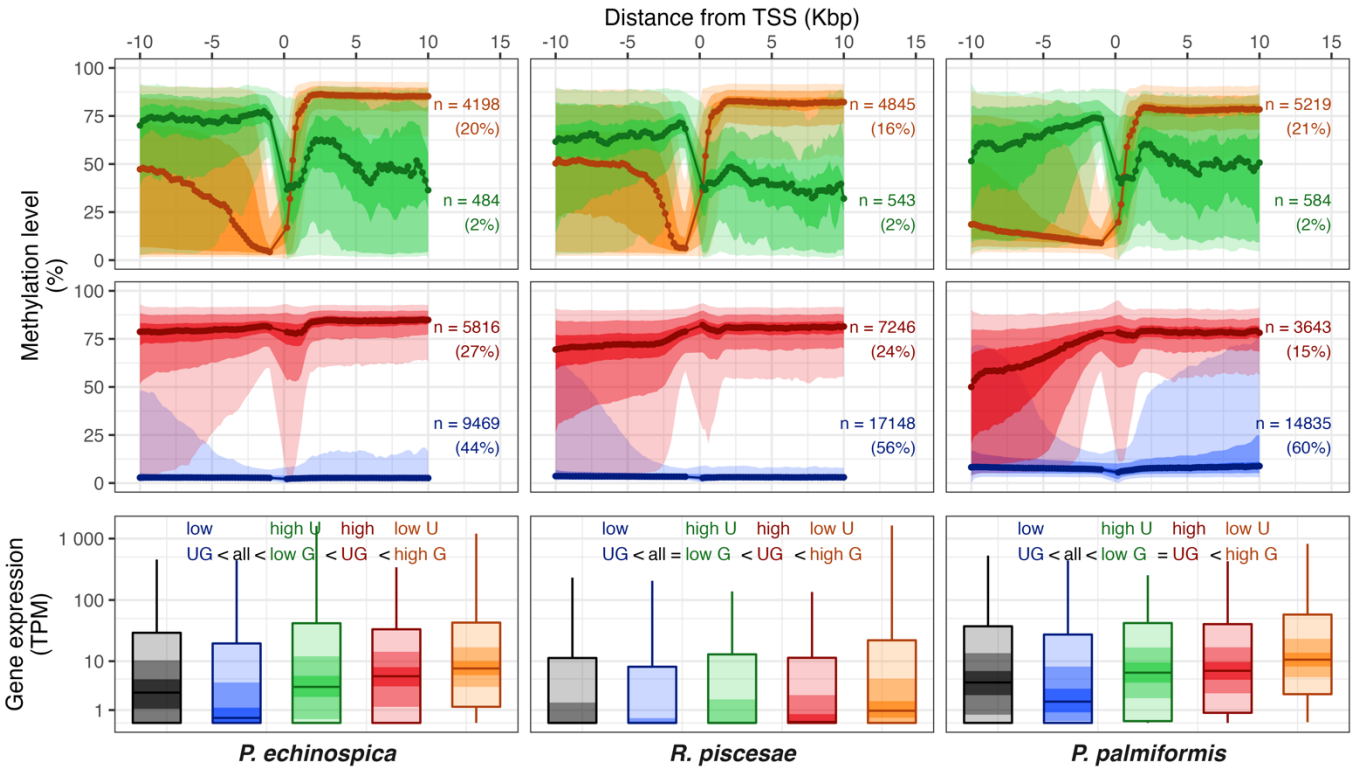

Figure 4.5 Methylation level around the transcription start site (TSS) of genes classified into four groups according to their 1Kbp upstream and gene methylation state. Genes with missing methylation data were not included. The bottom row represents the normalised gene expression (transcripts per million; TPM) for the gene groups (the whiskers extend to 95% of the data). The 'high U low G' group likely contains genes misclassified because of natural stochasticity or noisy methylation data. Moreover, *R. piscesae* and *P. palmiformis* transcriptomes include data from different individuals and different tissues than the ones surveyed for methylation, whereas *P. echinospica* transcriptome was obtained from the same individual and tissues as the epigenome. Differences in gene expression across groups were tested using pairwise KS tests (p-value threshold=0.05) and are presented above the boxes. Gene group description: low UG = low-intermediate upstream and gene methylation (mean coverage ≤ 75% and mean depth ≤ 50%); high U low G = high upstream methylation (mean coverage > 75% and mean depth > 50%) and low-intermediate gene methylation; high UG = high upstream and gene methylation; low U high G = low-intermediate upstream and high gene methylation. The methylation coverage and depth thresholds were selected to match density clusters in the data (Figure S4.13); n = group gene count.

Methylation profiles of orthologous genes

A total of 15,559 gene orthologs were identified across the three worm genomes, including 5,132 core orthologs represented by a single gene model in all three species. Gene body methylation could be characterised in all species for 4,942 of these reliable core orthologs (Figure S4.30). Among them, 1,185 had contrasting levels of gene-body methylation across species and 3,036 had high levels of DNA methylation in all species. The methylation state of gene promoters across species was also characterised for 2,906 of the 3,036 orthologs with conserved hypermethylated gene-bodies (Figure S4.30). Based on the comparison of the relative expression of orthologs in *P. echinospica* and *P. palmiformis*, the association of gene-body hypermethylation and promoter de-methylation with higher expression was confirmed (Figure S4.31, and Figure S4.32). Moreover, the methylation levels of promoters and gene-bodies were tightly matched by their respective methylation-driven mutational biases (CpG observed/expected) but not the gene length and GC content (Figure S4.33-Figure S4.36). The epigenetic and genetic signatures of gene-body and promoter methylation were more strongly coupled for genes with conserved methylation across species compared with genes having contrasting methylation profiles (Figure 4.6, Figure S4.37, and Figure S4.38). In addition, the coupling strength was correlated with the phylogeny, and it was stronger across worm families than amongst the two Siboglinid tubeworms. These results show that the methylation states of a large number of genes and, to a lesser extent, their promoters, are stable and persistent over evolutionary time. Genes with conserved gene-body hypermethylation did not represent a random subsample of all core orthologs (hypergeometric test p-value <0.01) and were most enriched in genes associated with translation and protein folding and degradation functions, particularly in ribosomal proteins, chaperones and proteins associated with the ubiquitin-proteasome pathway (Figure 4.7). Together with the genes of the energy metabolism (ATPase and NADH dehydrogenase subunits), orthologs associated with protein homeostasis were also over-represented amongst orthologs with conserved unmethylated promoters (Figure 4.7).

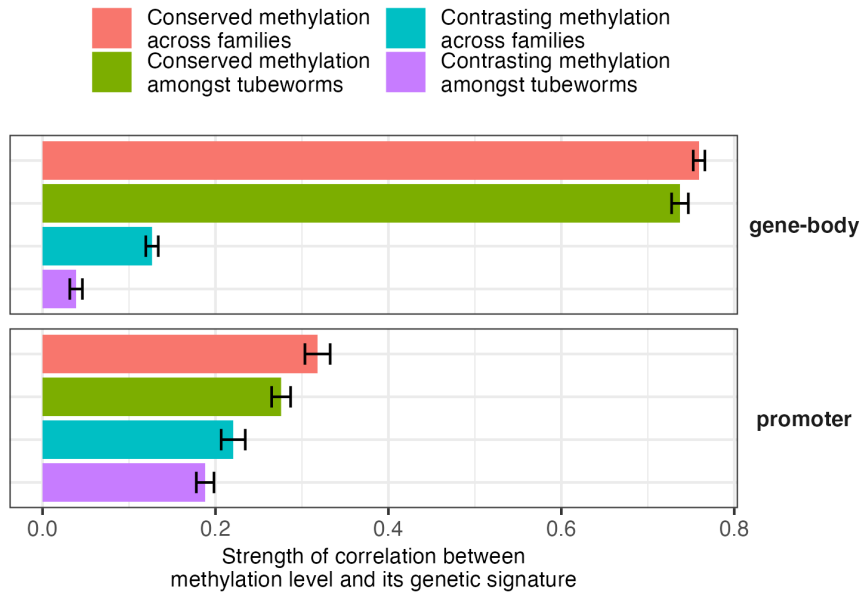

Figure 4.6 Stronger genetic signatures are associated with persistent methylation across taxa. For each gene group, the correlation strength was estimated by the goodness of fit of the logistic regression of CpG observed/expected on methylation level. The mean (bar height) and standard deviation (error bar) of 1,000 bootstrap samples are represented.

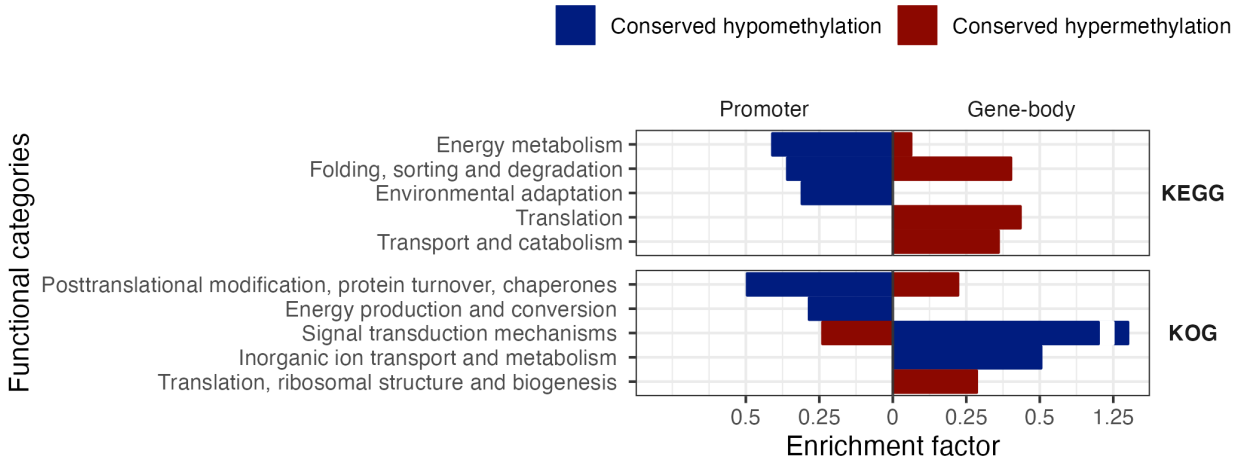

Figure 4.7 Most enriched functional categories amongst genes with conserved hypermethylated gene bodies and conserved promoter methylation. The top three functional categories (based on enrichment factor) for each of the gene-body and promoter gene sets are shown. The enrichment factor is defined as the normalised differences between expected and observed relative abundances of functional categories. The significance of the enrichment for each functional category was tested with Fisher tests followed by Holms correction. Non-significant enrichments (Fisher test p-value ≥ 0.05) are not shown.

Discussion

The methylome of three polychaete species (*P. echinospica*, *R. piscesae* and *P. palmiformis*) from deep-sea ecosystems was characterised for the first time. Based on the identification of the key genes of the DNA methylation metabolism in their genomes (two of which were newly assembled in the present study) and available transcriptomes, these worms possessed and expressed the machinery for DNA methylation, but interesting differences were observed across taxa. Considering that the DNA methylation estimates obtained from Nanopore data were as accurate as those recovered from WGBS at the genomic resolution of genes, this third-generation sequencing technology was used to describe the epigenomic landscape across the whole worm genomes.
Mobile genetic elements were not ubiquitously targeted by methylation, although a close examination of well-defined LTR-retrotransposons revealed differences in methylation load in comparison to their genomic context. Furthermore, our results strongly support the hypotheses that gene body methylation strengthens, and promoter methylation hinders gene expression in these worms. Indeed, the DNA methylation of gene bodies and promoters were positively and negatively correlated with expression. Lastly, we found evidence for the presence of conserved gene and promoter methylation profiles across the species which could reflect epigenetic adaptations to the deep-sea environment.

Siboglinids and alvinellids possess the complete and active machinery for setting up and removing DNA methylation

All key genes for the DNA methylation metabolism were present in the worms' genomes, although they were not ubiquitously expressed in the somatic tissues of adult worms. DNMT1, DNMT3, and TET were fragmented and poorly represented in the transcriptomic data of muscle and dermal tissues. In both vertebrates (e.g., cattle and *Medaka* fish) and invertebrates (e.g., oyster *Crassostrea gigas*, the model polychaete species *P. dumerilii*), these enzymes were predominantly expressed in oocytes and early embryos (Wang *et al.* 2014; Duan *et al.* 2019; Wang and Bhandari 2019; Planques *et al.* 2021). Our results are consistent with the observations made in other animals, in which most of the DNA methylation marks were removed during the gametogenesis and reset during the development (Smallwood and Kelsey 2012). Epigenetic marks set up during the development are suspected to contribute to the thermal acclimation of *R. piscesae* and *Paralvinella*

species (Dilly *et al.* 2012; Tunnicliffe *et al.* 2014). Future studies on the DNA methylation dynamics in gametogenesis and early development may provide more information on the polychaetes epigenetic landscapes of their ontogenic and adaptive roles.

Putative metabolic differences in DNA metabolism also exist across the two families. Unlike the alvinellids, tubeworms possess two enzymes for the remethylation of homocysteine to methionine (BHMT and MTR). BHMT homologs can be found in other invertebrates, including polychaetes, suggesting a secondary loss in alvinellids. Putative independent duplications of methionine adenosyltransferases (MAT) genes were also observed in tubeworms and alvinellids, respectively. This finding is remarkable, because MAT, which is responsible for synthesizing the main methyl donor S-adenosylmethionine is a highly conserved enzyme, and only a few paralogs have been discovered so far (Sanchez-Perez *et al.* 2008; Riesgo *et al.* 2012). The products of BHMT, MTR, and MAT are involved in several metabolic functions beyond the epigenetic metabolism (Finkelstein 1990; Fontecave *et al.* 2004; Stipanuk 2020) and in polychaetes, their specific roles still remain unclear.

Siboglinids and alvinellids possess intermediate levels of DNA methylation

The methylome surveys of the three deep-sea polychaetes revealed that they possessed intermediate levels of DNA methylation (35%-61% of methylated CpGs) similar to that of early-branching arthropods (de Mendoza *et al.* 2020). DNA methylation in the arthropod species *Stegodynphus dumicola*, *Strigamia maritima*, and *Procambarus virginalis*, ranges from 17 to 76% of methylated CpGs (de Mendoza *et al.* 2020) whereas it is typically present in less than 30% of CpGs in insects (< 10% in holometabola) (Provataris *et al.* 2018).

Transposable elements carry weak DNA methylation signatures

DNA methylation may act in silencing TE activity via RNA-directed targeting mechanisms (Deniz *et al.* 2019). TEs are categorised in two clades according to their mode of propagation. DNA transposons move from one genomic location to another as DNA segments, whereas retrotransposons propagate through RNA intermediates, which are reverse-transcribed into DNA (Kojima 2019). In invertebrates, the methylation state of TEs varies widely across element types and species (Lyko *et al.* 2010; Wang *et al.* 2014; Gatzmann *et al.* 2018; de Mendoza, Hatleberg, *et al.* 2019; Lewis *et al.* 2020; Ying *et al.* 2022), suggesting that the silencing of TEs by DNA

methylation is a modular function acquired multiple times in their evolutionary histories (de Mendoza *et al.* 2020). The analyses of mobile DNA methylation did not provide strong support for the hypothesis that TE proliferation is broadly repressed by DNA methylation in the polychaetes. The methylation levels of mobile elements were generally lower (> 85% hypomethylated CpGs TEs in the worms) than those of vertebrates (< 3% hypomethylated CpGs in humans [Pehrsson *et al.* 2019]) or plants (~25 % of hypomethylated TEs in *Arabidopsis* [Ahmed *et al.* 2011]) and matched that of their genomic context.

Well-defined intact LTR-retrotransposons were detected in the worms' genomes. Their abundance falls within the same range as that in other protostomes (Kim *et al.* 1994; de la Chaux and Wagner 2011; Thomas-Bulle *et al.* 2018; Aroh and Halanych 2021). Most LTR-retrotransposons (> 90%) belonged to the Gypsy superfamily, as previously reported for vestimentiferan tubeworms (Aroh and Halanych 2021; de Oliveira *et al.* 2022; Wang *et al.* 2023), molluscs (Thomas-Bulle *et al.* 2018), arthropods (Kim *et al.* 1994; Kaminker *et al.* 2002; Pelisson *et al.* 2002; Xu *et al.* 2005; Piednoël *et al.* 2013; de Mendoza, Pflueger, *et al.* 2019) and nematodes (de la Chaux and Wagner 2011). Nevertheless, a higher abundance of LTR-retrotransposons was observed in *P. echinospica* than that in the other worms, which appears to be correlated with its larger genome size (almost double that of the close-relative siboglinids *R. piscesae* [Wang *et al.* 2023], *L. luymesi* [Li *et al.* 2019], and *R. pachyptila* [de Oliveira *et al.* 2022; Moggioli *et al.* 2023]). The high abundance of LTR-retrotransposons in *P. echinospica* and their insertion time distribution suggests a recent wave of invasion that contributes to the unique genome expansion among species (Sun *et al.* 2021).

Younger and presumably more mobile LTR-retrotransposons did not display elevated methylation as observed in mammals (Barau *et al.* 2016) and cnidaria (Ying *et al.* 2022) but strong variation of DNA methylation load was observed in LTR-retrotransposons boundaries. Such contrasting methylation densities between TEs and their immediate surroundings have also been observed in other species (Lyko *et al.* 2010; Lewis *et al.* 2020; Ying *et al.* 2022) but are hard to interpret without a comprehensive dataset of invertebrate methylomes. This finding was obtained possibly because methylation density variation along the DNA sequence rather than absolute methylation level allows the recognition of certain transposable elements, but this hypothesis needs to be further tested.

Gene body methylation is associated with increased gene expression

Gene body methylation was observed in these three species and was positively correlated to transcript abundance, thus supporting that it strengthens gene expression. In rat embryos, Neri *et al.* (2017) demonstrated that the high density of intragenic methylation marks induced by the de novo DNA methyltransferase 3b (DNMT3b) reduces spurious transcription initiation, thus increasing transcription yield, by preventing the aberrant intragenic binding of the RNA polymerase. Although a similar mechanism of transcription stabilisation in invertebrates has not been determined, the presence of DNMT3 (a homolog to DNMT3b) in the worms' genomes supports the hypothesis that gene bodies are specifically targeted by the DNA methylation machinery. DNMT gene knockout experiments have shown that somatic gene expression in plants and insects was not necessarily affected by gene body methylation (Zhang *et al.* 2006; Bewick *et al.* 2019). Similar experimental approaches in other model polychaete species are needed to assess if DNA methylation exerts direct epigenetic control on gene expression in this group. Since deep-sea worms have not been successfully reared in the laboratory, it will be very hard to conduct *in-vitro* experiments, although the animals can be fixed *in situ* (Yan *et al.* 2022).

Promoter methylation is associated with decreased gene expression

In vertebrates, gene promoter methylation silences expression by impeding the binding of transcription factors (Suzuki and Bird 2008). A similar gene silencing mechanism was observed in the mud crab *Scylla paramamosain* (Jiang *et al.* 2020). In the deep-sea worms investigated in this study, a negative correlation between the promoter methylation and expression amongst highly methylated genes was observed. This association could reflect a gene silencing mechanism similar to that of vertebrates, although the relative weakness of the association in our samples suggests that promoter methylation may play a supporting role to other gene silencing mechanisms (Keller *et al.* 2016). Marsh and Pasqualone (2014) showed that DNA methylation in the Antartic polychaete species *Spiophanes tcherniai* was linked to thermal acclimation. Future study of the physiological role of DNA methylation in deep-sea worms exposed to strong spatial and temporal thermal gradients, such as those inhabiting hydrothermal vents, could bring new insights into the mechanisms of animal's adaptation to environmental disturbances in the deep-sea habitats.

Conserved patterns of DNA methylation in the deep-sea worms may reflect adaptations to hydrostatic pressure

Methylation profiles, which were conserved across species (for both gene-bodies and promoter regions), were matched by a strong genetic signal, indicating their persistence over evolutionary times (methylated cytosines have a higher C→A mutation rate [Coulondre *et al.* 1978; Ehrlich *et al.* 1986]) and pervasiveness (the mutations must affect the germ cells to be passed on to future generations). The epigenetic makeup of an individual is partly encoded in its genome and is therefore under the influence of natural selection (Angers *et al.* 2020). Thus, polychaetes species that are endemic to deep-sea chemosynthetic ecosystems are expected to have acquired adaptive epigenetic buffering strategies aside from behavioral coping methods and molecular alterations to extreme habitats.

In the three species studied, genes with a conserved pattern of promoter hypomethylation and gene-body hypermethylation (conducive to upregulation) displayed strong methylation-driven mutational biases and were disproportionately associated with the maintenance of protein homeostasis. Chaperones, which are important for protein folding stability in the deep-sea (Cario *et al.* 2016; Ritchie *et al.* 2018; Weber *et al.* 2020), were highly methylated in the worms. Likewise, the ubiquitin-proteasome pathway, which is responsible for degrading misfolded or damaged proteins and plays an important role in cold-water adapted species (Todgham *et al.* 2017), was overrepresented amongst genes with hypomethylated promoters in all three worms. Our samples were not fixed in situ, and thus, we cannot exclude that the methylation profiles of the worms partly reflect the decompression stress they suffered during their recovery to the ship (Yan *et al.* 2022). However, methylomes tend to be much more stable in time than transcriptomes (Strader *et al.* 2020) and thus, environmental stress during the short timeframe of samples recovery (a few hours at most) is unlikely to have had substantial consequences on worms' methylomes. Further comparative studies on deep-sea invertebrates and their shallow-water relatives could combine transcriptomic and epigenomic data to shed new light onto the mechanisms of adaptation to deep-sea environment.

Conclusion

Deep-sea polychaetes possess a fully functional DNA methylation metabolism, and their genomes are moderately methylated. In *P. echinospica*, *R. piscesae* and *P. palmiformis*, DNA methylation does not specifically target transposable elements but is highly concentrated in genes. Among the three worms, gene expression is positively correlated to gene body methylation and, to a lesser extent, negatively correlated to promoter methylation. These chemical modifications of the DNA, which appear to exert control on gene expression, could play roles in metabolic acclimation at the individual level. Finally, the persistent and conserved epigenetic profiles of genes responsible for maintaining homeostasis suggest DNA methylation plays an important adaptive role at the population and species level. The complete genomes and epigenomes that were assembled for three endemic polychaetes would serve as a valuable reference for future investigations of the ecological and evolutionary roles of DNA methylation in deep-sea animals.

Acknowledgments

We thank Verena Tunnicliffe, Catherine Stevens, Rachel Boschen-Rose and the crew members of the CCGS John P. Tully and ROV ROPOS for providing the *R. piscesae* and *P. palmiformis* specimen. We acknowledge Jack (Yick Hang) Kwan and Yi Yang for their help in nucleic acid extraction. We are also grateful to Alex de Mendoza for his comments on this study's preliminary results, Chema Martin Duran, Ekin Tillic and Josefin Stiller for providing additional polychaetes' genomic and transcriptomic assemblies, Sandra Ann Binning, Sophie Breton, Marie-Anne Cambon-Bonavita and three anonymous reviewers for their helpful comments on the manuscript. This study was supported by grants awarded to P-YQ from the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) [2021HJ01, SMSEGL20SC01], the Major Basic and Applied Research Projects of Guangdong Province [2019B030302004-04], and a GRF and CRF of HKSAR government [16101822, C2013-22G] and grants awarded to BA by the Natural Science and Engineering Research Council of Canada [N238600] and the Digital Research Alliance of Canada [RRG #4257]. MP received support from the Natural Science and Engineering Research Council's Alexander Graham Bell fellowship and the Fonds de recherche du Québec – Nature et Technologie and Quebec Center for Biodiversity Science's international internship award.

Chapter 5 – General conclusions

My general objective was to determine what factors influence the evolution and connectivity of symbiotic bacteria, and what the mechanisms of resilience in deep-sea chemosynthetic ecosystems were at the level of the populations and species.

By conducting comparative genomic, genetic and epigenomic analyses on CBE's bacterial symbionts and their hosts, I showed that (1) genetic drift (affected by inter-specific recombination rates) and selection were both equally important in shaping the genetic makeup of vertically transmitted symbionts, (2) symbiotic bacteria with a free-living stage exhibit a segregation pattern primarily influenced by geographical distance and/or bottom current patterns, rather than the physicochemical conditions of their respective chemosynthetic habitats, and (3) epigenetics, and in particular DNA methylation, is important to gene regulation in polychaetes inhabiting CBEs and could have adaptive properties at the level of species.

This chapter offers a concise summary of the principal findings derived from each study conducted within this thesis. This summary is followed by a perspective on how the research presented in this thesis and new genetic approaches can contribute to future conservation endeavors in the deep-sea.

Summary of the main findings

Study 1 (Chapter 2)

Table 5.1 Summary of objectives and hypotheses pertaining to the first study.

STUDY 1: COMPARATIVE GENOMICS OF THE SYMBIONTS OF VESICOMYID CLAMS		
Theme: *How have deep-sea taxa evolved?*	*Keywords:* *Molecular evolution, symbiosis*	*Model:* *Vertically-transmitted endosymbionts of vesicomyid clams*
Main Objective	**Determine the relative contribution of neutral and selective processes to the evolution of vesicomyid clams endosymbionts**	
Sub-Objective 1 : **Determine if the vesicomyid symbiosis is typical of that of vertically transmitted symbionts**		
Hypothesis	The vesicomyid symbionts represent an intermediate state of reductive genome evolution	
VALIDATED		
Sub-Objective 2 : **Determine whether the symbionts genome evolution is affected by neutral processes**		
Hypothesis	Genetic drift strongly affects the genome evolution of clam symbionts	
VALIDATED		
Sub-Objective 3 : **Determine if selective processes also play an important role in symbiont genome evolution**		
Hypothesis	Arms race evolution affects the clam symbionts	
NOT VALIDATED; Positive selection did not preferentially affect genes related to host-symbiont communications		

This study used the vesicomyid clams holobiont model to address a fundamental question in evolutionary biology: what is the interplay between neutral and selective evolutionary forces throughout the transition from free-living bacterial cells to fully integrated cellular organelles?

The main objective was to determine the relative contribution of neutral and selective processes to the evolution of the vesicomyid clams endosymbionts (Table 5.1). When this project started, two complete genomes of vesicomyid clam symbionts had been published (Kuwahara *et al.* 2007; Newton *et al.* 2007). Two more were released (Ip *et al.* 2020; Russell *et al.* 2020) by the time of the study's publication. Our first goal was thus to assess whether the genomes of the vesicomyid symbionts were typical of that of vertically transmitted bacteria and bared the signatures of RGE. By comparing nine new symbiont genomes to these already published, the hosts mitochondria, and the free-living bacteria outgroup, we confirmed that vesicomyid symbionts represent an intermediate state of RGE. Notably, they presented reduced size, CG %, and gene content. Contrasting differences amongst Clade I and Clade II symbionts also supported preliminary observations (Kuwahara *et al.* 2011) that the *Gigas* and *Ruthia* holobiont clades were on different evolutionary paths. Our second goal was to assess whether the symbionts' genome evolution was affected by neutral processes. The pervasive shift of symbiont dN/dS towards neutrality and reduction in codon usage bias indicated that genetic drift strongly affect the genome evolution of clam symbionts. Interestingly, we also found a correlation between inter-specific recombination and the intensity of RGE amongst the clam symbionts. This highlights the importance of genetic exchanges in reducing the effects of genetic drift. Our last goal was to determine if selective processes, including arms race dynamics, also played an role in the clam symbionts evolution. We expected to preferentially detect episodic positive selection in genes that affect host-symbiont communications but found instead that all functional categories were affected by diversifying selection. In light of RGE, these results may indicate the prevalence of compensatory mutations to face the drift-driven accumulation of deleterious mutations. Last but not least, we identified putative functional differences related to symbiont nitrogen and sulfur physiology, and dependency on environmental vitamin B12 appear to have maintained and possibly driven holobiont niche segregation.

In summary, we showed by applying comparative genomics that while nearly-neutral processes have been the motor of reductive genome evolution in vesicomyid symbionts, selective processes

have also played a crucial role in maintaining symbiont functional integrity throughout their evolutive history.

Study 2 (Chapter 3)

Table 5.2 Summary of objectives and hypotheses pertaining to the second study.

STUDY 2: DIVERSITY AND POPULATION STRUCTURE OF *R. PISCESAE* SYMBIONTS		
Theme: *How connected are hydrothermal vents?*	*Keywords:* *Bacterial population genetics, connectivity*	*Model:* *Populations of Ridgeia piscesae symbionts on the Juan de Fuca ridge*
Main Objective	**Characterize the structure of the populations of *R. piscesae* symbionts on the Juan de Fuca ridge**	
Sub-Objective 1 :	**Determine the extent of genetic diversity in the symbionts**	
Hypothesis	There are multiple strains of symbionts in the environment	
VALIDATED		
Sub-Objective 2 :	**Determine how the symbiont populations are partitioned**	
Hypothesis	Symbionts cluster into multiple ecotypes according to environment	
NOT VALIDATED; Symbionts appeared to cluster according to patterns of deep-sea circulation instead		

Bacterial symbionts are the main (and sometimes sole) food providers for the habitat-forming invertebrate species at CBEs. Hence, this study of bacterial population genetics asked a key question for CBEs conservation: how are communities connected at the local scale?

The study's aim was to characterize the structure of the populations of *R. piscesae* symbionts (*Ca.* Endoriftia persephone) on the Juan de Fuca ridge (Table 5.2). To do so, we first had to determine the extent of the symbiont genetic diversity. Along with the universal 16S gene marker, we used CRISPR and three other gene markers to discriminate between strains. CRISPR had previously been used to follow the proliferation of known and culturable bacterial pathogens but this study is the first to use it on an unculturable bacterial species. We found that unlike the 16S gene which showed almost no polymorphism, CRISPR revealed more than 150 different strains including 10 dominant ones. These strains were found across individual hosts indicating that they were not the result of host-specific 'somatic' mutations but truly represented the environmental diversity of *Ca.* Endorifita persephone. Furthermore, the patterns of strain diversity uncovered with CRISPR matched these of the other polymorphic gene markers. The presence of older but not newer CRISPR spacers in vicariant populations of *Ca*. Endorifita persephone confirmed this marker accurately retraced the evolutive history of the bacteria and thus was suitable for population genetics. Hence, the CRISPR-derived patterns of diversity were then further analysed to determine how the symbiont populations were partitioned, whether it be by habitat type or geographic distance. We expected the symbiont communities to cluster into distinct ecotypes but observed that

the broad environmental conditions of the milieu (characterized by distinct temperatures, oxygen and hydrogen sulfide levels) did not explain the symbiont genetic structure. Instead, this structure seemed to match known patterns of regional deep-sea circulation.

Taken together, the results of this study demonstrate that CRISPR is a good marker to reveal the strain-level diversity of natural bacterial populations and that physical barriers affect the partitioning of CBEs symbionts.

Study 3 (Chapter 4)

Table 5.3 Summary of objectives and hypotheses pertaining to the third study.

STUDY 3: FUNCTIONAL ROLES OF DNA METHYLATION IN THREE POLYCHAETE SPECIES		
Theme: *How resilient are CBE species to environmental changes?*	*Keywords:* *Epigenomes, DNA methylation, Evolution,*	*Model:* *Paraescarpia echinospica, Ridgeia piscesae and Paralvinella palmiformis*
Main Objective	**Assess if DNA methylation is an important epigenetic mechanism in deep-sea worms**	
Sub-Objective 1 : **Assess whether the worms possess a functional metabolism for DNA methylation**		
Hypothesis	*The worms possess a functional DNA methylation metabolism*	
VALIDATED		
Sub-Objective 2 : **Evaluate if the methylome can be obtained from third generation sequencing technology**		
Hypothesis	*DNA methylation is accurately detected by third generation sequencing*	
VALIDATED; Provided that coverage and window size thresholds are set		
Sub-Objective 3 : **Test established hypotheses about the roles of DNA methylation**		
Hypothesis 1	*Gene-body methylation upregulates gene expression*	
VALIDATED		
Hypothesis 2	*Promoter methylation down regulates gene expression*	
VALIDATED		
Hypothesis 3	*Transposable elements are silenced by DNA methylation*	
NOT VALIDATED; Transposable elements did not appear to be specifically targeted by DNA methylation		

How resilient are CBEs species to unpredictable environmental changes? In many species, resilience is enabled by physiological and phenotypic plasticity. This plasticity is itself enabled by epigenetic mechanisms. Hence, epigenetic mechanisms are likely to play an important role in CBE invertebrates.

This study constitutes the first genome-wide characterization of DNA methylation in the phylum Annelida and the first methylomics survey of a deep-sea species. Its goal was to assess if DNA methylation was an important epigenetic mechanism in deep-sea polychaete worms (Table 5.3). We first had to determine whether the worms possessed a functional metabolism for DNA methylation. Investigating the genomes and transcriptomes of three polychaete species, we found that all key genes of the DNA methylation metabolism were present and expressed in these species.

Given that Nanopore is increasingly being used for genome sequencing projects, our second objective was to benchmark this technology for recovering methylomics data. To do so, we compared the DNA methylation profiles obtained through Nanopore sequencing and the gold-standard (but more expensive) WGBS approach. We found that the per CpG methylation levels derived from both methods were highly correlated but established two important thresholds for insuring optimal estimation of methylation levels; the per-base methylation call coverage should be above 10, otherwise, methylation should be averaged over 1000 Kbp genomic windows to increase accuracy without significantly impacting sensitivity. The last objective of this study was to test established hypotheses about the putative roles of DNA methylation in invertebrates. By correlating the genome-wide patterns of DNA methylation in the three worm species investigated to the levels of transcriptomic expression, and gene functions, we were able to confirmed that gene-body methylation was correlated to upregulation whereas promoter was, to a lesser extent correlated to gene down-regulation. However, comparing the methylation on transposable elements to their genetic context and their estimated insertion times, we found no evidence that they were specifically targeted by the DNA methylation machinery. Lastly, a comparative analysis of orthologous genes across the three worm species revealed that genes with a methylation pattern conducive to upregulation were preferentially associated to protein homeostasis functions. This observation could reflect a mechanism of epigenetic adaptation to increased hydrostatic pressure in the deep-sea but further studies would need to be conducted to test this new hypothesis.

Overall, we showed in this study that DNA methylation is indeed present in deep-sea polychaetes (at least in the species investigated) and that it plays an important functional role.

Perspective: towards effective conservation guidelines

The benthic area beyond national jurisdiction, simply called “The Area”, represents more than 50% of the area of the world’s oceans (Figure 5.1) and includes most of the known CBEs. Mining regulations are currently being drafted by the ISA (International Seabed Authority 2019). and will require that mining permit applicants provide 1) a detailed environmental impact assessment, and 2) monitoring plan. The methodologies and thresholds informing environmental assessments and monitoring plans will be resource- and site-specific, and it is up to the scientific community to develop the appropriate standards and guidelines.

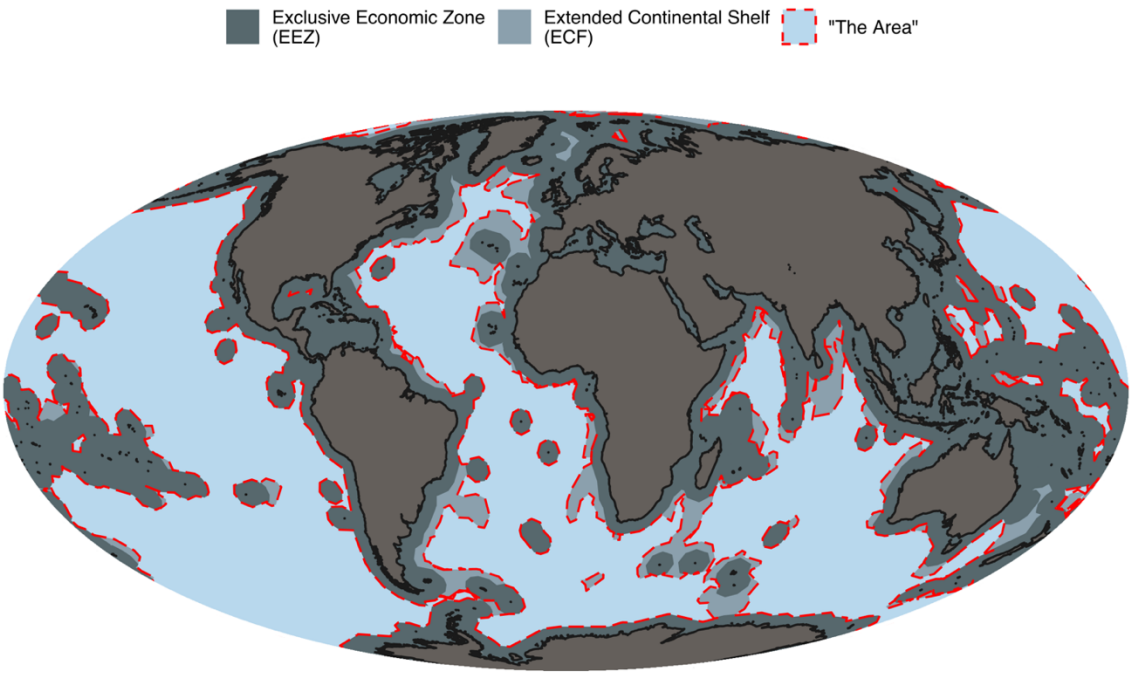

Figure 5.1 Global map of "The Area". Data from marineregions.org including most recent list of ECF (Flanders Marine Institute 2022).

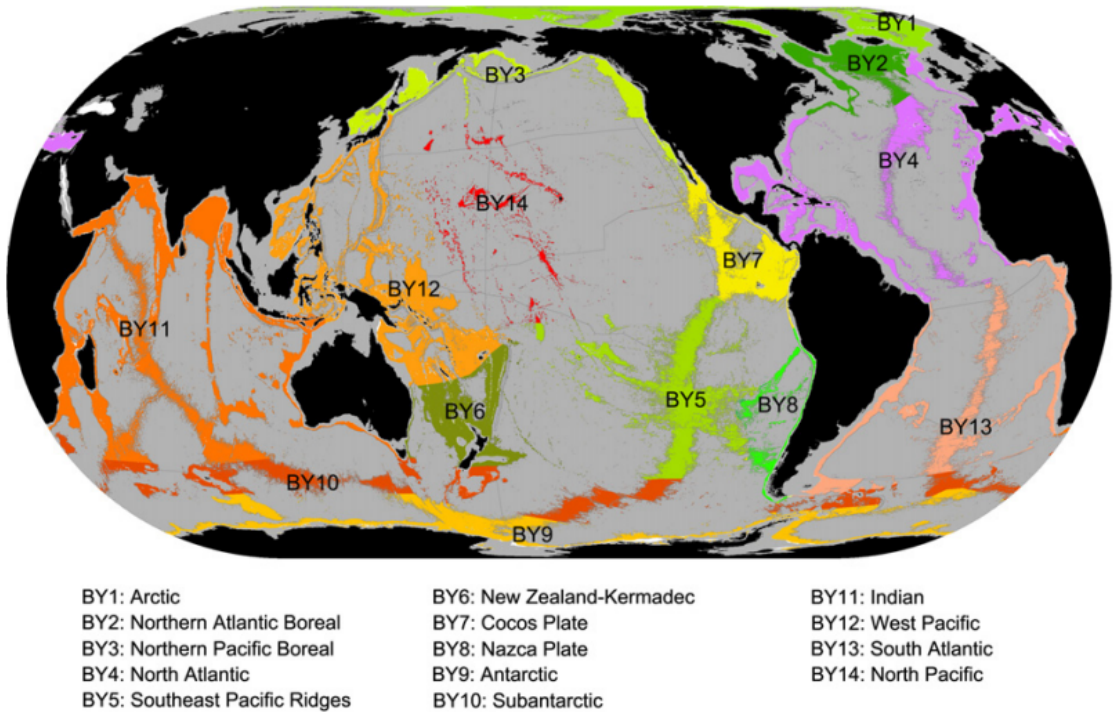

Figure 5.2 Deep-sea benthic biogeographic regions. From Watling *et al.* (2013)

Globally, CBE communities are grouped in several biogeographic zones (Bachraty *et al.* 2009; Levin *et al.* 2016; Watling *et al.* 2013) which roughly map to the different oceanographic basins and likely reflect broad patterns of connectivity and environmental conditions (Figure 5.2). Deep-sea CBEs are therefore highly diverse, and it is critical to keep in mind that ecological inferences are not necessarily transposable across basins. For instance, an important factor shaping the ecology of animal communities in mid-ocean ridge hydrothermal vents is the rate at which ridges spread apart. High volcanic activity at fast-spreading ridges can wipe out pre-existing chimney structures after just a few decades (Nees *et al.* 2009) and is therefore selecting for short-lived but overall highly resilient animal communities (Hessler *et al.* 1988; Juniper and Tunnicliffe 1997; Sarrazin *et al.* 1997). On the other hand, venting sites may be active for several hundreds of thousand years in slow spreading ridges (Jamieson *et al.* 2013) and will host more stable, and likely less resilient, communities of slow-growing fauna (Zhou *et al.* 2018).

Nevertheless, the increased pace of exploration of CBEs worldwide and the resulting growing body of literature about these ecosystems offers valuable insights for conservation purposes. Even though their individual scopes were limited, the studies that I have carried and contributed to over the course of my doctoral studies highlight a number of important general rules that will help formulate specific recommendations. In this the next section, I will showcase how this research informs the development of upcoming mining standards and guidelines and discuss how emerging environmental genetics approaches can help us better assess the resilience of putative mining sites and monitor the impacts of human activities on the seafloor.

Lessons from the research presented in this thesis

The symbiont transmission mode of foundation species has consequence on the resilience of their host and the way different CBEs should be managed

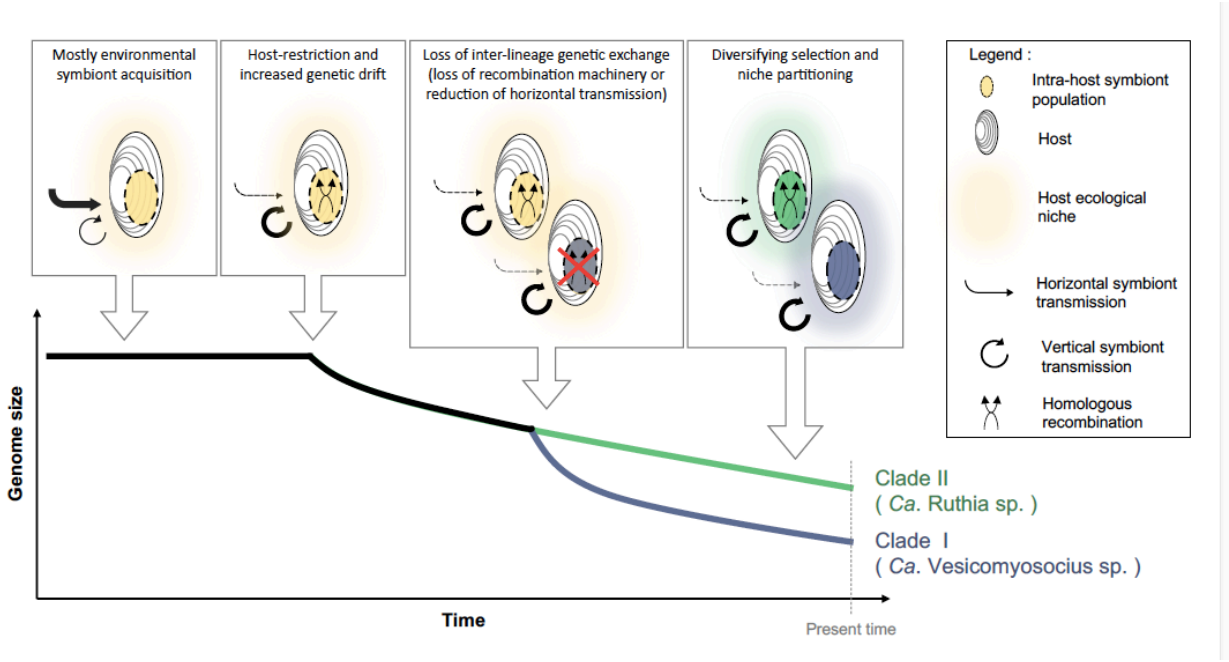

Figure 5.3 Summary of the evolutive history of vesicomyid clams symbionts evolution.

Chapter 2's study demonstrates how the loss of recombination in vertically transmitted symbionts can rapidly and drastically alter the genomic makeup of mutualistic bacterial symbionts. Through a combination of genetic drift and selective processes, the recently acquired symbionts of deep-sea vesicomyid clams have in the relatively short time span of 45 million years, become highly specialized plastid-like organelles (Figure 5.3). Thanks to the maintained specificity of their symbiotic association, clam holobiont populations have become highly adapted to the specific ecological niche they occupy but in turn, may be less resilient in the face of changing environmental conditions. Clade I holobionts for instance, appear to be constrained to higher quality habitats (higher levels of hydrogen sulfide) possibly because of the drift-driven degeneration or selective differential erosion of the symbionts genome. Hence, small shifts in the physico-chemical conditions of the milieu could wipe out entire populations and recolonization by new recruits would not be possible even under high larval connectivity unless the initial environmental conditions were restored. Large fields of dead clams are often observed in area of low hydrocarbon seepage

suggesting mass vesicomyid death and quick temporal successions follow changes in environmental conditions (Guillon *et al.* 2017; Hessler *et al.* 1988; Heyl *et al.* 2007).

As a general recommendation, it follows that at sites dominated by foundation species with a dominant vertical symbiont transmission mode, particular efforts should be made to minimize environmental disturbances and to ensure environmental conditions are not permanently altered by marine operations. The design of protected areas should in these cases be informed by the host species' patterns of connectivity.

On the other hand of the symbiont transmission spectrum, CBE animals which acquire their symbiotic partners from the environment at each generation, such as tubeworms of the species *R. piscesae* (Chapter 3) benefit from a degree of partner choice. In this case, hosts can associate with multiple bacterial strains (or even species) issued from an environmental pool of symbionts that are 1) genetically diverse and 2) locally adapted. The high intra-host and inter-host strain diversity of *Ridgeia*-associated *Ca.* Endoriftia persephone uncovered in the study of Chapter 3 are indirect evidence of this phenomenon. Because they were based on a few genetic markers, the genetic data presented in Chapter 3 cannot inform us about putative metabolic and functional differences that exist between *Ca.* Endoriftia persephone strains. However, in Patra *et al.* (2022), we found that the genome of *Bathymodiolus thermophilus*'s symbionts issued from EPR 9N but no other hydrothermal vent sites, possessed a hydrogenase operon. We emitted the hypothesis that the EPR 9N *B. thermophilus* symbionts are uniquely able to derive energy from the oxidation of di-hydrogen and that metabolic capability would provide the holobiont with a unique phenotype and possibly broaden or shift its ecological niche. Similarly, in Lan *et al.* (2022), we found that the symbionts of the scaly-foot snails which follow a mixed mode of transmission, were structured locally and possessed different metabolic repertoires possibly linked to hydrothermal fluid chemistry. Many other studies of CBE holobionts have linked symbiont diversity to functional breadth (Aubé *et al.* 2022; Breusing *et al.* 2022; Cambon-Bonavita *et al.* 2021; Dick 2019; Duperron *et al.* 2006; Zimmermann *et al.* 2014).

Taken together, these studies suggest holobionts that are dependent on horizontally transferred symbionts may be able to thrive in a broader range of habitats and thus may be more resilient to environmental shifts. That being said, the decoupling of the host and symbiont genetic diversity

and dispersal necessitates that not only the host but also the symbiont demographics be taken into account when planning and monitoring mining operations (see next section).

The connectivity patterns of keystone bacteria should be taken into account in mitigations plans and protected areas network designs

High throughput sequencing technologies gives us unprecedented access to the genetic diversity of uncultured bacterial populations and interrogate multiple loci and even whole genomes. In hydrothermal vents, recent population-level whole-genome shotgun (WGS) metagenomic studies have uncovered striking variation in the genetic makeup of host-associated bacterial species coming from different sites of the same spreading ridge (Breusing *et al.* 2022; Jang *et al.* 2022; Lan *et al.* 2022). These studies offer valuable insights into the metabolic breadth of various symbiont species and their large-scale biogeography (hundreds-thousands of kilometers) but not their connectivity at the local or regional scale (tens-hundreds of kilometers). Indeed, such inferences of dispersal necessitate to resolve individual haplotypes which in haploid bacteria correspond to individual cells. These bacterial haplotypes or strains cannot be singled out from WGS metagenomic data because of the short length of their sequenced reads. In Chapter 3's study, we overcome this limitation by using CRISPR as a single hypervariable genetic marker (one CRISPR haplotype = one bacterial strain). The CRISPR-based regional genetic structure of *Ca.* Endoriftia persephone associated with *R. piscesae* indicated that distance and/or physical barriers (*e.g.* topography, deep-ocean current patterns) best explained the distribution of strains along the Endeavour segment of the Juan de Fuca ridge (which is less than 200 km in length). This study therefore underlines the importance of characterising the connectivity patterns of bacterial populations even across CBE sites that are geographically close to each other.

A recent study, Jang *et al.* (2022) used high throughput sequencing of six functional genes to characterize the population structure of the horizontally transmitted *Bathymodiolus* mussels symbionts in two ocean basins, the Central Indian Ocean and the eastern Pacific Ocean. The authors found the structure of the symbiont populations followed a stepping stone model within the 1600 km of the slow-spreading Central Indian Ridge (CIR) but not along the >4000 km of the fast-spreading eastern Pacific ridges. Furthermore, applying isolation with migration models to the CIR host and symbiont populations, they found contrasting gene flows; southward for the hosts and northward for the symbionts. The results of this study carry two important implications; 1) that

inferences of symbionts populations structure are not transferable across ocean basins and have to be evaluated on a per-case basis, and 2) that the host and symbiont dispersal potential and connectivity patterns are independent. The study of Jang *et al.* (2022) and that of Chapter 3 together make a strong plea for the systematic inclusion of key-stone symbionts in future environmental impact assessments and design of marine protected areas.

The study of deep-sea epigenetics can shed novel light onto the mechanisms of adaptation and resilience in CBEs.

Elucidating the short- and long-term mechanisms of physiological resilience of CBEs populations is important to predict the effects of human impacts in these environments. The article presented in Chapter 4 proposes that new mechanisms of resilience can be uncovered by the study of an epigenetic mechanism: DNA methylation.

In Chapter 4's study, we presented the genome-wide methylomes of three deep-sea polychaetes: *Paraescarpia echinospica*, *Ridgeia piscesae*, and *Paralvinella palmiformis*. These are not only the first complete epigenomes for deep-sea species but also the first for the phylum Annelida, thus filling an important knowledge gap of the invertebrate epigenomic landscape. The genomes of invertebrate species range from broadly hypermethylated (in sponges) to completely devoid of methylation (in some nematodes) (de Mendoza *et al.* 2020). Our DNA methylation surveys showed deep-sea polychaete species possess intermediate levels of DNA methylation among protostomes. By combining epigenomic, genomic and transcriptomic data, we confirmed that DNA methylation played an important role in modulating gene expression in these worms. Interestingly, we found that functions pertaining to protein homeostasis were epigenetically targeted for upregulation in all three species, possibly unmasking an adaptive mechanism of epigenetic buffering to increased hydrostatic pressure. DNA methylation is also present in bacteria. In Patra *et al.* (2022), we detected DNA methylation in the genome of the 9N symbionts of *Bathymodiolus thermophilus* but its role remains unknown.

Other epigenetic mechanisms such as histone modifications and microRNAs could also help us understand resilience in CBEs and in Angers *et al.* (2020), I further argue that conceptualizing the microbiome as an epigenetic trait of the holobiont would offer a valuable conceptual framework to assess how species cope with unpredictable environmental variations. The associations between

the hosts and their environmental symbionts may be more or less stringent but even rich microbiomes composed of facultative symbionts show a degree of host specificity. In Lee *et al.* (2021) for instance, we showed evidence of co-cladistic associations between the dominant members of the gill microbiomes of three mollusc and two crustacean species issued from the same vent site in the Tonga arc. These results support a large body of literature showing that the composition of the environmentally acquired microbiome of animals is issued from the integration of both host and environmental factors (Bordenstein and Theis 2015; Gilbert 2014) and provides a physiological plasticity to the holobiont.

I posit that additional studies of the DNA methylation and microbiomes of CBE benthic megafauna at the population level will uncover valuable information on species-specific epigenetic buffering strategies. Two of such strategies which are particularly relevant to the heterogeneous and unpredictable nature of the benthic megafauna habitats in CBEs are phenotypic plasticity (favoring specific environmentally-driven variation in conditions that are variable but predictable) and bet-hedging (favoring stochastic variation in unpredictable conditions) (Leung *et al.* 2016). Indeed, most benthic invertebrates will disperse during their larval stage and settle (more or less selectively) in a new environment away from that of their parents (Mullineaux *et al.* 2010; Yahagi *et al.* 2017). Furthermore, these environments are composed of a multitude of micro habitats that may have vastly different biotic and abiotic characteristics and can themselves be highly erratic (*e.g.* hydrothermal vent sulfide chimneys [Sarrazin *et al.* 1997]). Thus, it can be hypothesized that the degree to which animals are able to predict the environmental conditions they settle in will favor one strategy over the other (Chevin and Lande 2011; Leung *et al.* 2016).

Filling knowledge gaps through environmental genetic surveys

The genetic material shed by species into the environment is degraded in the span of hours or days (Barnes *et al.* 2014). Hence, the genotyping of environmental samples, such as water, soil, or air, referred to as environmental DNA (eDNA), provides a snapshot of the species present in an environment without actually seeing or capturing them (Barnes *et al.* 2014). eDNA approaches predate the advent of sequencing technologies (Pace *et al.* 1986). Their use, was initially contained to the field of microbiology but the emergence of high throughput sequencing technologies, the increased automatization of their workflow, and the miniaturisation of their hardware have led to

their widespread adoption in a variety of research fields including marine sciences. Today, eDNA methods are used in a broad range of applications including endangered species detection, invasive species management, and biodiversity monitoring (Senapati *et al.* 2019). Applied to the deep-sea, these methods could make a revolution because they considerably reduce the cost of sampling which is the main factor limiting research (Amon *et al.* 2022). In the following paragraphs, I will discuss how eDNA barcoding could be used to 1) better estimate larval dispersal (Figure 5.4A), and 2) characterize taxonomic diversity (Figure 5.4B); two previously identified key knowledge gaps (Miller *et al.* 2018; Perez *et al.* 2021; Sarradin *et al.* 2017).

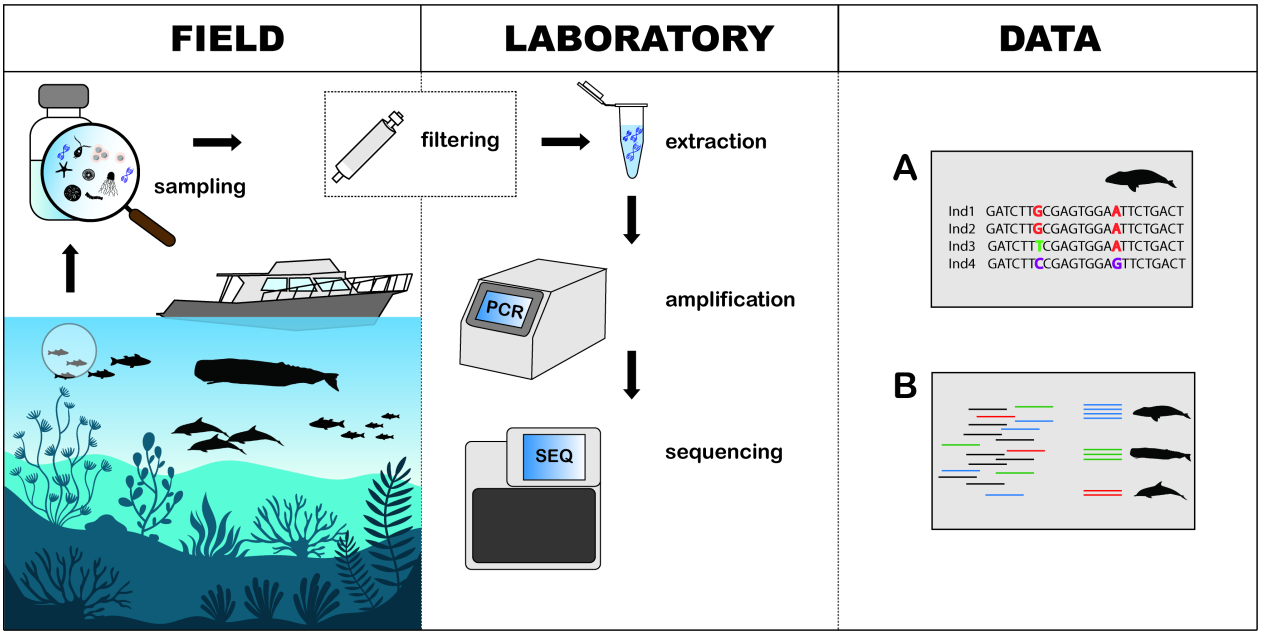

Figure 5.4 Schematic representation of eDNA workflow and its applications. A) genotyping for population genetics studies, B) metabarcoding for biodiversity characterisation. Adapted from Székely *et al.* (2022).

eDNA for estimating larval dispersal

The genetic fluxes between the populations of CBE animals provide information on their connectivity on the time scale of thousands of years and may not be relevant for conservation which necessitate to estimate population connectivity on the time scale of decades (Lowe and Allendorf 2010). Most macrofaunal species in CBEs possess pelagic larva but their dispersal is particularly difficult to estimate. Indirect assessments of larval dispersal combine information about larval physiology (e.g. its swimming ability, whether it can feed in the water column) to oceanic current modelling but often lack validation (Arellano and Young 2011; Brooke and Young 2009). Meanwhile, obtaining direct evidence of larval distribution is very challenging because larvae are

extremely diluted in the water column, easily damaged during sampling, and hard to identify based on morphotypic characteristics (Kim *et al.* 2022; Mullineaux *et al.* 2005). eDNA methodology could overcome these limitations because water samples are much easier to obtain than zooplankton samples, genetic markers do not require the recovery of whole or intact individuals, and the primers designed to amplify the key genetic markers of most macrofaunal species of interest already exist. For instance, such primers exist for 27/35 of the described species found in the Indian Ocean's hydrothermal vents (NCBI's inquiries against the species list from Perez *et al.* [2021]). However, eDNA approaches to larval detection have important limitations. Firstly, DNA lasts much longer in the deep-sea than in shallow waters because of low temperatures and pH (Barnes *et al.* 2014; McCartin *et al.* 2022). In McCartin *et al.* (2022), quantifiable levels of eDNA from *Lophelia* corals could persist for over two weeks in the deep-sea but for less than 7 days in the warm surface waters. Thus, samples issued from the deep-sea have a higher risk of false positive detection of larvae since DNA from dead individuals or their wastes can be transported by oceanic currents over very long distances in that time frame; in the range of tens to hundreds of km according to Mitarai's (2016) estimates of passive dispersal at 1000m in the western Pacific Ocean. Environmental RNA which degrades much faster than eDNA should therefore be used when dealing with deep-sea samples (Giroux *et al.* 2022). Secondly, methodologies for quantitative measurements of species abundance using qPCR or ddPCR exist for some types of environmental data (Doyle *et al.* 2017; Pinheiro *et al.* 2012; Senapati *et al.* 2019) including bulk zooplankton samples (Breton *et al.* 2022; Garcia-Vazquez *et al.* 2021; Uthicke *et al.* 2018) but they require strong baseline knowledge (e.g. the minimal DNA concentration needed to detect one individual [Klymus *et al.* 2020]), are tedious to develop and not always accurate when tested in the field (Fonseca 2018; Yates *et al.* 2019). Lasty, both quantitative and qualitative applications of eDNA/eRNA methods have high rate of false negative detection. Nucleic acids from the species of interest may fail to be captured in the sampling even when present in the environment or may not sufficiently amplify against background nucleic acids even when present in the sample (Ruppert *et al.* 2019). This constraint can however partly be overcome by scaling up environmental sampling (the volume and number of samples taken from the environment) and genetic sampling (the sequencing depth) which is easy and inexpensive to do. International efforts by scientists and contractors in 'The Area" to sample eDNA across depths and space (e.g. on- and off-axis in spreading ridges) using a standardized methodology would strengthen inferences of larval

distributions and advance our understanding of their ecology. Such endeavour could be coordinated by the ISA as it already oversees the implementation of the standardized acquisition of environmental baseline data by contractors as part of their exploration contracts (International Seabed Authority 2010).

eDNA for characterising biological diversity

The characterisation of CBE taxonomic diversity, and the biotic and abiotic factors that control it is essential to predict the effects of seabed mining. Unfortunately, the benthic fauna in CBEs remains largely undescribed, especially at the micro and meio scales. By using universal primers to amplify highly conserved marker genes, it is possible to detect multiple species of various sizes from environmental DNA and thus shed light on this dark diversity. In one of the first application of eDNA to CBEs, Cowart *et al.* (2020) analyzed the community composition at the Monségur hydrothermal vent (Mid-Atlantic Ridge) using traditional (recovering, identifying and counting specimens) and eDNA methods. All of the 22 metazoan taxa that were identified to the genus or species level from morphology and DNA barcoding were recovered in the eDNA samples of sulfide and rocky substrates. Furthermore, the eDNA datasets included additional species of mobile macrofauna (a snail and a polychaete) and meiofauna (a nematode) which were present in high abundance but were possibly gone at the time of sampling or too small to have been identified by traditional methods. DNA from non-vent animal species represented less than 10% of the eDNA datasets but because of high novelty at vents, these sequences may belong to yet undescribed vent species that had been assigned to the closest (non-vent) taxon within the NCBI GenBank database (Cowart *et al.* 2020). Other than metazoans, bacteria and protists were also abundant in the eDNA datasets representing between 15 and 70% of Operational Taxonomic Units (OTUs), but the authors did not analyzed these data further. Taken together, the results of this study highlight the power of eDNA to recover the taxonomic composition of benthic communities including its small, cryptic, and transient or rare species. Nevertheless, there are important limitations to eDNA application for community ecology. The first is that eDNA metabarcoding results are most accurate when based on incidence (presence/absence) because many factors can ultimately affect the number of reads recovered for a specific OTU or species; its size, the rate of decay of its DNA (itself dependent on environmental conditions), primer relative affinity, etc. (Barnes *et al.* 2014). Hence, species abundances cannot be estimated from eDNA datasets. Secondly, a good reference

database is needed to assign accurate taxonomic classification to the OTUs and therefore eDNA methods still rely on traditional deep-sea fauna inventories. There inventories are however growing rapidly and OTUs can still be informative even when not assigned to a specific taxon. In Cowart *et al.* (2020) for instance, the authors used the full OTU repertoire to analyze beta-diversity and compare communities across distinct vent habitats (habitats near active venting: sulfide chimneys and base of sulfide edifice, periphery of the vent, and inactive chimneys) and found contrasting community compositions across them but not across the different years of sampling. Moving forward, the development of a comprehensive CBEs metabarcoding database and a standardized methodology informed by strong baseline studies will be key to make eDNA approaches globally adopted and part of the diagnostic toolkit for assessing mining-related environmental impacts.

Conclusion

The three studies that I conducted during my doctoral studies contributed to filling knowledge gaps about the evolution and ecology of the unique benthic fauna inhabiting deep-sea CBEs. By comparing the genomes of 11 species of vesicomyid clam symbionts, I demonstrated that their reduced genome evolution was driven by a demographic bottleneck caused by host restriction and vertical transmission. Despite the strong relative effect of genetic drift on the symbiont evolution, the genetic signatures of selective processes were also evident but much remains to be uncovered about how exactly these processes contribute to the holobiont adaptation and affect its fitness and resilience to environmental changes. In another symbiotic model, I shed light onto the regional genetic diversity of a keystone bacterial species and the factors contributing to its spatial distribution using a novel hypervariable marker: CRISPR. Because this gene is widely spread amongst bacteria, it has the potential to reveal many new species-specific patterns of connectivity that will be particularly informative for designing marine protected networks. Lastly, I used species issued from CBEs to conduct the first genome-wide survey of DNA methylation in polychaetes (phylum: Annelida) and uncovered two putative functional roles for this epigenetic mechanism in the worms. Epigenomic comparative analyses also presented the epigenome as a novel target for deep-sea adaptation which should be investigated further in the future. Ultimately, I hope that the research unraveling from the threads of this doctoral research and the adoption of new genetic surveying approaches such as eDNA will advance efforts for sustainable development and help protect these wonderful ecosystems.

References

Ahmed, I., Sarazin, A., Bowler, C., Colot, V., and Quesneville, H. 2011. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in *Arabidopsis*. Nucleic Acids Research **39**(16): 6919–6931. doi:10.1093/nar/gkr324.

Amon, D.J., Rotjan, R.D., Kennedy, B.R.C., Alleng, G., Anta, R., Aram, E., Edwards, T., Creary-Ford, M., Gjerde, K.M., Gobin, J., Henderson, L.-A., Hope, A., Ali, R.K., Lanser, S., Lewis, K., Lochan, H., MacLean, S., Mwemwenikarawa, N., Phillips, B., Rimon, B., Sarjursingh, S.-A., Teemari, T., Tekiau, A., Turchik, A., Vallès, H., Waysang, K., and Bell, K.L.C. 2022. My Deep Sea, My Backyard: a pilot study to build capacity for global deep-ocean exploration and research. Philosophical Transactions of the Royal Society B: Biological Sciences **377**(1854): 20210121. doi:10.1098/rstb.2021.0121.

Anantharaman, K., Breier, J.A., Sheik, C.S., and Dick, G.J. 2013. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. Proceedings of the National Academy of Sciences **110**(1): 330–335. doi:10.1073/pnas.1215340110.

Andersson, J.O., and Andersson, S.G. 1999. Genome degradation is an ongoing process in Rickettsia. Molecular Biology and Evolution **16**(9): 1178–1191. doi:10.1093/oxfordjournals.molbev.a026208.

Angers, B., Perez, M., Menicucci, T., and Leung, C. 2020. Sources of epigenetic variation and their applications in natural populations. Evolutionary Applications **13**(6): 1262–1278. doi:10.1111/eva.12946.

Arellano, S.M., and Young, C.M. 2011. Temperature and salinity tolerances of embryos and larvae of the deep-sea mytilid mussel "*Bathymodiolus*" *childressi*. Marine Biology **158**(11): 2481–2493. doi:10.1007/s00227-011-1749-9.

Aroh, O., and Halanych, K.M. 2021. Genome-wide characterization of LTR retrotransposons in the non-model deep-sea annelid *Lamellibrachia luymesi*. BMC Genomics **22**(1): 466. doi:10.1186/s12864-021-07749-1.

Atwater, T., and Stock, J. 1998. Pacific-North America plate tectonics of the neogene southwestern United States: an update. International Geology Review **40**(5): 375–402. doi:10.1080/00206819809465216.

Aubé, J., Cambon-Bonavita, M.-A., Velo-Suárez, L., Cueff-Gauchard, V., Lesongeur, F., Guéganton, M., Durand, L., and Reveillaud, J. 2022. A novel and dual digestive symbiosis scales up the nutrition and immune system of the holobiont *Rimicaris exoculata*. Microbiome **10**(1): 1–17. BioMed Central.

Audzijonyte, A., Krylova, E.M., Sahling, H., and Vrijenhoek, R.C. 2012. Molecular taxonomy reveals broad trans-oceanic distributions and high species diversity of deep-sea clams

(Bivalvia: Vesicomyidae: Pliocardiinae) in chemosynthetic environments. Systematics and Biodiversity **10**(4): 403–415. doi:10.1080/14772000.2012.744112.

Baas-Becking, L.G.M. 1934. Geobiologie; of inleiding tot de milieukunde. WP Van Stockum and Zoon NV.

Bachmann, N.L., Petty, N.K., Ben Zakour, N.L., Szubert, J.M., Savill, J., and Beatson, S.A. 2014. Genome analysis and CRISPR typing of *Salmonella enterica* serovar Virchow. BMC Genomics **15**: 389. doi:10.1186/1471-2164-15-389.

Bachraty, C., Legendre, P., and Desbruyères, D. 2009. Biogeographic relationships among deep-sea hydrothermal vent faunas at global scale. Deep-Sea Research Part I: Oceanographic Research Papers **56**(8): 1371–1378. doi:10.1016/j.dsr.2009.01.009.

Bao, W., Kojima, K.K., and Kohany, O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA **6**(1): 11. doi:10.1186/s13100-015-0041-9.

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Hérault, Y., Guillou, F., and Bourc'his, D. 2016. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science **354**(6314): 909–912. AAAS. doi:10.1126/science.aah5143.

Barco, R.A., Garrity, G.M., Scott, J.J., Amend, J.P., Nealson, K.H., and Emerson, D. 2020. A genus definition for bacteria and archaea based on a standard genome relatedness index. mBio **11**(1): e02475-19. American Society for Microbiology. doi:10.1128/mBio.02475-19.

Barnes, M.A., Turner, C.R., Jerde, C.L., Renshaw, M.A., Chadderton, W.L., and Lodge, D.M. 2014. Environmental Conditions Influence eDNA Persistence in Aquatic Systems. Environmental Science and Technology **48**(3): 1819–1827. American Chemical Society. doi:10.1021/es404734p.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315**(5819): 1709–1712. doi:10.1126/science.1138140.

Barry, J.P., and Kochevar, R.E. 1998. A tale of two clams: differing chemosynthetic life styles among vesicomyids in Monterey Bay cold seeps. Cahiers de Biologie Marine **39**: 329–331.

Bártová, E., Krejčí, J., Harničarová, A., Galiová, G., and Kozubek, S. 2008. Histone modifications and nuclear architecture: a review. Journal of Histochemistry and Cytochemistry **56**(8): 711–721. doi:10.1369/jhc.2008.951251.

Baumann, P. 2005. Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annual Review of Microbiology **59**(1): 155–189. doi:10.1146/annurev.micro.59.030804.121041.

Baumgartner, M., Roffler, S., Wicker, T., and Pernthaler, J. 2017. Letting go: bacterial genome reduction solves the dilemma of adapting to predation mortality in a substrate-restricted environment. The ISME Journal **11**(10): 2258–2266. doi:10.1038/ismej.2017.87.

Beaulieu, S.E. 2015. InterRidge global database of active submarine hydrothermal vent fields: prepared for InterRidge, Version 3.4. Available from http://vents-data.interridge.org [accessed 6 March 2018].

Beauruelle, C., Pastuszka, A., Horvath, P., Perrotin, F., Mereghetti, L., and Lanotte, P. 2017. CRISPR: a useful genetic feature to follow vaginal carriage of group B *Streptococcus*. Frontiers in Microbiology **8**. doi:10.3389/fmicb.2017.01981.

Bennett, G.M., and Moran, N.A. 2015. Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. Proceedings of the National Academy of Sciences **112**(33): 10169–10176. doi:10.1073/pnas.1421388112.

Berger, S.L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. 2009. An operational definition of epigenetics. Genes and Development **23**(7): 781–783. doi:10.1101/gad.1787609.

Bewick, A.J., Sanchez, Z., Mckinney, E.C., Moore, A.J., Moore, P.J., and Schmitz, R.J. 2019. Dnmt1 is essential for egg production and embryo viability in the large milkweed bug, *Oncopeltus fasciatus*. Epigenetics and Chromatin **12**(1): 6. doi:10.1186/s13072-018-0246-5.

Bicho, R.C., Scott-Fordsmand, J.J., and Amorim, M.J.B. 2020. Developing an epigenetics model species - From blastula to mature adult, life cycle methylation profile of *Enchytraeus crypticus* (Oligochaete). Science of the Total Environment **732**: 139079. doi:10.1016/j.scitotenv.2020.139079.

Boden, R., Scott, K.M., Williams, J., Russel, S., Antonen, K., Rae, A.W., and Hutt, L.P. 2017. An evaluation of *Thiomicrospira*, *Hydrogenovibrio* and *Thioalkalimicrobium*: reclassification of four species of *Thiomicrospira* to each *Thiomicrorhabdus* gen. nov. and *Hydrogenovibrio*, and reclassification of all four species of *Thioalkalimicrobium* to *Thiomicrospira*. International Journal of Systematic and Evolutionary Microbiology **67**(5): 1140–1151. doi:10.1099/ijsem.0.001855.

Boer, J.L., Mulrooney, S.B., and Hausinger, R.P. 2014. Nickel-dependent metalloenzymes. Archives of Biochemistry and Biophysics **544**(0): 142–152. doi:10.1016/j.abb.2013.09.002.

Bordenstein, S.R., and Reznikoff, W.S. 2005. Mobile DNA in obligate intracellular bacteria. Nature Reviews Microbiology **3**(9): 688. doi:10.1038/nrmicro1233.

Bordenstein, S.R., and Theis, K.R. 2015. Host biology in light of the microbiome: ten principles of holobionts and hologenomes. PLOS Biology **13**(8): e1002226. doi:10.1371/journal.pbio.1002226.

Borovok, I., Gorovitz, B., Schreiber, R., Aharonowitz, Y., and Cohen, G. 2006. Coenzyme B12 controls transcription of the Streptomyces class Ia ribonucleotide reductase nrdABS operon via a riboswitch mechanism. Journal of Bacteriology **188**(7): 2512–2520. doi:10.1128/JB.188.7.2512-2520.2006.

Boschen, R.E., Collins, P.C., Tunnicliffe, V., Carlsson, J., Gardner, J.P.A., Lowe, J., McCrone, A., Metaxas, A., Sinniger, F., and Swaddling, A. 2016. A primer for use of genetic tools in selecting and testing the suitability of set-aside sites protected from deep-sea seafloor massive sulfide mining activities. Ocean and Coastal Management **122**: 37–48. doi:10.1016/j.ocecoaman.2016.01.007.

Bossdorf, O., Richards, C.L., and Pigliucci, M. 2008. Epigenetics for ecologists. Ecology Letters **11**(2): 106–115. doi:10.1111/j.1461-0248.2007.01130.x.

Brand, G.L., Horak, R.V., Bris, N.L., Goffredi, S.K., Carney, S.L., Govenar, B., and Yancey, P.H. 2007. Hypotaurine and thiotaurine as indicators of sulfide exposure in bivalves and vestimentiferans from hydrothermal vents and cold seeps. Marine Ecology **28**(1): 208–218.

Breton, B.-A.A., Beaty, L., Bennett, A.M., Kyle, C.J., Lesbarrères, D., Vilaça, S.T., Wikston, M.J.H., Wilson, C.C., and Murray, D.L. 2022. Testing the effectiveness of environmental DNA (eDNA) to quantify larval amphibian abundance. Environmental DNA **4**(6): 1229–1240. doi:10.1002/edn3.332.

Breusing, C., Johnson, S.B., Vrijenhoek, R.C., and Young, C.R. 2019. Host hybridization as a potential mechanism of lateral symbiont transfer in deep-sea vesicomyid clams. Molecular Ecology **28**(21): 4697–4708. doi:10.1111/mec.15224.

Breusing, C., Franke, M., and Young, C.R. 2020. Intra-host symbiont diversity in eastern Pacific cold seep tubeworms identified by the 16S-V6 region, but undetected by the 16S-V4 region. PLoS ONE **15**(1): e0227053. doi:10.1371/journal.pone.0227053.

Breusing, C., Genetti, M., Russell, S.L., Corbett-Detig, R.B., and Beinart, R.A. 2022. Horizontal transmission enables flexible associations with locally adapted symbiont strains in deep-sea hydrothermal vent symbioses. Proceedings of the National Academy of Sciences **119**(14): e2115608119. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.2115608119.

Bright, M., and Lallier, F.H. 2010. The biology of vestimentiferan tubeworms. Oceanography and Marine Biology **48**: 213–266.

Brooke, S.D., and Young, C.M. 2009. Where do the embryos of *Riftia pachyptila* develop? Pressure tolerances, temperature tolerances, and buoyancy during prolonged embryonic dispersal. Deep Sea Research Part II: Topical Studies in Oceanography **56**(19): 1599–1606. doi:10.1016/j.dsr2.2009.05.003.

Brownlie, J.C., Adamski, M., Slatko, B., and McGraw, E.A. 2007. Diversifying selection and host adaptation in two endosymbiont genomes. BMC Evolutionary Biology **7**(1): 68. doi:10.1186/1471-2148-7-68.

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

Burke, G.R., and Moran, N.A. 2011. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. Genome Biology and Evolution **3**: 195–208. doi:10.1093/gbe/evr002.

Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. 2016. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications **7**: ncomms10613. doi:10.1038/ncomms10613.

Bushnell, B., Rood, J., and Singer, E. 2017. BBMerge − Accurate paired shotgun read merging via overlap. PLoS ONE **12**(10): e0185056. doi:10.1371/journal.pone.0185056.

Butterfield, D.A., McDuff, R.E., Lilley, M.D., Massoth, G.J., and Lupton, J.E. 1990. Geochemistry of hydrothermal fluids from Axial seamount hydrothermal emissions study vent field, Juan de Fuca ridge: subseafloor boiling and subsequent fluid-rock interaction. Journal of Geophysical Research **95**(B8), 12895-12921. doi:10.1029/JB095iB08p12895.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods **13**(7): 581−583. doi:10.1038/nmeth.3869.

Cambon-Bonavita, M.-A., Aubé, J., Cueff-Gauchard, V., and Reveillaud, J. 2021. Niche partitioning in the *Rimicaris exoculata* holobiont: the case of the first symbiotic Zetaproteobacteria. Microbiome **9**(1): 87. doi:10.1186/s40168-021-01045-6.

Cario, A., Jebbar, M., Thiel, A., Kervarec, N., and Oger, P.M. 2016. Molecular chaperone accumulation as a function of stress evidences adaptation to high hydrostatic pressure in the piezophilic archaeon *Thermococcus barophilus*. Scientific Reports **6**(1): 29483. doi:10.1038/srep29483.

Carney, S.L., Peoples, J.R., Fisher, C.R., and Schaeffer, S.W. 2002. AFLP analyses of genomic DNA reveal no differentiation between two phenotypes of the vestimentiferan tubeworm, *Ridgeia piscesae*. Cahiers de Biologie Marine **43**(3/4): 363–366.

Cary, S.C., and Giovannoni, S.J. 1993. Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. Proceedings of the National Academy of Sciences **90**(12): 5695–5699. doi:10.1073/pnas.90.12.5695.

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical Research **63**(3): 213–227. doi:10.1017/S0016672300032365.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. 2019. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics: btz848. doi:10.1093/bioinformatics/btz848.

Chen, N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics **5**(1): 4.10.1-4.10.14. doi:10.1002/0471250953.bi0410s05.

Chevin, L.-M., and Lande, R. 2011. Adaptation to marginal habitats by evolution of increased phenotypic plasticity. Journal of Evolutionary Biology **24**(7): 1462–1476. doi:10.1111/j.1420-9101.2011.02279.x.

Childress, J.J., Fisher, C.R., Favuzzi, J.A., and Sanders, N.K. 1991. Sulfide and carbon dioxide uptake by the hydrothermal vent clam, *Calyptogena magnifica*, and its chemoautotrophic symbionts. Physiological Zoology **64**(6): 1444–1470. doi:10.1086/physzool.64.6.30158224.

Cho, J.-C., and Tiedje, J.M. 2000. Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. Applied and Environmental Microbiology **66**(12): 5448–5456. doi:10.1128/AEM.66.12.5448-5456.2000.

Chong, R.A., and Moran, N.A. 2018. Evolutionary loss and replacement of *Buchnera*, the obligate endosymbiont of aphids. The ISME Journal **12**(3): 898. doi:10.1038/s41396-017-0024-6.

Chong, R.A., Park, H., and Moran, N.A. 2019. Genome evolution of the obligate endosymbiont *Buchnera aphidicola*. Molecular Biology and Evolution **36**(7): 1481–1489. doi:10.1093/molbev/msz082.

Corliss, J., Dymond, J, Gordon, LI, Herzen, RPV, Ballard, RD, Green, K, Williams, D, Bainbridge, A, Crane, K, and Vanandel, TH. 1979. Submarine thermal springs on the Galapagos Rift. Science **203**(4385): 1073-1083.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**(5673): 775–780. doi:10.1038/274775a0.

Cowart, D.A., Matabos, M., Brandt, M.I., Marticorena, J., and Sarrazin, J. 2020. Exploring environmental DNA (eDNA) to assess biodiversity of hard substratum faunal communities on the Lucky Strike vent field (Mid-Atlantic Ridge) and investigate recolonization dynamics after an induced disturbance. Frontiers in Marine Science **6**: 783. doi: 10.3389/fmars.2019.00783

Cruaud, P., Decker, C., Olu, K., Arnaud-Haond, S., Papot, C., Baut, J.L., Vigneron, A., Khripounoff, A., Gayet, N., Cathalot, C., Caprais, J.-C., Pignet, P., Godfroy, A., and Cambon-Bonavita, M.-A. 2019. Ecophysiological differences between vesicomyid species and metabolic capabilities of their symbionts influence distribution patterns of the deep-sea clams. Marine Ecology **40**(0): e12541. doi:10.1111/maec.12541.

Cuvelier, D., Sarradin, P.-M., Sarrazin, J., Colaço, A., Copley, J.T., Desbruyères, D., Glover, A.G., Santos, R.S., and Tyler, P.A. 2011. Hydrothermal faunal assemblages and habitat characterisation at the Eiffel Tower edifice (Lucky Strike, Mid-Atlantic Ridge). Marine Ecology **32**(2): 243–255. doi:10.1111/j.1439-0485.2010.00431.x.

Dale, C., and Moran, N.A. 2006. Molecular interactions between bacterial symbionts and their hosts. Cell **126**(3): 453–465. doi:10.1016/j.cell.2006.07.014.

Darling, A.E., Mau, B., and Perna, N.T. 2010. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE **5**(6): e11147. doi:10.1371/journal.pone.0011147.

de la Chaux, N., and Wagner, A. 2011. BEL/Pao retrotransposons in metazoan genomes. BMC Evolutionary Biology **11**(1): 154. doi:10.1186/1471-2148-11-154.

de la Rocha, C.L., and Passow, U. 2014. The biological pump. *In* Treatise on Geochemistry, Second Edition, vol. 8. Holland H.D. and Turekian K.K. pp. 93–122. doi:10.1016/B978-0-08-095975-7.00604-5.

de Mendoza, A., Hatleberg, W.L., Pang, K., Leininger, S., Bogdanovic, O., Pflueger, J., Buckberry, S., Technau, U., Hejnol, A., Adamska, M., Degnan, B.M., Degnan, S.M., and Lister, R. 2019a. Convergent evolution of a vertebrate-like methylome in a marine sponge. Nature Ecology and Evolution **3**(10): 1464–1473. doi:10.1038/s41559-019-0983-2.

de Mendoza, A., Pflueger, J., and Lister, R. 2019b. Capture of a functionally active methyl-CpG binding domain by an arthropod retrotransposon family. Genome Research **29**(8): 1277–1286. doi:10.1101/gr.243774.118.

de Mendoza, A., Lister, R., and Bogdanovic, O. 2020. Evolution of DNA methylome diversity in eukaryotes. Journal of Molecular Biology **432**(6): 1687–1705. doi:10.1016/j.jmb.2019.11.003.

de Oliveira, A.L., Mitchell, J., Girguis, P., and Bright, M. 2022. Novel insights on obligate symbiont lifestyle and adaptation to chemosynthetic environment as revealed by the giant tubeworm genome. Molecular Biology and Evolution **39**(1): msab347. doi:10.1093/molbev/msab347.

Decker, C., Olu, K., Arnaud-Haond, S., and Duperron, S. 2013. Physical proximity may promote lateral acquisition of bacterial symbionts in vesicomyid clams. PLoS ONE **8**(7): e64830. doi:10.1371/journal.pone.0064830.

Deniz, Ö., Frost, J.M., and Branco, M.R. 2019. Regulation of transposable elements by DNA modifications. Nature Reviews Genetics **20**(7): 417–431. doi:10.1038/s41576-019-0106-6.

Desbruyères, D., and Laubier, L. 1989. *Paralvinella hessleri*, new species of Alvinellidae (Polychaeta) from the Mariana Back-Arc basin hydrothermal vents. Proceedings of the Biological Society of Washington **102**(3): 761–767.

Di Meo, C.A., Wilbur, A.E., Holben, W.E., Feldman, R.A., Vrijenhoek, R.C., and Cary, S.C. 2000. Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. Applied and Environmental Microbiology **66**(2): 651–658.

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

Dilly, G.F., Young, C.R., Lane, W.S., Pangilinan, J., and Girguis, P.R. 2012. Exploring the limit of metazoan thermal tolerance via comparative proteomics: thermally induced changes in protein abundance by two hydrothermal vent polychaetes. Proceedings of the Royal Society B: Biological Sciences **279**(1741): 3347–3356. doi:10.1098/rspb.2012.0098.

Doyle, J.R., McKinnon, A.D., and Uthicke, S. 2017. Quantifying larvae of the coralivorous seastar *Acanthaster* cf. *solaris* on the Great Barrier Reef using qPCR. Marine Biology **164**(8): 176. doi:10.1007/s00227-017-3206-x.

Dray, S., and Dufour, A.-B. 2007. The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software **22**(4): 1–20. doi:10.18637/jss.v022.i04.

Duan, J.E., Jiang, Z.C., Alqahtani, F., Mandoiu, I., Dong, H., Zheng, X., Marjani, S.L., Chen, J., and Tian, X.C. 2019. Methylome dynamics of bovine gametes and in vivo early embryos. Frontiers in Genetics **10**: 512. doi: 10.3389/fgene.2019.00512.

Dubilier, N., Bergin, C., and Lott, C. 2008. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nature Reviews Microbiology **6**(10): 725–740. doi:10.1038/nrmicro1992.

Duperron, S., Bergin, C., Zielinski, F., Blazejak, A., Pernthaler, A., McKiness, Z.P., DeChaine, E., Cavanaugh, C.M., and Dubilier, N. 2006. A dual symbiosis shared by two mussel species, *Bathymodiolus* azoricus and *Bathymodiolus puteoserpentis* (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. Environmental Microbiology **8**(8): 1441–1447.

Duperron, S., De Beer, D., Zbinden, M., Boetius, A., Schipani, V., Kahil, N., and Gaill, F. 2009. Molecular characterization of bacteria associated with the trophosome and the tube of *Lamellibrachia* sp., a siboglinid annelid from cold seeps in the eastern Mediterranean. FEMS Microbiology Ecology **69**(3): 395–409.

Durkin, A., Fisher, C.R., and Cordes, E.E. 2017. Extreme longevity in a deep-sea vestimentiferan tubeworm and its implications for the evolution of life history strategies. The Science of Nature **104**(7–8): 63. doi:10.1007/s00114-017-1479-z.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**(5): 1792–1797. doi:10.1093/nar/gkh340.

Ehrlich, M., Norris, K.F., Wang, R.Y., Kuo, K.C., and Gehrke, C.W. 1986. DNA cytosine methylation and heat-induced deamination. Bioscience Reports **6**(4): 387–393. doi:10.1007/BF01116426.

Ellis, J.I., Clark, M.R., Rouse, H.L., and Lamarche, G. 2017. Environmental management frameworks for offshore mining: the New Zealand approach. Marine Policy **84**: 178–192. doi:10.1016/j.marpol.2017.07.004.

Elsaied, H., and Naganuma, T. 2001. Phylogenetic diversity of ribulose-1, 5-bisphosphate carboxylase/oxygenase large-subunit genes from deep-sea microorganisms. Applied and Environmental Microbiology **67**(4): 1751−1765.

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

Excoffier, L., Smouse, P.E., and Quattro, J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**(2): 479–491. doi: 10.1093/genetics/131.2.479

Excoffier, L., and Lischer, H.E.L. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10**(3): 564−567. doi:10.1111/j.1755-0998.2010.02847.x.

Fabre, L., Zhang, J., Guigon, G., Le Hello, S., Guibert, V., Accou-Demartin, M., de Romans, S., Lim, C., Roux, C., Passet, V., Diancourt, L., Guibourdenche, M., Issenhuth-Jeanjean, S., Achtman, M., Brisse, S., Sola, C., and Weill, F.-X. 2012. CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. PLoS ONE **7**(5): e36995. doi:10.1371/journal.pone.0036995.

Felsenstein, J. 1974. The evolutionary advantage of recombination. Genetics **78**(2): 737–756.

Fenchel, T. 2003. Biogeography for Bacteria. Science **301**(5635): 925–926. doi:10.1126/science.1089242.

Ferri, E., Bain, O., Barbuto, M., Martin, C., Lo, N., Uni, S., Landmann, F., Baccei, S.G., Guerrero, R., Lima, S. de S., Bandi, C., Wanji, S., Diagne, M., and Casiraghi, M. 2011. New Insights into the evolution of *Wolbachia* infections in filarial nematodes inferred from a large range of screened species. PLoS ONE **6**(6): e20843. doi:10.1371/journal.pone.0020843.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews Genetics **9**(2): 102–114. doi:10.1038/nrg2290.

Finkelstein, J.D. 1990. Methionine metabolism in mammals. Journal of Nutritional Biochemistry **1**(5): 228–237. doi:10.1016/0955-2863(90)90070-2.

Finlay, B.J., and Esteban, G.F. 2004. Ubiquitous dispersal of free-living microorganisms. Microbial Diversity and Bioprospecting: 216–224. doi:10.1128/9781555817770.ch21.

Flanders Marine Institute. 2022. Maritime Boundaries Geodatabase: Extended Continental Shelves, version 1. Available from https://www.marineregions.org/.

Flores, K., Wolschin, F., Corneveaux, J.J., Allen, A.N., Huentelman, M.J., and Amdam, G.V. 2012. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. BMC Genomics **13**: 480. doi:10.1186/1471-2164-13-480.

Fonseca, V.G. 2018. “Pitfalls in relative abundance estimation using eDNA metabarcoding.” Molecular Ecology Resources **18**(5): 923–926. doi:10.1111/1755-0998.12902.

Fontaine, F.J., Wilcock, W.S.D., Foustoukos, D.E., and Butterfield, D.A. 2009. A Si-Cl geothermobarometer for the reaction zone of high-temperature, basaltic-hosted mid-ocean ridge hydrothermal systems. Geochemistry, Geophysics, Geosystems **10**(5): Q05009. doi:10.1029/2009GC002407.

Fontanillas, E., Galzitskaya, O.V., Lecompte, O., Lobanov, M.Y., Tanguy, A., Mary, J., Girguis, P.R., Hourdez, S., and Jollivet, D. 2017. Proteome evolution of deep-sea hydrothermal vent alvinellid polychaetes supports the ancestry of thermophily and subsequent adaptation to cold in some lineages. Genome Biology and Evolution **9**(2): 279–296. doi:10.1093/gbe/evw298.

Fontecave, M., Atta, M., and Mulliez, E. 2004. S-adenosylmethionine: nothing goes to waste. Trends in Biochemical Sciences **29**(5): 243–249. doi:10.1016/j.tibs.2004.03.007.

Forget, N.L., Perez, M., and Juniper, S.K. 2014. Molecular study of bacterial diversity within the trophosome of the vestimentiferan tubeworm *Ridgeia piscesae*. Marine Ecology **36**: 35–44. doi:10.1111/maec.12169.

Fouquet, Y., Stackelberg, U. von, Charlou, J.L., Donval, J.P., Foucher, J.P., Erzinger, J., Herzig, P., Mühe, R., Wiedicke, M., Soakai, S., and Whitechurch, H. 1991. Hydrothermal activity in the Lau back-arc basin: sulfides and water chemistry. Geology **19**(4): 303–306.

Funk, D.J., Wernegreen, J.J., and Moran, N.A. 2001. Intraspecific variation in symbiont genomes: bottlenecks and the aphid-*Buchnera* association. Genetics **157**(2): 477–489.

Gagnière, N., Jollivet, D., Boutet, I., Brélivet, Y., Busso, D., Da Silva, C., Gaill, F., Higuet, D., Hourdez, S., Knoops, B., Lallier, F., Leize-Wagner, E., Mary, J., Moras, D., Perrodou, E., Rees, J.-F., Segurens, B., Shillito, B., Tanguy, A., Thierry, J.-C., Weissenbach, J., Wincker, P., Zal, F., Poch, O., and Lecompte, O. 2010. Insights into metazoan evolution from *Alvinella pompejana* cDNAs. BMC Genomics **11**: 634. doi:10.1186/1471-2164-11-634.

Garcia-Vazquez, E., Georges, O., Fernandez, S., and Ardura, A. 2021. eDNA metabarcoding of small plankton samples to detect fish larvae and their preys from Atlantic and Pacific waters. Scientific Reports **11**(1): 7224. doi:10.1038/s41598-021-86731-z.

Gardebrecht, A., Markert, S., Sievert, S.M., Felbeck, H., Thürmer, A., Albrecht, D., Wollherr, A., Kabisch, J., Le Bris, N., and Lehmann, R. 2012. Physiological homogeneity among the endosymbionts of *Riftia pachyptila* and *Tevnia jerichonana* revealed by proteogenomics. The ISME Journal **6**(4): 766–776.

Gatzmann, F., Falckenhayn, C., Gutekunst, J., Hanna, K., Raddatz, G., Carneiro, V.C., and Lyko, F. 2018. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. Epigenetics and Chromatin **11**(1): 57. doi:10.1186/s13072-018-0229-6.

Gavery, M.R., and Roberts, S.B. 2014. A context dependent role for DNA methylation in bivalves. Briefings in Functional Genomics **13**(3): 217–222. doi:10.1093/bfgp/elt054.

German, C.R., Ramirez-Llodra, E., Baker, M.C., Tyler, P.A., and Committee, and the C.S.S. 2011. Deep-water chemosynthetic ecosystem research during the census of marine life decade and beyond: a proposed deep-ocean road map. PLoS ONE **6**(8): e23259. doi:10.1371/journal.pone.0023259.

Gerrish, P.J., and Lenski, R.E. 1998. The fate of competing beneficial mutations in an asexual population. Genetica **102**(0): 127. doi:10.1023/A:1017067816551.

Gil, R., and Belda, E. 2008. Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*. International Microbiology **11**(1): 41–48. doi:10.2436/20.1501.01.43.

Gilbert, S.F. 2014. A holobiont birth narrative: the epigenetic transmission of the human microbiome. Frontiers in Genetics **5**: 282. doi:10.3389/fgene.2014.00282.

Gillespie, J.H. 2000. Genetic drift in an infinite population: The pseudohitchhiking model. Genetics **155**(2): 909–919. doi:10.1093/genetics/155.2.909.

Giovannoni, S.J., Cameron Thrash, J., and Temperton, B. 2014. Implications of streamlining theory for microbial ecology. The ISME Journal **8**(8): 1553–1565. doi:10.1038/ismej.2014.60.

Girguis, P.R., and Lee, R.W. 2006. Thermal preference and tolerance of alvinellids. Science **312**(5771): 231–231. doi:10.1126/science.1125286.

Giroux, M.S., Reichman, J.R., Langknecht, T., Burgess, R.M., and Ho, K.T. 2022. Environmental RNA as a tool for marine community biodiversity assessments. Scientific Reports **12**(1): 17782. doi:10.1038/s41598-022-22198-w.

Glaeser, S.P., and Kämpfer, P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Systematic and Applied Microbiology **38**(4): 237–245. doi:10.1016/j.syapm.2015.03.007.

Glöckner, F.O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. 2000. Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of Actinobacteria. Applied and Environmental Microbiology **66**(11): 5053–5065. doi:10.1128/AEM.66.11.5053-5065.2000.

Godde, J.S., and Bickerton, A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. Journal of Molecular Evolution **62**(6): 718–729. doi:10.1007/s00239-005-0223-z.

Goffredi, S.K., and Barry, J.P. 2002. Species-specific variation in sulfide physiology between closely related vesicomyid clams. Marine Ecology Progress Series **225**: 227–238. doi:10.3354/meps225227.

Goldman, N., and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution **11**(5): 725–736. doi:10.1093/oxfordjournals.molbev.a040153.

Gollner, S., Kaiser, S., Menzel, L., Jones, D.O.B., Brown, A., Mestre, N.C., van Oevelen, D., Menot, L., Colaço, A., Canals, M., Cuvelier, D., Durden, J.M., Gebruk, A., Egho, G.A., Haeckel, M., Marcon, Y., Mevenkamp, L., Morato, T., Pham, C.K., Purser, A., Sanchez-Vidal, A., Vanreusel, A., Vink, A., and Martinez Arbizu, P. 2017. Resilience of benthic deep-sea fauna to mining activities. Marine Environmental Research **129**: 76–101. doi:10.1016/j.marenvres.2017.04.010.

González, J.C., Peariso, K., Penner-Hahn, J.E., and Matthews, R.G. 1996. Cobalamin-Independent Methionine Synthase from *Escherichia coli*: a zinc metalloenzyme. Biochemistry **35**(38): 12228–12234. doi:10.1021/bi9615452.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology **29**(7): 644–652. doi:10.1038/nbt.1883.

Grelon, D., Morineaux, M., Desrosiers, G., and Juniper, S.K. 2006. Feeding and territorial behavior of *Paralvinella sulfincola*, a polychaete worm at deep-sea hydrothermal vents of the Northeast Pacific Ocean. Journal of Experimental Marine Biology and Ecology **329**(2): 174–186. doi:10.1016/j.jembe.2005.08.017.

Grzymski, J.J., and Dussaq, A.M. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. The ISME Journal **6**(1): 71–80. doi:10.1038/ismej.2011.72.

Guillon, E., Menot, L., Decker, C., Krylova, E., and Olu, K. 2017. The vesicomyid bivalve habitat at cold seeps supports heterogeneous and dynamic macrofaunal assemblages. Deep Sea Research Part I: Oceanographic Research Papers **120**: 1–13. doi:10.1016/j.dsr.2016.12.008.

Gunderson, F.F., and Cianciotto, N.P. 2013. The CRISPR-associated gene cas2 of *Legionella pneumophila* is required for intracellular infection of Amoebae. mBio **4**(2). doi:10.1128/mBio.00074-13.

Halfmann, R., and Lindquist, S. 2010. Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. Science **330**(6004): 629–632. doi:10.1126/science.1191081.

Han, Y., Zhang, D., Wang, C., and Zhou, Y. 2021. Out of the Pacific: a new alvinellid worm (Annelida: Terebellida) from the northern Indian Ocean hydrothermal vents. Frontiers in Marine Science **8**: 669918. doi:10.3389/fmars.2021.669918.

Hansen, A.K., and Moran, N.A. 2014. The impact of microbial symbionts on host plant utilization by herbivorous insects. Molecular Ecology **23**(6): 1473–1496. doi:10.1111/mec.12421.

Harmer, T.L., Rotjan, R.D., Nussbaumer, A.D., Bright, M., Ng, A.W., DeChaine, E.G., and Cavanaugh, C.M. 2008. Free-living tube worm endosymbionts found at deep-sea vents. Applied and Environmental Microbiology **74**(12): 3895–3898.

Harrell, F.E.J. 2019. Package 'hmisc.' CRAN2018 **2019**: 235–236.

Hashimoto, J., Ohta, S., Gamo, T., Chiba, H., Yamaguchi, T., Tsuchida, S., Okudaira, T., Watabe, H., Yamanaka, T., and Kitazawa, M. 2001. First hydrothermal vent communities from the Indian Ocean discovered. Journal of Zoology **18**(5): 717–721. doi:10.2108/zsj.18.717.

Held, N.L., Herrera, A., Cadillo-Quiroz, H., and Whitaker, R.J. 2010. CRISPR associated diversity within a population of *Sulfolobus islandicus*. PLoS ONE **5**(9): e12988. doi:10.1371/journal.pone.0012988.

Herbeck, J.T., Funk, D.J., Degnan, P.H., and Wernegreen, J.J. 2003. A conservative test of genetic drift in the endosymbiotic bacterium *Buchnera*: slightly deleterious mutations in the chaperonin groEL. Genetics **165**(4): 1651–1660.

Hershberg, R., and Petrov, D.A. 2010. Evidence that mutation is universally biased towards AT in Bacteria. PLoS Genetics **6**(9): e1001115. doi:10.1371/journal.pgen.1001115.

Hessler, R.R., Smithey, W.M., Boudrias, M.A., Keller, C.H., Lutz, R.A., and Childress, J.J. 1988. Temporal change in megafauna at the Rose Garden hydrothermal vent (Galapagos Rift; eastern tropical Pacific). Deep Sea Research Part A. Oceanographic Research Papers **35**(10): 1681–1709. doi:10.1016/0198-0149(88)90044-1.

Heyl, T.P., Gilhooly, W.P., Chambers, R.M., Gilchrist, G.W., Macko, S.A., Ruppel, C.D., and Dover, C.L.V. 2007. Characteristics of vesicomyid clams and their environment at the Blake Ridge cold seep, South Carolina, USA. Marine Ecology Progress Series **339**: 169–184. doi:10.3354/meps339169.

Hilário, A., Young, C.M., and Tyler, P.A. 2005. Sperm storage, internal fertilization, and embryonic dispersal in vent and seep tubeworms (Polychaeta: Siboglinidae: Vestimentifera). The Biological Bulletin **208**(1): 20–28. doi:10.2307/3593097.

Hill, W.G., and Robertson, A. 1966. The effect of linkage on limits to artificial selection. Genetical Research **8**(3): 269–294.

Hinzke, T., Kleiner, M., Breusing, C., Felbeck, H., Häsler, R., Sievert, S.M., Schlüter, R., Rosenstiel, P., Reusch, T.B.H., Schweder, T., and Markert, S. 2019. Host-microbe interactions in the chemosynthetic *Riftia pachyptila* symbiosis. mBio **10**(6): 20.

Hitchins, M.P. 2015. Constitutional epimutation as a mechanism for cancer causality and heritability? Nature Reviews Cancer **15**(10): 625–634. doi:10.1038/nrc4001.

Ho, P.-T., Park, E., Hong, S.G., Kim, E.-H., Kim, K., Jang, S.-J., Vrijenhoek, R.C., and Won, Y.-J. 2017. Geographical structure of endosymbiotic bacteria hosted by *Bathymodiolus* mussels at eastern Pacific hydrothermal vents. BMC Evolutionary Biology **17**(1): 121.

Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12**(1): 491. doi:10.1186/1471-2105-12-491.

Hotopp, J.C.D., Clark, M.E., Oliveira, D.C.S.G., Foster, J.M., Fischer, P., Torres, M.C.M., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R.V., Shepard, J., Tomkins, J., Richards, S., Spiro, D.J., Ghedin, E., Slatko, B.E., Tettelin, H., and Werren, J.H. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science **317**(5845): 1753–1756. doi:10.1126/science.1142490.

Howe, D.K., and Denver, D.R. 2008. Muller's Ratchet and compensatory mutation in *Caenorhabditis briggsae* mitochondrial genome evolution. BMC Evolutionary Biology **8**(1): 62. doi:10.1186/1471-2148-8-62.

Hughes, D. 2000. Co-evolution of the tuf genes links gene conversion with the generation of chromosomal inversions. Journal of Molecular Biology **297**(2): 355–364. doi:10.1006/jmbi.2000.3587.

Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A.C.C., von Dohlen, C.D., Fukatsu, T., and McCutcheon, J.P. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell **153**(7): 1567–1578. doi:10.1016/j.cell.2013.05.040.

Ihaka, R., and Gentleman, R. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics **5**(3): 299–314. doi:10.1080/10618600.1996.10474713.

Ikuta, T., Kanae, I., Akihiro, T., Tsuneyoshi, K., Haruko, K., Aoki Yui, Takaki Yoshihiro, Nagai Yukiko, Ozawa Genki, Yamamoto Masahiro, Deguchi Ryusaku, Fujikura Katsunori, Maruyama Tadashi, and Yoshida Takao. 2016. Surfing the vegetal pole in a small population: extracellular vertical transmission of an "intracellular" deep-sea clam symbiont. Royal Society Open Science **3**(5): 160130. doi:10.1098/rsos.160130.

International Seabed Authority. 2010. ISBA/16/A/12/Rev.1: Decision of the assembly of the international seabed authority relating to the regulations on prospecting and exploration for polymetallic sulphides in the area. Available from https://www.isa.org.jm/mining_code/2010-isba-16-a-12-rev-1/.

International Seabed Authority. 2018. ISBA/24/A/4: Consideration, with a view to adoption, of the draft strategic plan of the International Seabed Authority for the period 2019–2023. Available from https://www.isa.org.jm/document/isba24a4.

International Seabed Authority. 2019, March. ISBA/25/C/WP.1: Draft regulations on exploitation of mineral resources in the Area. Available from https://isa.org.jm/files/files/documents/isba_25_c_wp1-e_0.pdf.

Ip, J.C.-H., Xu, T., Sun, J., Li, R., Chen, C., Lan, Y., Han, Z., Zhang, H., Wei, J., Wang, H., Tao, J., Cai, Z., Qian, P.-Y., and Qiu, J.-W. 2020. Host-endosymbiont genome integration in a deep-sea chemosymbiotic clam. Molecular Biology and Evolution **38**(2): 502–518. doi:10.1093/molbev/msaa241.

Itoh, T., Martin, W., and Nei, M. 2002. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proceedings of the National Academy of Sciences **99**(20): 12944–12948. doi:10.1073/pnas.192449699.

Jaenisch, R., and Bird, A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics **33**(3s): 245–254. doi:10.1038/ng1089.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications **9**(1): 5114. doi:10.1038/s41467-018-07641-9.

Jamieson, J.W., Hannington, M.D., Clague, D.A., Kelley, D.S., Delaney, J.R., Holden, J.F., Tivey, M.K., and Kimpe, L.E. 2013. Sulfide geochronology along the Endeavour Segment of the Juan de Fuca Ridge. Geochemistry, Geophysics, Geosystems **14**(7): 2084–2099. doi:10.1002/ggge.20133.

Jang, S.-J., Chung, Y., Jun, S., and Won, Y.-J. 2022. Connectivity and divergence of symbiotic bacteria of deep-sea hydrothermal vent mussels in relation to the structure and dynamics of mid-ocean ridges. Frontiers in Marine Science **9**: 845965. doi: 10.3389/fmars.2022.845965

Jeremias, G., Barbosa, J., Marques, S.M., Asselman, J., Gonçalves, F.J.M., and Pereira, J.L. 2018. Synthesizing the role of epigenetics in the response and adaptation of species to climate change in freshwater ecosystems. Molecular Ecology **27**(13): 2790-2806. doi:10.1111/mec.14727.

Jiang, Q., Lin, D., Huang, H., Wang, G., and Ye, H. 2020. DNA methylation inhibits the expression of CFSH in mud crab. Frontiers in Endocrinology **11**: 163. doi: 10.3389/fendo.2020.00163

Johnson, S.B., Krylova, E.M., Audzijonyte, A., Sahling, H., and Vrijenhoek, R.C. 2017. Phylogeny and origins of chemosynthetic vesicomyid clams. Systematics and Biodiversity **15**(4): 346–360. doi:10.1080/14772000.2016.1252438.

Jones, D.S., Schaperdoth, I., and Macalady, J.L. 2016. Biogeography of sulfur-oxidizing *Acidithiobacillus* populations in extremely acidic cave biofilms. The ISME Journal **10**(12): 2879–2891. doi:10.1038/ismej.2016.74.

Joye, S.B. 2020. The geology and biogeochemistry of hydrocarbon seeps. Annual Review of Earth and Planetary Sciences **48**(1): 205–231. doi:10.1146/annurev-earth-063016-020052.

Jun, J., Won, Y.-J., and Vrijenhoek, R.C. 2016. Complete mitochondrial genome of the hydrothermal vent tubeworm, *Ridgeia piscesae* (Polychaeta, Siboglinidae). Mitochondrial DNA Part A: DNA Mapping, Sequencing, and Analysis **27**(2): 1123–1124. doi:10.3109/19401736.2014.933330.

Juniper, S.K., and Tunnicliffe, V. 1997. Crustal accretion and the hot vent ecosystem. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **355**(1723): 459–474. doi:10.1098/rsta.1997.0017.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M., Ashburner, M., and Celniker, S.E. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biology **3**(12): 0084.1. doi:10.1186/gb-2002-3-12-research0084.

Kamvar, Z.N., Tabima, J.F., and Grünwald, N.J. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ **2**: e281. doi:10.7717/peerj.281.

Keller, T.E., Han, P., and Yi, S.V. 2016. Evolutionary transition of promoter and gene body DNA methylation across invertebrate–vertebrate boundary. Molecular Biology and Evolution **33**(4): 1019–1028. doi:10.1093/molbev/msv345.

Kim, A., Terzian, C., Santamaria, P., Pélisson, A., Purd'homme, N., and Bucheton, A. 1994. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proceedings of the National Academy of Sciences **91**(4): 1285–1289. doi:10.1073/pnas.91.4.1285.

Kim, M., Oh, H.-S., Park, S.-C., and Chun, J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal of Systematic and Evolutionary Microbiology **64**(2): 346–351. doi:10.1099/ijs.0.059774-0.

Kim, M., Kang, J.-H., and Kim, D. 2022. Holoplanktonic and meroplanktonic larvae in the surface waters of the Onnuri vent field in the Central Indian Ridge. Journal of Marine Science and Engineering **10**(2): 158. doi:10.3390/jmse10020158.

Kimura, H., Higashide, Y., and Naganuma, T. 2003. Endosymbiotic microflora of the vestimentiferan tubeworm (*Lamellibrachia* sp.) from a bathyal cold seep. Marine Biotechnology **5**(6): 593–603.

Kimura, M., and Weiss, G.H. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49**(4): 561–576.

Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., Gerlach, D., Kriventseva, E.V., Elsik, C.G., Graur, D., Hill, C.A., Veenstra, J.A., Walenz, B., Tubío, J.M.C., Ribeiro, J.M.C., Rozas, J., Johnston, J.S., Reese, J.T., Popadic, A., Tojo, M., Raoult, D., Reed, D.L., Tomoyasu, Y., Kraus, E., Mittapalli, O., Margam, V.M., Li, H.-M., Meyer, J.M., Johnson, R.M., Romero-Severson, J., VanZee, J.P., Alvarez-Ponce, D., Vieira, F.G., Aguadé, M., Guirao-Rico, S., Anzola, J.M., Yoon, K.S., Strycharz, J.P., Unger, M.F., Christley, S., Lobo, N.F., Seufferheld, M.J., Wang, N., Dasch, G.A., Struchiner, C.J., Madey, G., Hannick, L.I., Bidwell, S., Joardar, V., Caler, E., Shao, R., Barker, S.C., Cameron, S., Bruggner, R.V., Regier, A., Johnson, J., Viswanathan, L., Utterback, T.R., Sutton, G.G., Lawson, D., Waterhouse, R.M., Venter, J.C., Strausberg, R.L., Berenbaum, M.R., Collins, F.H., Zdobnov, E.M., and Pittendrigh, B.R. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proceedings of the National Academy of Sciences **107**(27): 12168–12173. doi:10.1073/pnas.1003379107.

Klose, J., Polz, M.F., Wagner, M., Schimak, M.P., Gollner, S., and Bright, M. 2015. Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. Proceedings of the National Academy of Sciences **112**(36): 11300–11305. doi:10.1073/pnas.1501160112.

Klose, R.J., and Bird, A.P. 2006. Genomic DNA methylation: the mark and its mediators. Trends in Biochemical Sciences **31**(2): 89–97. doi:10.1016/j.tibs.2005.12.008.

Klymus, K.E., Merkes, C.M., Allison, M.J., Goldberg, C.S., Helbing, C.C., Hunter, M.E., Jackson, C.A., Lance, R.F., Mangan, A.M., Monroe, E.M., Piaggio, A.J., Stokdyk, J.P., Wilson, C.C., and Richter, C.A. 2020. Reporting the limits of detection and quantification for environmental DNA assays. Environmental DNA **2**(3): 271–282. doi:10.1002/edn3.29.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., and Ding, L. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics **25**(17): 2283–2285. doi:10.1093/bioinformatics/btp373.

Koga, R., and Moran, N.A. 2014. Swapping symbionts in spittlebugs: evolutionary replacement of a reduced genome symbiont. The ISME Journal **8**(6): 1237–1246. doi:10.1038/ismej.2013.235.

Kojima, K.K. 2019. Structural and sequence diversity of eukaryotic transposable elements. Genes and Genetic Systems **94**(6): 233–252. doi:10.1266/ggs.18-00024.

Kovanen, S. m., Kivistö, R. i., Rossi, M., and Hänninen, M.-L. 2014. A combination of MLST and CRISPR typing reveals dominant *Campylobacter jejuni* types in organically farmed laying hens. Journal of Applied Microbiology **117**(1): 249–257. doi:10.1111/jam.12503.

Krylova, E.M., and Sahling, H. 2010. Vesicomyidae (Bivalvia): current taxonomy and distribution. PLoS ONE **5**(4): e9957. doi:10.1371/journal.pone.0009957.

Kuhner, M.K., and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution. doi:10.1093/oxfordjournals.molbev.a040126.

Kuno, S., Sako, Y., and Yoshida, T. 2014. Diversification of CRISPR within coexisting genotypes in a natural population of the bloom-forming cyanobacterium *Microcystis aeruginosa*. Microbiology **160**(5): 903–916. doi:10.1099/mic.0.073494-0.

Kuo, C.-H., Moran, N.A., and Ochman, H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Research **19**(8): 1450–1454. doi:10.1101/gr.091785.109.

Kupczok, A., and Bollback, J.P. 2013. Probabilistic models for CRISPR spacer content evolution. BMC Evolutionary Biology **13**(1): 54. doi:10.1186/1471-2148-13-54.

Kupczok, A., Landan, G., and Dagan, T. 2015. The contribution of genetic recombination to CRISPR array evolution. Genome Biology and Evolution **7**(7): 1925–1939. doi:10.1093/gbe/evv113.

Kuwahara, H., Yoshida, T., Takaki, Y., Shimamura, S., Nishi, S., Harada, M., Matsuyama, K., Takishita, K., Kawato, M., Uematsu, K., Fujiwara, Y., Sato, T., Kato, C., Kitagawa, M., Kato, I., and Maruyama, T. 2007. Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. Current Biology **17**(10): 881–886. doi:10.1016/j.cub.2007.04.039.

Kuwahara, H., Takaki, Y., Yoshida, T., Shimamura, S., Takishita, K., Reimer, J.D., Kato, C., and Maruyama, T. 2008. Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams. Extremophiles **12**(3): 365–374. doi:10.1007/s00792-008-0141-2.

Kuwahara, H., Takaki, Y., Shimamura, S., Yoshida, T., Maeda, T., Kunieda, T., and Maruyama, T. 2011. Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of *Calyptogena* clams. BMC Evolutionary Biology **11**(1): 285. doi:10.1186/1471-2148-11-285.

Kvist, S., and Oceguera-Figueroa, A. 2021. Phylum Annelida. *In* Invertebrate Zoology. CRC Press. pp. 311-328

Lambert, J.D., and Moran, N.A. 1998. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. Proceedings of the National Academy of Sciences **95**(8): 4458–4462.

Lan, Y., Sun, J., Chen, C., Wang, H., Xiao, Y., Perez, M., Yang, Y., Kwan, Y.H., Sun, Y., Zhou, Y., Han, X., Miyazaki, J., Watsuji, T., Bissessur, D., Qiu, J.-W., Takai, K., and Qian, P.-Y. 2022. Endosymbiont population genomics sheds light on transmission mode, partner

specificity, and stability of the scaly-foot snail holobiont. The ISME Journal **16**(9): 2132-2143. doi:10.1038/s41396-022-01261-4.

Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4): 357–359. doi:10.1038/nmeth.1923.

Larget, B.R., Kotha, S.K., Dewey, C.N., and Ané, C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics **26**(22): 2910–2911. doi:10.1093/bioinformatics/btq539.

Lathe, W.C., and Bork, P. 2001. Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Letters **502**(3): 113–116. doi:10.1016/S0014-5793(01)02639-4.

Latorre, A., and Manzano-Marín, A. 2017. Dissecting genome reduction and trait loss in insect endosymbionts. Annals of the New York Academy of Sciences **1389**(1): 52–75. doi:10.1111/nyas.13222.

Lawrence, J.G., Hendrix, R.W., and Casjens, S. 2001. Where are the pseudogenes in bacterial genomes? Trends in Microbiology **9**(11): 535–540. doi:10.1016/S0966-842X(01)02198-9.

Le Bris, N., and Gaill, F. 2007. How does the annelid *Alvinella pompejana* deal with an extreme hydrothermal environment? Reviews in Environmental Science and Biotechnology **6**(1–3): 197. doi:10.1007/s11157-006-9112-1.

Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., Hartmann, R.K., Prohaska, S.J., and Stadler, P.F. 2014. Orthology detection combining clustering and synteny for very large datasets. PLoS ONE **9**(8): e105015. doi:10.1371/journal.pone.0105015.

Lee, R.W., Robert, K., Matabos, M., Bates, A.E., and Juniper, S.K. 2015. Temporal and spatial variation in temperature experienced by macrofauna at Main Endeavour hydrothermal vent field. Deep-Sea Research Part I: Oceanographic Research Papers **106**: 154–166. doi:10.1016/j.dsr.2015.10.004.

Lee, Y., Kwak, H., Shin, J., Kim, S.-C., Kim, T., and Park, J.-K. 2019. A mitochondrial genome phylogeny of Mytilidae (Bivalvia: Mytilida). Molecular Phylogenetics and Evolution **139**: 106533. doi:10.1016/j.ympev.2019.106533.

Lee, W.-K., Juniper, S.K., Perez, M., Ju, S.-J., and Kim, S.-J. 2021. Diversity and characterization of bacterial communities of five co-occurring species at a hydrothermal vent on the Tonga Arc. Ecology and Evolution **11**(9): 4481-4493**.** doi:10.1002/ece3.7343.

Lelièvre, Y., Sarrazin, J., Marticorena, J., Schaal, G., Day, T., Legendre, P., Hourdez, S., and Matabos, M. 2017. Biodiversity and trophic ecology of hydrothermal vent fauna associated with tubeworm assemblages on the Juan de Fuca Ridge. Biogeosciences Discussions **2017**: 1–34. doi:10.5194/bg-2017-411.

Leung, C., Breton, S., and Angers, B. 2016. Facing environmental predictability with different sources of epigenetic variation. Ecology and Evolution **6**(15): 5234–5245. doi:10.1002/ece3.2283.

Levesque, C., Juniper, S.K., and Marcus, J. 2003. Food resource partitioning and competition among alvinellid polychaetes of Juan de Fuca Ridge hydrothermal vents. Marine Ecology Progress Series **246**: 173–182.

Levin, L.A., Baco, A.R., Bowden, D.A., Colaco, A., Cordes, E.E., Cunha, M.R., Demopoulos, A.W.J., Gobin, J., Grupe, B.M., Le, J., Metaxas, A., Netburn, A.N., Rouse, G.W., Thurber, A.R., Tunnicliffe, V., Dover, V., Lee, C., Vanreusel, A., and Watling, L. 2016. Hydrothermal vents and methane seeps: rethinking the sphere of influence. Frontiers in Marine Science **3**: 72. doi:10.3389/fmars.2016.00072.

Lewis, S.H., Ross, L., Bain, S.A., Pahita, E., Smith, S.A., Cordaux, R., Miska, E.A., Lenhard, B., Jiggins, F.M., and Sarkies, P. 2020. Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. PLoS Genetics **16**(6): e1008864. doi:10.1371/journal.pgen.1008864.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16): 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, Y., Kocot, K.M., Whelan, N.V., Santos, S.R., Waits, D.S., Thornhill, D.J., and Halanych, K.M. 2017. Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. Zoologica Scripta **46**(2): 200–213. doi:10.1111/zsc.12201.

Li, Y., Liles, M.R., and Halanych, K.M. 2018. Endosymbiont genomes yield clues of tubeworm success. The ISME Journal **12**(11): 2785-2795. doi:10.1038/s41396-018-0220-z.

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

Liu, H., Cai, S., Zhang, H., and Vrijenhoek, R.C. 2016. Complete mitochondrial genome of hydrothermal vent clam *Calyptogena magnifica*. Mitochondrial DNA Part A: DNA Mapping, Sequencing, and Analysis **27**(6): 4333–4335. doi:10.3109/19401736.2015.1089488.

Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., and Li, S. 2021. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biology **22**(1): 295. doi:10.1186/s13059-021-02510-z.

Lowe, W.H., and Allendorf, F.W. 2010. What can genetics tell us about population connectivity? Molecular Ecology **19**(15): 3038–3051. doi:10.1111/j.1365-294X.2010.04688.x.

Luo, H., Huang, Y., Stepanauskas, R., and Tang, J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nature Microbiology **2**(8): 1–9. doi:10.1038/nmicrobiol.2017.91.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., and Maleszka, R. 2010. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol **8**(11): e1000506. doi:10.1371/journal.pbio.1000506.

Lyko, F. 2018. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nature Reviews Genetics **19**(2): 81–92. doi:10.1038/nrg.2017.80.

Mandrioli, M. 2007. A new synthesis in epigenetics: towards a unified function of DNA methylation from invertebrates to vertebrates. Cellular and Molecular Life Sciences **64**(19–20): 2522. doi:10.1007/s00018-007-7231-7.

Markert, S., Arndt, C., Felbeck, H., Becher, D., Sievert, S.M., Hügler, M., Albrecht, D., Robidart, J., Bench, S., Feldman, R.A., Hecker, M., and Schweder, T. 2007. Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. Science **315**(5809): 247–250. doi:10.1126/science.1132913.

Marsh, A.G., and Pasqualone, A.A. 2014. DNA methylation and temperature stress in an Antarctic polychaete, *Spiophanes tcherniai*. Frontiers in Physiology **5**: 173. doi:10.3389/fphys.2014.00173.

Martínez-Cano, D.J., Reyes-Prieto, M., Martínez-Romero, E., Partida-Martínez, L.P., Latorre, A., Moya, A., and Delaye, L. 2015. Evolution of small prokaryotic genomes. Frontiers in Microbiology **5**: 742. doi:10.3389/fmicb.2014.00742.

Massicotte, R., Whitelaw, E., and Angers, B. 2011. DNA methylation: a source of random variation in natural populations. Epigenetics **6**(4): 421-427.

McCartin, L.J., Vohsen, S.A., Ambrose, S.W., Layden, M., McFadden, C.S., Cordes, E.E., McDermott, J.M., and Herrera, S. 2022. Temperature controls eDNA persistence across physicochemical conditions in seawater. Environmental Science and Technology **56**(12): 8629–8639. doi:10.1021/acs.est.2c01672.

McCaw, B.A., Stevenson, T.J., and Lancaster, L.T. 2020. Epigenetic responses to temperature and climate. Integrative and Comparative Biology **60**:1469–1480. doi:10.1093/icb/icaa049.

McCutcheon, J.P., and Moran, N.A. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proceedings of the National Academy of Sciences **104**(49): 19392–19397. doi:10.1073/pnas.0708855104.

McCutcheon, J.P., and Moran, N.A. 2012. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology **10**(1): 13–26. doi:10.1038/nrmicro2670.

McHugh, D. 1989. Population structure and reproductive biology of two sympatric hydrothermal vent polychaetes, *Paralvinella pandorae* and *P. palmiformis*. Marine Biology 103(1): 95–106. doi:10.1007/BF00391068.

McManus, D.A., Holmes, M.L., Carson, B., and Barr, S.M. 1972. Late Quaternary tectonics, northern end of Juan de Fuca Ridge (Northeast Pacific). Marine Geology **12**(2): 141−164. doi:10.1016/0025-3227(72)90025-4.

McMullin, E.R., Hourdez, S., Schaeffer, S.W., and Fisher, C.R. 2003. Phylogeny and biogeography of deep sea vestimentiferan tubeworms and their bacterial symbionts. Symbiosis **34**(1): 1–41.

Meier, D.V., Pjevac, P., Bach, W., Hourdez, S., Girguis, P.R., Vidoudez, C., Amann, R., and Meyerdierks, A. 2017. Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. The ISME Journal **11**(7): 1545–1558. doi:10.1038/ismej.2017.37.

METI. 2017. World's first success in continuous ore lifting test for seafloor polymetallic sulphides. Available from http://www.meti.go.jp/english/press/2017/0926_004.html [accessed 1 June 2018].

Miller, K.A., Thompson, K.F., Johnston, P., and Santillo, D. 2018. An overview of seabed mining including the current state of development, environmental impacts, and knowledge gaps. Frontiers in Marine Science **4**: 418.

Mira, A., Ochman, H., and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics **17**(10): 589–596. doi:10.1016/S0168-9525(01)02447-7.

Mitarai, S., Watanabe, H., Nakajima, Y., Shchepetkin, A.F., and McWilliams, J.C. 2016. Quantifying dispersal from hydrothermal vent fields in the western Pacific Ocean. Proceedings of the National Academy of Sciences **113**(11): 2976–2981. doi:10.1073/pnas.1518395113.

Moggioli, G., Panossian, B., Sun, Y., Thiel, D., Martín-Zamora, F.M., Tran, M., Clifford, A.M., Goffredi, S.K., Rimskaya-Korsakova, N., Jékely, G., Tresguerres, M., Qian, P.-Y., Qiu, J.-W., Rouse, G.W., Henry, L.M., and Martín-Durán, J.M. 2023. Distinct genomic routes underlie transitions to specialised symbiotic lifestyles in deep-sea annelid worms. Nature Communications **14**(1): 2814. doi:10.1038/s41467-023-38521-6.

Moller, A.G., and Liang, C. 2017. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. PeerJ **5**: e3788. doi:10.7717/peerj.3788.

MolluscaBase. 2019. MolluscaBase. Vesicomyidae Dall and Simpson, 1901. Available from http://www.marinespecies.org/aphia.php?p=taxdetailsandid=23140#distributions [accessed 30 April 2019].

Moran, N.A., Munson, M.A., Baumann, P., and Ishikawa, H. 1993. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proceedings of the Royal Society B: Biological Sciences **253**(1337): 167–171. doi:10.1098/rspb.1993.0098.

Moran, N.A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proceedings of the National Academy of Sciences **93**(7): 2873–2878.

Moran, N.A., McCutcheon, J.P., and Nakabachi, A. 2008. Genomics and evolution of heritable bacterial symbionts. Annual Review of Genetics **42**(1): 165–190. doi:10.1146/annurev.genet.41.110306.130119.

Moran, N.A., McLaughlin, H.J., and Sorek, R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science **323**(5912): 379–382. doi:10.1126/science.1167140.

Morris, R.M., Rappé, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature **420**(6917): 806. doi:10.1038/nature01240.

Morris, J.J., Lenski, R.E., and Zinser, E.R. 2012. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. mBio **3**(2): e00036-12. doi:10.1128/mBio.00036-12.

Morris, M.R.J. 2014. Plasticity-Mediated Persistence in New and Changing Environments. International Journal of Evolutionary Biology **2014**: 1–18. doi:10.1155/2014/416497.

Muller, H.J. 1964. The relation of recombination to mutational advance. Supports Open Access Visit journal website Mutation Research: Fundamental and Molecular Mechanisms of Mutagenesis **1**(1): 2–9. doi:10.1016/0027-5107(64)90047-8.

Mullineaux, L.S., Mills, S.W., Sweetman, A.K., Beaudreau, A.H., Metaxas, A., and Hunt, H.L. 2005. Vertical, lateral and temporal structure in larval distributions at hydrothermal vents. Marine Ecology Progress Series **293**: 1–16. doi:10.3354/meps293001.

Mullineaux, L.S., Adams, D.K., Mills, S.W., and Beaulieu, S.E. 2010. Larvae from afar colonize deep-sea hydrothermal vents after a catastrophic eruption. Proceedings of the National Academy of Sciences **107**(17): 7829–7834. doi:10.1073/pnas.0913187107.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S.L. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetics **8**(7): e1002764. doi:10.1371/journal.pgen.1002764.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., and Scheffler, K. 2013. FUBAR: A fast, unconstrained bayesian approximation for inferring selection. Molecular Biology and Evolution **30**(5): 1196–1205. doi:10.1093/molbev/mst030.

Nautilus Minerals. 2018, February 27. Presse release: nautilus announces preliminary economic assessment for its solwara 1 project. Available from http://www.nautilusminerals.com/irm/PDF/1973_0/NautilusAnnouncesPreliminaryEconomicAssessmentforitsSolwara1Project.

Nees, H.A., Lutz, R.A., Shank, T.M., and Luther III, G.W. 2009. Pre- and post-eruption diffuse flow variability among tubeworm habitats at 9°50′ north on the East Pacific Rise. Deep Sea Research Part II: Topical Studies in Oceanography **56**(19–20): 1607–1615. doi:10.1016/j.dsr2.2009.05.007.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. 2017. Intragenic DNA methylation prevents spurious transcription initiation. Nature **543**(7643): 72–77. doi:10.1038/nature21373.

Newbold, L.K., Robinson, A., Rasnaca, I., Lahive, E., Soon, G.H., Lapied, E., Oughton, D., Gashchak, S., Beresford, N.A., and Spurgeon, D.J. 2019. Genetic, epigenetic and microbiome characterisation of an earthworm species (*Octolasion lacteum*) along a radiation exposure gradient at Chernobyl. Environmental Pollution **255**: 113238. doi:10.1016/j.envpol.2019.113238.

Newton, I.L.G., Woyke, T., Auchtung, T.A., Dilly, G.F., Dutton, R.J., Fisher, M.C., Fontanez, K.M., Lau, E., Stewart, F.J., Richardson, P.M., Barry, K.W., Saunders, E., Detter, J.C., Wu, D., Eisen, J.A., and Cavanaugh, C.M. 2007. The *Calyptogena magnifica* chemoautotrophic symbiont genome. Science **315**(5814): 998–1000. doi:10.1126/science.1138438.

Newton, I.L., Girguis, P.R., and Cavanaugh, C.M. 2008. Comparative genomics of vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. BMC Genomics **9**(1): 585. doi:10.1186/1471-2164-9-585.

Newton, I.L.G., and Bordenstein, S.R. 2010. Correlations between bacterial ecology and mobile DNA. Current Microbiology **62**(1): 198–208. doi:10.1007/s00284-010-9693-3.

Nikoh, N., McCutcheon, J.P., Kudo, T., Miyagishima, S., Moran, N.A., and Nakabachi, A. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLOS Genetics **6**(2): e1000827. doi:10.1371/journal.pgen.1000827.

Nowack, E.C.M., Vogel, H., Groth, M., Grossman, A.R., Melkonian, M., and Glöckner, G. 2011. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. Molecular Biology and Evolution **28**(1): 407–422. doi:10.1093/molbev/msq209.

Nussbaumer, A.D., Fisher, C.R., and Bright, M. 2006. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. Nature **441**(7091): 345–348. doi:10.1038/nature04793.

O'Dea, R.E., Noble, D.W.A., Johnson, S.L., Hesselson, D., and Nakagawa, S. 2016. The role of non-genetic inheritance in evolutionary rescue: epigenetic buffering, heritable bet hedging and epigenetic traps. Environmental Epigenetics **2**(1): dvv014.

Ogunlaja, A., Sharma, V., Ghai, M., and Lin, J. 2020. Molecular characterization and DNA methylation profile of *Libyodrilus violaceous* from oil polluted soil. Molecular Biology Research Communications **9**(2): 45–53. doi:10.22099/mbrc.2019.35242.1449.

Olins, H.C., Rogers, D.R., Preston, C., Ussler, W.I., Pargett, D., Jensen, S., Roman, B., Birch, J.M., Scholin, C.A., Haroon, M.F., and Girguis, P.R. 2017. Co-registered geochemistry and metatranscriptomics reveal unexpected distributions of microbial activity within a hydrothermal vent field. Frontiers in Microbiology **8**. doi:10.3389/fmicb.2017.01042.

Ozawa, G., Shimamura, S., Takaki, Y., Takishita, K., Ikuta, T., Barry, J.P., Maruyama, T., Fujikura, K., and Yoshida, T. 2017a. Ancient occasional host switching of maternally transmitted bacterial symbionts of chemosynthetic vesicomyid clams. Genome Biology and Evolution **9**(9): 2226–2236.

Ozawa, G., Shimamura, S., Takaki, Y., Yokobori, S.-I., Ohara, Y., Takishita, K., Maruyama, T., Fujikura, K., and Yoshida, T. 2017b. Updated mitochondrial phylogeny of pteriomorph and heterodont bivalvia, including deep-sea chemosymbiotic *Bathymodiolus* mussels, vesicomyid clams and the thyasirid clam *Conchocele* cf. *bisecta*. Marine Genomics **31**: 43–52. doi:10.1016/j.margen.2016.09.003.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. *In* Advances in Microbial Ecology. *Edited by* K.C. Marshall. Springer US, Boston, MA. pp. 1–55. doi:10.1007/978-1-4757-0611-6_1.

Papke, R.T., and Ward, D.M. 2004. The importance of physical isolation to microbial diversification. FEMS Microbiology Ecology **48**(3): 293–303. doi:10.1016/j.femsec.2004.03.013.

Patra, A.K., Cho, H.H., Kwon, Y.M., Kwon, K.K., Sato, T., Kato, C., Kang, S.G., and Kim, S.-J. 2016. Phylogenetic relationship between symbionts of tubeworm *Lamellibrachia* satsuma and the sediment microbial community in Kagoshima Bay. Ocean Science Journal **51**(3): 317–332. doi:10.1007/s12601-016-0028-6.

Patra, A.K., Perez, M., Jang, S.-J., and Won, Y.-J. 2022. A regulatory hydrogenase gene cluster observed in the thioautotrophic symbiont of *Bathymodiolus* mussel in the East Pacific Rise. Scientific Reports **12**(1): 22232. doi:10.1038/s41598-022-26669-y.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. 2017. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nature Methods **14**(4): 417–419. doi:10.1038/nmeth.4197.

Peek, A.S., Gustafson, R.G., Lutz, R.A., and Vrijenhoek, R.C. 1997. Evolutionary relationships of deep-sea hydrothermal vent and cold-water seep clams (Bivalvia: Vesicomyidae): results from the mitochondrial cytochrome oxidase subunit I. Marine Biology **130**(2): 151–161. doi:10.1007/s002270050234.

Peek, A.S., Feldman, R.A., Lutz, R.A., and Vrijenhoek, R.C. 1998. Cospeciation of chemoautotrophic bacteria and deep sea clams. Proceedings of the National Academy of Sciences **95**(17): 9962–9966. doi:10.1073/pnas.95.17.9962.

Pehrsson, E.C., Choudhary, M.N.K., Sundaram, V., and Wang, T. 2019. The epigenomic landscape of transposable elements across normal human development and anatomy. Nature Communications **10**(1): 5640. doi:10.1038/s41467-019-13555-x.

Pelisson, A., Mejlumian, L., Robert, V., Terzian, C., and Bucheton, A. 2002. *Drosophila* germline invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. Insect Biochemistry and Molecular Biology **32**(10): 1249–1256. doi:10.1016/S0965-1748(02)00088-7.

Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J.M., Silva, F.J., Moya, A., and Latorre, A. 2006. A small microbial genome: the end of a long symbiotic relationship? Science **314**(5797): 312–313.

Perez, M., and Juniper, S.K. 2016. Insights into symbiont population structure among three vestimentiferan tubeworm host species at eastern Pacific spreading centers. Applied and Environmental Microbiology **82**(17): 5197–5205. doi:10.1128/AEM.00953-16.

Perez, M., and Juniper, S.K. 2018. Is the trophosome of *Ridgeia piscesae* monoclonal? Symbiosis **74**(1): 55–65. doi:10.1007/s13199-017-0490-7.

Perez, M., Sun, J., Xu, Q., and Qian, P.-Y. 2021. Structure and connectivity of hydrothermal vent communities along the mid-ocean ridges in the west Indian Ocean: a review. Frontiers in Marine Science **8**: 1434. doi:10.3389/fmars.2021.744874.

Perez, M., Wang, H., Angers, B., and Qian, P.-Y. 2022. Complete mitochondrial genome of *Paralvinella palmiformis* (Polychaeta: Alvinellidae). Mitochondrial DNA Part B: Resources **7**(5): 786–788. doi:10.1080/23802359.2022.2071652.

Piednoël, M., Donnart, T., Esnault, C., Graça, P., Higuet, D., and Bonnivard, E. 2013. LTR-retrotransposons in *R. exoculata* and other crustaceans: the outstanding success of GalEa-Like Copia elements. PLoS ONE **8**(3): e57675. doi:10.1371/journal.pone.0057675.

Pinheiro, L.B., Coleman, V.A., Hindson, C.M., Herrmann, J., Hindson, B.J., Bhat, S., and Emslie, K.R. 2012. Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Analytical Chemistry **84**(2): 1003–1011. doi:10.1021/ac202578x.

Plague, G.R., Dunbar, H.E., Tran, P.L., and Moran, N.A. 2008. Extensive proliferation of transposable elements in heritable bacterial symbionts. Journal of Bacteriology **190**(2): 777–779. doi:10.1128/JB.01082-07.

Planques, A., Kerner, P., Ferry, L., Grunau, C., Gazave, E., and Vervoort, M. 2021. DNA methylation atlas and machinery in the developing and regenerating annelid *Platynereis dumerilii*. BMC Biology **19**(1): 148. doi:10.1186/s12915-021-01074-5.

Polzin, J., Arevalo, P., Nussbaumer, T., Polz, M.F., and Bright, M. 2019. Polyclonal symbiont populations in hydrothermal vent tubeworms and the environment. Proceedings of the Royal Society B: Biological Sciences **286**(1896): 20181281. doi:10.1098/rspb.2018.1281.

Prescott, L.M., Harley, J.P., and Klein, D.A. 2003. Microbiologie. De Boeck. Available from http://books.google.ca/books?id=_M4tHiD8VXgC.

Provataris, P., Meusemann, K., Niehuis, O., Grath, S., and Misof, B. 2018. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. Genome Biology and Evolution **10**(4): 1185–1197. doi:10.1093/gbe/evy066.

Puetz, L. 2014. Connectivity within a metapopulation of the foundation species, *Ridgeia piscesae* Jones (Annelida, Siboglinidae), from the Endeavour hydrothermal vents marine protected area on the Juan de Fuca Ridge. Thesis. Available from https://dspace.library.uvic.ca//handle/1828/5337 [accessed 22 April 2017].

Quinlan, A.R., and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6): 841–842. doi:10.1093/bioinformatics/btq033.

Ratner, H.K., Sampson, T.R., and Weiss, D.S. 2015. I can see CRISPR now, even when phage are gone: a view on alternative CRISPR-Cas functions from the prokaryotic envelope. Current Opinion in Infectious Diseases **28**(3): 267–274. doi:10.1097/QCO.0000000000000154.

Reed, D.C., Breier, J.A., Jiang, H., Anantharaman, K., Klausmeier, C.A., Toner, B.M., Hancock, C., Speer, K., Thurnherr, A.M., and Dick, G.J. 2015. Predicting the response of the deep-ocean microbiome to geochemical perturbations by hydrothermal vents. The ISME Journal **9**(8): 1857–1869. doi:10.1038/ismej.2015.4.

Reveillaud, J., Anderson, R., Reves-Sohn, S., Cavanaugh, C., and Huber, J.A. 2018. Metagenomic investigation of vestimentiferan tubeworm endosymbionts from Mid-Cayman Rise reveals new insights into metabolism and diversity. Microbiome **6**: 19. doi:10.1186/s40168-018-0411-x.

Riesgo, A., Andrade, S.C.S., Sharma, P.P., Novo, M., Pérez-Porro, A.R., Vahtera, V., González, V.L., Kawauchi, G.Y., and Giribet, G. 2012. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Frontiers in Zoology **9**(1): 33. doi:10.1186/1742-9994-9-33.

Rio, R.V.M., Attardo, G.M., and Weiss, B.L. 2016. Grandeur alliances: symbiont metabolic integration and obligate arthropod hematophagy. Trends in Parasitology **32**(9): 739–749. doi:10.1016/j.pt.2016.05.002.

Rispe, C., Delmotte, F., Ham, R.C.H.J. van, and Moya, A. 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Research **14**(1): 44–53. doi:10.1101/gr.1358104.

Ritchie, H., Jamieson, A.J., and Piertney, S.B. 2018. Heat-shock protein adaptation in abyssal and hadal amphipods. Deep-Sea Research Part II: Topical Studies in Oceanography **155**: 61–69. doi:10.1016/j.dsr2.2018.05.003.

Roberts, M.S., and Cohan, F.M. 1995. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. Evolution **49**(6): 1081–1094. doi:10.1111/j.1558-5646.1995.tb04435.x.

Roberts, S.B., and Gavery, M.R. 2012. Is there a relationship between DNA methylation and phenotypic plasticity in invertebrates? Frontiers in Physiology **2**. doi:10.3389/fphys.2011.00116.

Robidart, J.C., Bench, S.R., Feldman, R.A., Novoradovsky, A., Podell, S.B., Gaasterland, T., Allen, E.E., and Felbeck, H. 2008. Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. Environmental Microbiology **10**(3): 727–737.

Roeselers, G., Newton, I.L.G., Woyke, T., Auchtung, T.A., Dilly, G.F., Dutton, R.J., Fisher, M.C., Fontanez, K.M., Lau, E., Stewart, F.J., Richardson, P.M., Barry, K.W., Saunders, E., Detter, J.C., Wu, D., Eisen, J.A., and Cavanaugh, C.M. 2010. Complete genome sequence of *Candidatus* Ruthia magnifica. Standards in Genomic Sciences **3**(2): 163–173. doi:10.4056/sigs.1103048.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology **61**(3): 539–542. doi:10.1093/sysbio/sys029.

Ruan, L., Xu, H., Lin, W., Shi, H., Cui, Z., and Xu, X. 2017. A novel beta-galactose-specific lectin of the tubeworm, *Ridgeia piscesae*, from the hydrothermal vent. Acta Oceanologica Sinica **36**(6): 61–67. doi:10.1007/s13131-017-1052-9.

Rubin-Blum, M., Tsadok, R., Shemesh, E., Goodman-Tchernov, B.N., Austin, J.A., Coleman, D.F., Ben-Avraham, Z., Gruber, D.F., and Tchernov, D. 2014. Distribution of the *Lamellibrachia* spp. (Siboglinidae, Annelida) and their trophosome endosymbiont phylotypes in the Mediterranean Sea. Marine Biology **161**(6): 1229–1239. doi:10.1007/s00227-014-2413-y.

Ruppert, K.M., Kline, R.J., and Rahman, M.S. 2019. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation **17**: e00547. doi:10.1016/j.gecco.2019.e00547.

Russell, S.L., and Cavanaugh, C.M. 2017. Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes. Molecular Biology and Evolution **34**(11): 2747–2761. doi:10.1093/molbev/msx188.

Russell, S.L. 2019. Transmission mode is associated with environment type and taxa across bacteria-eukaryote symbioses: a systematic review and meta-analysis. FEMS Microbiology Letters **366**(3): fnz013.

Russell, S.L., Pepper-Tunick, E., Svedberg, J., Byrne, A., Castillo, J.R., Vollmers, C., Beinart, R.A., and Corbett-Detig, R. 2020. Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. PLoS Genetics **16**(8): e1008935. doi:10.1371/journal.pgen.1008935.

Russo, V.E., Martienssen, R.A., and Riggs, A.D. 1996. Epigenetic mechanisms of gene regulation. Cold Spring Harbor Laboratory Press.

Sachidanandham, R., and Yew-Hoong Gin, K. 2008. A dormancy state in nonspore-forming bacteria. Applied Microbiology and Biotechnology **81**(5): 927. doi:10.1007/s00253-008-1712-y.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. Molecular cloning: a laboratory manual. Cold spring harbor laboratory press, New York, US.

Sampson, T.R., Saroj, S.D., Llewellyn, A.C., Tzeng, Y.-L., and Weiss, D.S. 2013. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. Nature **497**(7448): 254–257. doi:10.1038/nature12048.

Sanchez-Perez, G.F., Hampl, V., Simpson, A.G.B., and Roger, A.J. 2008. A new divergent type of eukaryotic methionine adenosyltransferase is present in multiple distantly related secondary algal lineages. Journal of Eukaryotic Microbiology **55**(5): 374–381. doi:10.1111/j.1550-7408.2008.00349.x.

Santoyo, G., and Romero, D. 2005. Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiology Reviews **29**(2): 169–183. doi:10.1016/j.fmrre.2004.10.004.

Sarradin, P.-M., Sarrazin, J., and Lallier, F.H. 2017. Les impacts environnementaux de l'exploitation minière des fonds marins : un état des lieux des connaissances, The environmental impact of mining the seabed : State of knowledge. Annales des Mines - Responsabilité et environnement **85**: 30–34. Available from https://www.cairn.info/resume.php?ID_ARTICLE=RE1_085_0030 [accessed 20 October 2017].

Sarrazin, J., Robigou, V., Juniper, S., and Delaney, J. 1997. Biological and geological dynamics over four years on a high-temperature sulfide structure at the Juan de Fuca Ridge hydrothermal observatory. Marine Ecology Progress Series **153**: 5–24.

Sarrazin, J., and Juniper, S.K. 1999. Biological characteristics of a hydrothermal edifice mosaic community. Marine Ecology Progress Series **185**: 1-19.

Savitskaya, E., Lopatina, A., Medvedeva, S., Kapustin, M., Shmakov, S., Tikhonov, A., Artamonova, I.I., Logacheva, M., and Severinov, K. 2017. Dynamics of *Escherichia coli* type I-E CRISPR spacers over 42 000 years. Molecular Ecology **26**(7): 2019–2026. doi:10.1111/mec.13961.

Senapati, D., Bhattacharya, M., Kar, A., Chini, D.S., Das, B.K., and Patra, B.C. 2019. Environmental DNA (eDNA): a promising biological survey tool for aquatic species

detection. Proceedings of the Zoological Society **72**(3): 211–228. doi:10.1007/s12595-018-0268-9.

Shah, V., and Morris, R.M. 2015. Genome sequence of "*Candidatus* Thioglobus autotrophica" strain EF1, a chemoautotroph from the SUP05 clade of marine Gammaproteobacteria. Genome Announcements **3**(5): e01156-15. doi:10.1128/genomeA.01156-15.

Shariat, N., and Dudley, E.G. 2014. CRISPRs: molecular signatures used for pathogen subtyping. Applied and Environmental Microbiology **80**(2): 430–439. doi:10.1128/AEM.02790-13.

Shariat, N., Timme, R.E., Pettengill, J.B., Barrangou, R., and Dudley, E.G. 2015. Characterization and evolution of *Salmonella* CRISPR-Cas systems. Microbiology **161**(2): 374–386. doi:10.1099/mic.0.000005.

Sharp, P.M., and Li, W.-H. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Research **14**(19): 7737–7749. doi:10.1093/nar/14.19.7737.

Shelton, A.N., Seth, E.C., Mok, K.C., Han, A.W., Jackson, S.N., Haft, D.R., and Taga, M.E. 2019. Uneven distribution of cobamide biosynthesis and dependence in bacteria predicted by comparative genomics. The ISME Journal **13**(3): 789–804. doi:10.1038/s41396-018-0304-9.

Shen, J., Lv, L., Wang, X., Xiu, Z., and Chen, G. 2017. Comparative analysis of CRISPR-Cas systems in *Klebsiella* genomes. Journal of Basic Microbiology, **57**(4), 325-336. doi:10.1002/jobm.201600589.

Shimamura, S., Kaneko, T., Ozawa, G., Matsumoto, M.N., Koshiishi, T., Takaki, Y., Kato, C., Takai, K., Yoshida, T., Fujikura, K., Barry, J.P., and Maruyama, T. 2017. Loss of genes related to Nucleotide Excision Repair (NER) and implications for reductive genome evolution in symbionts of deep-sea vesicomyid clams. PLoS ONE **12**(2): e0171274. doi:10.1371/journal.pone.0171274.

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

Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. Nature Methods **14**(4): 407–410. doi:10.1038/nmeth.4184.

Skennerton, C.T., Imelfort, M., and Tyson, G.W. 2013. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Research **41**(10): e105–e105. doi:10.1093/nar/gkt183.

Smallwood, S.A., and Kelsey, G. 2012. De novo DNA methylation: a germ cell perspective. Trends in Genetics **28**(1): 33–42. Elsevier. doi:10.1016/j.tig.2011.09.004.

Smith, C.R., and Baco, A.R. 2003. Ecology of whale falls at the deep-sea floor. Oceanography and Marine Biology **41**: 311–354.

Smith, C.R., Glover, A.G., Treude, T., Higgs, N.D., and Amon, D.J. 2015. Whale-fall ecosystems: recent insights into ecology, paleoecology, and evolution. Annual Review of Marine Science **7**: 571–596. doi:10.1146/annurev-marine-010213-135144.

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Molecular Biology and Evolution **32**(5): 1342–1353. doi:10.1093/molbev/msv022.

Sorek, R., Kunin, V., and Hugenholtz, P. 2008. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews Microbiology **6**(3): 181–186. doi:10.1038/nrmicro1793.

Southward, E.C., Tunnicliffe, V., and Black, M. 1995. Revision of the species of *Ridgeia* from northeast Pacific hydrothermal vents, with a redescription of *Ridgeia piscesae* Jones (Pogonophora: Obturata= Vestimentifera). Canadian Journal of Zoology **73**(2): 282–295.

Spiess, F.N., Macdonald, K.C., Atwater, T., Ballard, R., Carranza, A., Cordoba, D., Cox, C., Garcia, V.M.D., Francheteau, J., Guerrero, J., Hawkins, J., Haymon, R., Hessler, R., Juteau, T., Kastner, M., Larson, R., Luyendyk, B., Macdougall, J.D., Miller, S., Normark, W., Orcutt, J., and Rangin, C. 1980. East Pacific Rise: hot springs and geophysical experiments. Science **207**(4438): 1421–1433. doi:10.1126/science.207.4438.1421.

Stackebrandt, E., and Goebel, B.M. 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology **44**(4): 846–849. doi:10.1099/00207713-44-4-846.

Staley, J.T., and Gosink, J.J. 1999. Poles apart: biodiversity and biogeography of sea ice bacteria. Annual Review of Microbiology **53**(1): 189–215. doi:10.1146/annurev.micro.53.1.189.

Stevenson, L.H. 1977. A case for bacterial dormancy in aquatic systems. Microbial Ecology **4**(2): 127–133. doi:10.1007/BF02014283.

Stewart, F.J., Young, C.R., and Cavanaugh, C.M. 2008. Lateral symbiont acquisition in a maternally transmitted chemosynthetic clam endosymbiosis. Molecular Biology and Evolution **25**(4): 673–687. doi:10.1093/molbev/msn010.

Stewart, F.J., Young, C.R., and Cavanaugh, C.M. 2009. Evidence for homologous recombination in intracellular chemosynthetic clam symbionts. Molecular Biology and Evolution **26**(6): 1391–1404. doi:10.1093/molbev/msp049.

Stiller, J., Tilic, E., Rousset, V., Pleijel, F., and Rouse, G.W. 2020. Spaghetti to a tree: a robust phylogeny for Terebelliformia (Annelida) based on transcriptomes, molecular and morphological data. Biology **9**(4): 73. doi:10.3390/biology9040073.

Stipanuk, M.H. 2020. Metabolism of sulfur-containing amino acids: how the body copes with excess methionine, cysteine, and sulfide. Journal of Nutrition **150**: 2494S-2505S. doi:10.1093/jn/nxaa094.

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA **12**(1): 2. doi:10.1186/s13100-020-00230-y.

Strader, M.E., Kozal, L.C., Leach, T.S., Wong, J.M., Chamorro, J.D., Housh, M.J., and Hofmann, G.E. 2020. Examining the role of DNA methylation in transcriptomic plasticity of early stage sea urchins: developmental and maternal effects in a kelp forest herbivore. Frontiers in Marine Science **7**: 205. doi:10.3389/fmars.2020.00205.

Sudakaran, S., Kost, C., and Kaltenpoth, M. 2017. Symbiont acquisition and replacement as a source of ecological innovation. Trends in Microbiology **25**(5): 375–390. doi:10.1016/j.tim.2017.02.014.

Sul, W.J., Oliver, T.A., Ducklow, H.W., Amaral-Zettler, L.A., and Sogin, M.L. 2013. Marine bacteria exhibit a bipolar distribution. Proceedings of the National Academy of Sciences **110**(6): 2342–2347. doi:10.1073/pnas.1212424110.

Sun, Y., Sun, J., Yang, Y., Lan, Y., Ip, J.C.-H., Wong, W.C., Kwan, Y.H., Zhang, Y., Han, Z., Qiu, J.-W., and Qian, P.-Y. 2021. Genomic signatures supporting the symbiosis and formation of chitinous tube in the deep-sea tubeworm *Paraescarpia echinospica*. Molecular Biology and Evolution **38**(10): 4116–4134. doi:10.1093/molbev/msab203.

Suzuki, M.M., and Bird, A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics **9**(6): 465–476. doi:10.1038/nrg2341.

Székely, D., Cammen, K.M., and Olsen, M.T. 2022. Needles in an ocean haystack: using environmental DNA to study marine mammals in the North Atlantic. NAMMCO Scientific Publications **12**. doi:10.7557/3.6482.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A.K., Eriksson, A.-S., Wernegreen, J.J., Sandström, J.P., Moran, N.A., and Andersson, S.G.E. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296**(5577): 2376–2379. doi:10.1126/science.1071278.

Tesler, G. 2002. GRIMM: Genome rearrangements web server. Bioinformatics **18**(3): 492–493. doi:10.1093/bioinformatics/18.3.492.

Thomas-Bulle, C., Piednoël, M., Donnart, T., Filée, J., Jollivet, D., and Bonnivard, É. 2018. Mollusc genomes reveal variability in patterns of LTR-retrotransposons dynamics. BMC Genomics **19**(1): 821. doi:10.1186/s12864-018-5200-1.

Thornburg, C.C., Zabriskie, T.M., and McPhail, K.L. 2010. Deep-sea hydrothermal vents: potential hot spots for natural products discovery? Journal of Natural Products **73**(3): 489–499. doi:10.1021/np900662k.

Tilic, E., Sayyari, E., Stiller, J., Mirarab, S., and Rouse, G.W. 2020. More is needed—Thousands of loci are required to elucidate the relationships of the 'flowers of the sea' (Sabellida, Annelida). Molecular Phylogenetics and Evolution 151: 106892. doi:10.1016/j.ympev.2020.106892.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. 2017. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Research **45**(W1): W6–W11. Oxford Academic. doi:10.1093/nar/gkx391.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics **5**(2): 123. doi:10.1038/nrg1271.

Todgham, A.E., Crombie, T.A., and Hofmann, G.E. 2017. The effect of temperature adaptation on the ubiquitin–proteasome pathway in notothenioid fishes. Journal of Experimental Biology **220**(3): 369–378. doi:10.1242/jeb.145946.

Torrents, E. 2014. Ribonucleotide reductases: Essential enzymes for bacterial life. Frontiers in Cellular and Infection Microbiology **4**: 52. doi:10.3389/fcimb.2014.00052.

Touchon, M., Charpentier, S., Clermont, O., Rocha, E.P.C., Denamur, E., and Branger, C. 2011. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. Journal of Bacteriology **193**(10): 2460–2467. doi:10.1128/JB.01307-10.

Tsurumi, M., and Tunnicliffe, V. 2003. Tubeworm-associated communities at hydrothermal vents on the Juan de Fuca Ridge, Northeast Pacific. Deep Sea Research Part I: Oceanographic Research Papers **50**(5): 611–629. doi:10.1016/S0967-0637(03)00039-6.

Tunnicliffe, V., Garrett, J.F., and Johnson, H.P. 1990. Physical and biological factors affecting the behaviour and mortality of hydrothermal vent tubeworms (vestimentiferans). Deep Sea Research Part A. Oceanographic Research Papers **37**(1): 103–125. doi:10.1016/0198-0149(90)90031-P.

Tunnicliffe, V., Desbruyères, D., Jollivet, D., and Laubier, L. 1993. Systematic and ecological characteristics of *Paralvinella sulfincola* Desbruyères and Laubier, a new polychaete (family Alvinellidae) from northeast Pacific hydrothermal vents. Canadian Journal of Zoology **71**(2): 286–297. doi:10.1139/z93-041.

Tunnicliffe, V., Germain, C.S., and Hilário, A. 2014. Phenotypic variation and fitness in a metapopulation of tubeworms (*Ridgeia piscesae*, Jones) at hydrothermal vents. PLoS ONE **9**(10): e110578. doi:10.1371/journal.pone.0110578.

Tyson, G.W., and Banfield, J.F. 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environmental Microbiology **10**(1): 200–207. doi:10.1111/j.1462-2920.2007.01444.x.

UNCLOS - Part XI, Section 2. 1982. Available from https://www.un.org/depts/los/convention_agreements/texts/unclos/part11-2.htm [accessed 4 July 2019].

Urcuyo, I.A., Massoth, G.J., Julian, D., and Fisher, C.R. 2003. Habitat, growth and physiological ecology of a basaltic community of *Ridgeia piscesae* from the Juan de Fuca Ridge. Deep Sea Research Part I: Oceanographic Research Papers **50**(6): 763–780.

Urcuyo, I.A., Bergquist, D.C., MacDonald, I.R., VanHorn, M., and Fisher, C.R. 2007. Growth and longevity of the tubeworm *Ridgeia piscesae* in the variable diffuse flow habitats of the Juan de Fuca Ridge. Marine Ecology Progress Series **344**: 143–157.

Uthicke, S., Lamare, M., and Doyle, J.R. 2018. eDNA detection of corallivorous seastar (*Acanthaster* cf. *solaris*) outbreaks on the Great Barrier Reef using digital droplet PCR. Coral Reefs **37**(4): 1229–1239. doi:10.1007/s00338-018-1734-6.

Van Dover, C. 2002. Trophic relationships among invertebrates at the Kairei hydrothermal vent field (Central Indian Ridge). Marine Biology **141**(4): 761–772. doi:10.1007/s00227-002-0865-y.

Van Dover, C.L., Arnaud-Haond, S., Gianni, M., Helmreich, S., Huber, J.A., Jaeckel, A.L., Metaxas, A., Pendleton, L.H., Petersen, S., Ramirez-Llodra, E., Steinberg, P.E., Tunnicliffe, V., and Yamamoto, H. 2018. Scientific rationale and international obligations for protection of active hydrothermal vent ecosystems from deep-sea mining. Marine Policy **90**: 20–28. doi:10.1016/j.marpol.2018.01.020.

Veesenmeyer, J.L., Andersen, A.W., Lu, X., Hussa, E.A., Murfin, K.E., Chaston, J.M., Dillman, A.R., Wassarman, K.M., Sternberg, P.W., and Goodrich-Blair, H. 2014. NilD CRISPR RNA contributes to *Xenorhabdus nematophila* colonization of symbiotic host nematodes. Molecular Microbiology **93**(5): 1026–1042. doi:10.1111/mmi.12715.

Veller, C., Hayward, L.K., Hilbe, C., and Nowak, M.A. 2017. The Red Queen and King in finite populations. Proceedings of the National Academy of Sciences **114**(27): E5396–E5405. doi:10.1073/pnas.1702020114.

Vrijenhoek, R.C. 1997. Gene flow and genetic diversity in naturally fragmented metapopulations of deep-sea hydrothermal vent animals. Journal of Heredity **88**(4): 285–293. doi:10.1093/oxfordjournals.jhered.a023106.

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

Waddington, C.H. 1942. The epigenotype. International Journal of Epidemiology **41**(1): 10–13. doi:10.1093/ije/dyr184.

Waddington, C.H. 1957. The strategy of the genes. Routledge.

Wang, X., Li, Q., Lian, J., Li, L., Jin, L., Cai, H., Xu, F., Qi, H., Zhang, L., Wu, F., Meng, J., Que, H., Fang, X., Guo, X., and Zhang, G. 2014. Genome-wide and single-base resolution DNA methylomes of the Pacific oyster Crassostrea gigas provide insight into the evolution of invertebrate CpG methylation. BMC Genomics **15**(1): 1119. doi:10.1186/1471-2164-15-1119.

Wang, X., and Bhandari, R.K. 2019. DNA methylation dynamics during epigenetic reprogramming of medaka embryo. Epigenetics **14**(6): 611–622. doi:10.1080/15592294.2019.1605816.

Wang, H., Zhang, L., He, J., and Zhou, T. 2022. The development of natural gas hydrate exploitation technology from perspective of patents. Frontiers in Energy Research **10**: 187.

Wang, M., Ruan, L., Liu, M., Liu, Z., He, J., Zhang, L., Wang, Y., Shi, H., Chen, M., Yang, F., Zeng, R., He, J., Guo, C., and Chen, J. 2023. The genome of a vestimentiferan tubeworm (*Ridgeia piscesae*) provides insights into its adaptation to a deep-sea environment. BMC Genomics **24**(1): 72. doi:10.1186/s12864-023-09166-y.

Ward, B.B., and O'Mullan, G.D. 2002. worldwide distribution of N*itrosococcus oceani*, a marine ammonia-oxidizing γ-proteobacterium, detected by PCR and sequencing of 16S rRNA and amoA genes. Applied and Environmental Microbiology **68**(8): 4153–4157. doi:10.1128/AEM.68.8.4153-4157.2002.

Ward, B., Cael, B., Collins, S., and Young, C. 2020. Selective constraints on global plankton dispersal. Proceedings of the National Academy of Sciences **118**(10), e2007388118.

Watling, L., Guinotte, J., Clark, M.R., and Smith, C.R. 2013. A proposed biogeography of the deep ocean floor. Progress in Oceanography **111**: 91–112. doi:10.1016/j.pocean.2012.11.003.

Weber, A.A.-T., Hugall, A.F., and O'Hara, T.D. 2020. Convergent evolution and structural adaptation to the deep Ocean in the protein-folding chaperonin CCTα. Genome Biology and Evolution **12**(11): 1929–1942. doi:10.1093/gbe/evaa167.

Wernegreen, J.J., and Moran, N.A. 1999. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. Molecular Biology and Evolution **16**(1): 83–97.

Wernegreen, J.J., Richardson, A.O., and Moran, N.A. 2001. Parallel acceleration of evolutionary rates in symbiont genes underlying host nutrition. Molecular phylogenetics and evolution **19**(3): 479–485.

Wernegreen, J.J. 2011. Reduced selective constraint in endosymbionts: elevation in radical amino acid replacements occurs genome-wide. PLoS ONE **6**(12): e28905. doi:10.1371/journal.pone.0028905.

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., and Scheffler, K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Molecular Biology and Evolution **32**(3): 820–832. doi:10.1093/molbev/msu400.

Westra, E.R., Dowling, A.J., Broniewski, J.M., and Houte, S. van. 2016. Evolution and ecology of CRISPR. Annual Review of Ecology, Evolution, and Systematics **47**(1): 307–331. doi:10.1146/annurev-ecolsys-121415-032428.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. 2003.G Geographic barriers isolate endemic populations of hyperthermophilic Archaea. Science **301**(5635): 976–978. doi:10.1126/science.1086909.

Wirsen, C.O., Brinkhoff, T., Kuever, J., Muyzer, G., Molyneaux, S., and Jannasch, H.W. 1998. Comparison of a new T*hiomicrospira* strain from the mid-atlantic ridge with known hydrothermal vent isolates. Applied and Environmental Microbiology **64**(10): 4057–4059.

Wright, S. 1943. Isolation by distance. Genetics **28**(2): 114–138.

Xu, J.-S., Xia, Q.-Y., Li, J., Pan, G.-Q., and Zhou, Z.-Y. 2005. Survey of long terminal repeat retrotransposons of domesticated silkworm (*Bombyx mori*). Insect Biochemistry and Molecular Biology **35**(8): 921–929. doi:10.1016/j.ibmb.2005.03.014.

Yahagi, T., Watanabe, H., Shigeaki, K., and Yasunori, K. 2017. Do larvae from deep-sea hydrothermal vents disperse in surface waters? Ecology **98**(6): 1524–1534. doi:10.1002/ecy.1800.

Yan, G., Lan, Y., Sun, J., Xu, T., Wei, T., and Qian, P.-Y. 2022. Comparative transcriptomic analysis of in situ and onboard fixed deep-sea limpets reveals sample preparation-related differences. iScience **25**(4): 104092. doi:10.1016/j.isci.2022.104092.

Yang, L., Liu, Y., Zhang, H., Xiao, B., Guo, X., Wei, R., Xu, L., Sun, L., Yu, B., Leng, S., and Li, Y. 2019. The status of exploitation techniques of natural gas hydrate. Chinese Journal of Chemical Engineering **27**(9): 2133–2147. doi:10.1016/j.cjche.2019.02.028.

Yang, M., Gong, L., Sui, J., and Li, X. 2019. The complete mitochondrial genome of *Calyptogena* marissinica (Heterodonta: Veneroida: Vesicomyidae): insight into the deep-sea adaptive evolution of vesicomyids. PLoS ONE **14**(9): e0217952. doi:10.1371/journal.pone.0217952.

Yates, M.C., Fraser, D.J., and Derry, A.M. 2019. Meta-analysis supports further refinement of eDNA for monitoring aquatic species-specific abundance in nature. Environmental DNA **1**(1): 5–13. doi:10.1002/edn3.7.

Yin, S., Jensen, M.A., Bai, J., DebRoy, C., Barrangou, R., and Dudley, E.G. 2013. The evolutionary divergence of Shiga toxin-producing *Escherichia coli* is reflected in clustered regularly interspaced short palindromic repeat (CRISPR) spacer composition. Applied and Environmental Microbiology **79**(18): 5710–5720. doi:10.1128/AEM.00950-13.

Ying, H., Hayward, D.C., Klimovich, A., Bosch, T.C.G., Baldassarre, L., Neeman, T., Forêt, S., Huttley, G., Reitzel, A.M., Fraune, S., Ball, E.E., and Miller, D.J. 2022. The role of DNA methylation in genome defense in Cnidaria and other invertebrates. Molecular Biology and Evolution **39**(2): msac018. doi:10.1093/molbev/msac018.

Young, C.R., Fujio, S., and Vrijenhoek, R.C. 2008. Directional dispersal between mid-ocean ridges: deep-ocean circulation and gene flow in *Ridgeia piscesae*. Molecular Ecology **17**(7): 1718–1731.

Yuen, Z.W.-S., Srivastava, A., Daniel, R., McNevin, D., Jack, C., and Eyras, E. 2021. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nature Communications **12**(1): 3438. doi:10.1038/s41467-021-23778-6.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., and Ecker, J.R. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell **126**(6): 1189–1201. doi:10.1016/j.cell.2006.08.003.

Zhang, Z., Li, J., Cui, P., Ding, F., Li, A., Townsend, J.P., and Yu, J. 2012. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinformatics **13**(1): 43. doi:10.1186/1471-2105-13-43.

Zhao, Y., Xu, T., Law, Y.S., Feng, D., Li, N., Xin, R., Wang, H., Ji, F., Zhou, H., and Qiu, J.-W. 2020. Ecological characterization of cold-seep epifauna in the south China sea. Deep-Sea Research Part I: Oceanographic Research Papers **163**: 103361. doi:10.1016/j.dsr.2020.103361.

Zhou, Y., Zhang, D., Zhang, R., Liu, Z., Tao, C., Lu, B., Sun, D., Xu, P., Lin, R., Wang, J., and Wang, C. 2018. Characterization of vent fauna at three hydrothermal vent fields on the Southwest Indian Ridge: implications for biogeography and interannual dynamics on ultraslow-spreading ridges. Deep-Sea Research Part I: Oceanographic Research Papers **137**: 1–12. doi:10.1016/j.dsr.2018.05.001.

Zimmermann, J., Lott, C., Weber, M., Ramette, A., Bright, M., Dubilier, N., and Petersen, J.M. 2014. Dual symbiosis with co-occurring sulfur-oxidizing symbionts in vestimentiferan tubeworms from a Mediterranean hydrothermal vent. Environmental Microbiology **16**(12): 3638–3656. doi:10.1111/1462-2920.12427.

Zwart, G., Crump, B.C., Agterveld, M.P.K., Hagen, F., and Han, S.-K. 2002. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. Aquatic Microbial Ecology **28**(2): 141–155. doi:10.3354/ame028141.

Annex I. Chapter 2 supplementary materials

Supplementary Methods

Sample collection, DNA extraction, and metagenomic sequencing

Host (and symbiont) taxa examined in this study were chosen from the deepest diverging lineages within the Vesicomyidae that are distributed globally in the northern hemisphere and are representative of the known host diversity [1]. Mitochondrial and symbiont genome assemblies of *Abyssogena mariana* [2], *A. phaseoliformis* (Japan) [2], *Archivesica marissinica* [3, 4], *Calyptogena fausta* [5], *C. magnifica* [6, 7] and *Phreagena okutanii* [2, 8] were obtained from previous studies. New clam specimens were collected from eleven hydrocarbon seep and hydrothermal vent sites in the Pacific and Atlantic Ocean during research expeditions between 1994 and 2004 (Figure S2.1, Table S2.1). Upon recovery of the submersibles, samples were dissected and frozen at –80°C. DNA was extracted onshore from symbiont-bearing gill tissues with the DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) following manufacturer's instructions. Barcoded 2x300 bp metagenomic libraries for mixed host and symbiont DNA samples were prepared with the KAPA Hyperplus Library Preparation kit (KAPA Biosystems, Wilmington, MA, US) and sequenced on an Illumina MiSeq system at the National Oceanography Centre (Southampton, UK). Clam host species were identified via mitochondrial cytochrome-c-oxidase I (*COI*) sequencing using vesicomyid-specific primers [9]. Bacterial relatives with free-living phase of the SUP05 clade (*Bathymodiolus thermophilus* symbiont [Won *et al.* unpubl.], *Ca.* Thioglobus autotrophicus [10]) were selected as outgroups (Table S2.1).

Mitochondrial and symbiont genome reconstruction and annotation

After initial quality checks with FastQC v0.11.5 [11], reads were adapter-clipped with Trimmomatic v0.36 [12] and assembled *de novo* with Velvet v1.2.10 [13], SPAdes v3.13.1 [14] or Geneious v10.1.3 [15] using manual optimizations of k-mer size distribution and read depth. Mitochondrial genomes were assembled *de novo* with MitoBim v1.9 [16] using as seed a set of initial contigs constructed with the read mapping and assembly functions in Geneious. Scaffolding of the symbiont genomes was done with SSPACE v2.0 [17] and final circularization was performed

by re-mapping, extracting and reassembling reads that mapped to the extremities of contigs using BOWTIE2 v2.4.2 [18], SAMTOOLS v1.12 [19] and SPADES, respectively. Mitochondrial genome annotations were produced by GESEQ [20] using ARWEN v1.2.3 [21] for tRNA prediction, and manually curated with the aid of previously annotated mitochondrial genomes [2, 6] in GENEIOUS. Symbiont genome assemblies were annotated with RAST v2.0 [22] (Table S2.2). Genes were classified as pseudogenes with PSEUDOFINDER v1.0 (https://github.com/filip-husnik/pseudofinder/) if protein length was <80% of the average length of matches in the RefSeq database. We used this relatively conservative threshold to ensure exclusion of all potential pseudogenes from downstream selection and phylogenetic analyses. Pseudogenes annotated as fragmented were manually curated to distinguish actual fragmentation from functional gene copy number variations. Assembly quality and statistics were assessed with QUAST v5.0.0 [23], while taxonomic classification was performed with GTDB-TK v1.4.0 [24]. Average nucleotide identities and alignment fractions between genomes were calculated with FASTANI v1.32 [25].

Identification of orthologous groups and gene duplication events

Sequence homology between symbiont genomes was inferred via two independent, complementary methods. First, single-copy core orthologs were identified based on homology and position [26] with the function "getOrthologList" from MAUVE v2.4.0 [27], using a minimum identity of 35% and a minimum coverage of 51% (Table S2.3). Second, broader orthologous gene groups and gene duplications were determined with ORTHOFINDER v2.5.2 [28] using MAFFT v7.310 [29] for multiple sequence alignment, FASTTREE V2.1.11 [30] for gene tree inference and BLAST V2.9.0+ [31] for sequence search. A rooted species tree based on the core positional orthologs was used as prior information. Paralogous groups that originated before the divergence of the first extant clades were split into separate orthogroups (Table S2.4). An overview figure comparing genomic characteristics and relatedness of symbionts and bacterial relatives was produced with the COMPLEXHEATMAP package in R v4.0.3 [32] using Manhattan distances for clustering.

Genetic variation and phylogenomic analyses

Symbiont and mitochondrial intra-host genetic heterogeneities were estimated from the abundance of single nucleotide polymorphisms (SNPs). For each species, raw reads were mapped to the symbiont and mitochondrial genomes with BOWTIE2 (--very-sensitive-local) and SNPs were called

with VARSCAN v2.4.2 [33] using the following filters: --min-coverage 2 --min-reads2 1 --min-avg-qual 28 --min-var-freq 0.01. To remove false positives among symbiont SNPs, the "vcf" files were filtered with the accessory *fpfilter* script [33] using default parameters except for two: --max-var-mmqs 150 --max-mmqs-diff 100. To avoid the detection of putative false positive mitochondrial variants, we excluded the control region between ND6 and tRNA-Ala, which could not be fully resolved and contained repeats resulting in dubious mappings of the metagenomic reads. Because of low coverage for the mitochondrial genomes, the *fpfilter* script was not used. PROGRESSIVEMAUVE v2.4.0 [34] and GRIMM v2.01 [35] were used to identify large-scale structural differences among mitochondrial and symbiont genomes. For the host mitochondria, we concatenated alignments of 13 conserved protein-coding genes. For the symbionts and bacterial relatives, we extracted, realigned and concatenated 716 locally collinear blocks (LCBs) longer than 100 bp that were found in all bacterial genomes. Phylogenetic trees were produced from these core genes in MRBAYES v3.2.7a [36] using a GTR nucleotide substitution rate with a Gamma + I distribution across sites. The prior for the branch lengths was set to Unconstrained:Exp(50.0). Ten independent MCMC chains were each run for 2,000,000 generations after an initial 100,000 generations burn-in period. Trees were sampled every 10,000 generations to avoid autocorrelation. Parameter optimization for the MCMCs was performed by assessing convergence and mixing of both the continuous parameters of the model and the tree topologies using the R package RWTY v1.0.2 [37]. For the symbionts, additional neighbor-joining trees based on Jaccard distance were built from gene presence/absence patterns.

Bayesian concordance analyses

We used BUCKY v1.2 [38] to estimate the proportion of syntenic blocks – defined here as PROGRESSIVEMAUVE's LCBs – supporting each symbiont topology. Putative recombination breakpoints within the core LCBs ≥ 100 bp were identified with GARD v0.2 [39] based on AICc ratio tests and a 5% false positive discovery rate threshold. Input posterior distributions of LCB topologies were each obtained from 2,000 trees generated in MRBAYES with the same parameters as for the genome-wide tree construction. Two independent MCMC runs were carried out under the prior assumption that almost all genes shared the same topology (alpha = 0.001). MCMC runs

were updated 1,000,000 times after an initial 10% burn-in period. One cold and three heated chains (swapping frequency = 10) were used to improve mixing and convergence of all of the MCMC runs.

Host and symbiont evolutionary rates

We compared host and symbiont evolutionary rates by estimating the genome-wide divergence at synonymous sites between each host-symbiont pair. Nucleic and amino acid sequences of the 13 conserved mitochondrial and 555 non-recombining bacterial core protein-coding genes were extracted with BIOPYTHON v1.76 [40]. Amino acid sequences were then aligned with MUSCLE [41] and reverse translated into codon alignments using the "build" function from the BIOPYTHON "codonalign" package. The mitochondrial and bacterial codon-based alignments were each concatenated into two genome-wide alignments with lengths of 12,558 bp and 484,320 bp, respectively. We assessed substitution saturation by plotting transitions and transversions against adjusted F84 genetic distance. Pairwise synonymous substitution rates were computed using the Maximum-Likelihood method [42] implemented in the BIOPYTHON codonalign package. The source code was slightly modified to accommodate for ambiguous bases in the mitochondrial genomes.

Genome-wide screen for positive selection and changes in selective pressures

Episodic diversifying selection on individual lineages was identified on the core non-recombining protein-coding genes using the adaptive Branch-site Random Effects Likelihood method (aBSREL v2.2) [43] with corrections for multiple testing based on the Holm-Bonferroni procedure (alpha = 0.05). Changes in the strength of selection were inferred via two independent methods. First, we used the Codon Deviation Coefficient [44] to quantify codon usage bias on all core protein-coding genes because this index does not require *a priori* knowledge of gene expression and is not biased by GC content. Second, we used RELAX [45] to detect changes in the strength of selection between group pairs. We compared Clade I, Clade II, and both clades together to the outgroup. To support the inference of drift-driven RGE in the vesicomyid symbionts, we repeated the codon usage bias and RELAX analyses with additional 'outgroup' metagenome-assembled genomes of free-living SUP05 bacteria and closely related horizontally transmitted symbionts associated with deep-sea mussels (*Bathymodiolus* sp.) and sponges (*Suberites* sp.) (Table S2.5). To test whether genes under

episodic positive, relaxed or intensified (diversifying and purifying) selection represented a random subsample according to SEED categories [46], we estimated the probability of each distribution using the dmvhyper function from the EXTRADISTR v1.8.11 R package (https://cran.r-project.org/web/packages/extraDistr/index.html) and compared it to that of 100 distributions obtained from randomly sampling the non-recombining core gene dataset. Fisher's exact tests [47] were applied to find SEED categories that were over-represented in the genes under relaxed or intensified selection.

Tests for site-specific adaptive evolution in metabolic candidate genes

We assessed signatures of site-specific positive selection in 17 candidate genes that showed marked differences in presence/absence or duplication patterns between the two symbiont clades: vitamin B12 transporter component (*btuM*), cob(I)alamin adenosyltransferase (*btuR*), cysteine dioxygenase type I (*cdo*), putative cysteine sulfinic acid decarboxylase (*csad*), lactoylglutathione lyase (*gloA*), hydrogenase/urease accessory protein (*hupE*), isocitrate lyase (*icl*), methionine synthase and transcriptional activator (*metE*, *metH*, *metR*), dissimilatory/assimilatory nitrate reductase (*narGHIJ*, *nasA*), ribonucleotide reductase regulator (*nrdR*), and sulfide:quinone oxidoreductase type I (*sqrI*). Tests for pervasive and episodic diversifying selection were performed using Bayesian approximation and mixed-effects maximum likelihood approaches implemented in FUBAR v2.2 [48] and MEME v2.1.2 [49], respectively. Sequence alignments were partitioned according to recombination breakpoints identified with GARD. FUBAR analyses included 5 MCMC chains, with chain lengths of 2,000,000, a burn-in of 1,000,000 and a sample size of 100, while MEME analyses were run with default settings testing 1) the complete symbiont phylogeny and 2) only symbiont branches. Because site-level tests for positive selection are relatively conservative, we chose recommended p-value thresholds of 0.1 for MEME and posterior probability thresholds of 0.9 for FUBAR to assess statistical significance [50]. To assess the implications of site-specific selection in the investigated genes we predicted structural and functional features of the encoded proteins with PREDICTPROTEIN [51].

Supplementary Results

RGE signatures including additional outgroup genomes

The 'outgroup' genomes possess codon usage biases which are more variable but overall higher than those of the clam symbionts (Wilcoxon test p-value < 0.0001; Figure S2.10A). Codon usage bias is ordered as follows: Horizontally transmitted symbionts > free-living SUP05 ~ Clade II > Clade I (Figure S2.10B). Some genomes in the SUP05 clade contain strong reduction of codon usage bias compared to other members of the group (Figure S2.10A). Given their free-living lifestyle, lower codon usage bias in these bacteria likely results from selective processes rather than bottleneck-driven genetic drift [52–54]. Relaxation of selection tests further support the hypothesis of adaptive rather than neutral drivers of RGE in free-living SUP05. Indeed, even though they possess similar levels of codon usage bias, selection appears intensified in these bacteria compared to the symbionts associated with *Bathymodiolus* sp. and *Suberites* sp. (Figure S2.10C). Lower selection intensity in the *Bathymodiolus* sp. and *Suberites* sp. symbionts also signals an effect of genetic drift and highlights the consequence of host-symbiont integration (even if partial) on their evolution. Complete host-symbiont integration through vertical transmission appears to exacerbate genetic drift in the Clade I and II symbionts.

Energy metabolism

The genomes of all symbiont lineages contained genes for the oxidation of reduced sulfur compounds that serve as energy sources for chemoautotrophic growth [55]. All genomes encoded the sulfur oxidation (SOX) multienzyme pathway (without *soxCD*), the reversible dissimilatory sulfite reductase (rDSR) pathway as well as the adenosine 5'-phosphosulfate (APS) reductase pathway, indicating that these symbiont lineages are able to oxidize sulfide, thiosulfate and/or sulfite for energy production [56, 57]. In addition, all genomes comprised genes for sulfide:quinone oxido-reductase type I and VI (SQR), which can convert sulfide to sulfane sulfur [57]. With the exception of *Ca.* V. soyoae 2 and *Ca*. V. *okutanii*, the Clade I lineages contained two copies of the gene encoding SQR type I, whereas this gene was present as a single copy in Clade II.

Based on their gene content, it is likely that all symbiont lineages can use a variety of different enzymes to conserve energy via cross-membrane electron transport, including NADH-ubiquinone

oxidoreductase (Complex I), SQR, bacterial Rnf complex, cytochrome bc_1 complex (Complex III), terminal cbb3-type cytochrome-c-oxidase (Complex IV) and F_0F_1-type ATP synthase (Complex V).

Inorganic carbon fixation and biosynthetic processes

Members of both clades encoded a form II ribulose bisphosphate carboxylase (cbbM) and other key enzymes for carbon assimilation via the Calvin-Benson-Bassham cycle as well as a complete gene set for the non-oxidative branch of the pentose phosphate pathway. Both symbiont clades lacked the gene for sedoheptulose-bisphosphatase and might instead rely on a reversible pyrophosphate-dependent phosphofructokinase (PPi-PFK) to interconvert between sedoheptulose 1,7-bisphosphate and sedoheptulose 7-phosphate. PPi-PFK is likely also used to catalyze the phosphorylation of fructose-6-phosphate to fructose 1,6-bisphosphate during glycolysis, as the gene for its ATP-dependent homolog was absent in all vesicomyid symbiont genomes [7].

All symbiont lineages have the potential to further metabolize glycolytic intermediates and end products via a partial tricarboxylic acid (TCA) cycle and pentose phosphate pathway to produce precursors for the generation of several macronutrients, coenzymes and nucleotides.

Functional gene copies of oxoglutarate dehydrogenase and fumarate reductase appeared to be missing from all genomes. Both clades contained complete gene sets for the independent biosynthesis of 19 amino acids and a variety of enzyme cofactors, including most vitamins and their derivatives (e.g., coenzyme A, FAD, NAD^+), hemes and sirohemes, porphyrins, molybdopterin, ubiquinone and glutathione. The gene encoding homoserine kinase (*thrB*), an essential enzyme in threonine biosynthesis, was missing from all symbiont genomes [7], although it is possible that its function might be performed by a serine/threonine kinase that was present in genomes from both the Clade I and Clade II symbionts. Similarly, a separate gene for histidinol phosphatase involved in histidine biosynthesis was lacking from all symbiont genomes. However, the genomes of symbionts from both clades contained homologs of the *hisB* gene, which encodes a bifunctional imidazoleglycerol-phosphate dehydratase/histidinol-phosphatase. Pathways for the generation of retinol, cobalamin, ascorbic acid, cholecalciferol, menaquinone and tocopherol were incomplete, while protoheme biosynthesis appeared to occur through a novel form of protoporphyrinogen IX oxidase (HemJ), which has so far mostly been described in cyanobacteria

[58]. As previously noted, the *ubiD*/*ubiX* gene complex for ubiquinone biosynthesis was absent in all symbiont lineages [7]. The lack of UbiD/UbiX might be compensated through acquisition of metabolic intermediates from the host or through an alternative, currently undescribed pathway.

Methionine synthase

Clade I and Clade II symbionts appear to use different enzymes for the synthesis of methionine. The gene for the cobalamin-dependent homocysteine methyltransferase (*metH*) as well as genes for cobalamin (precursor) transport and conversion (*btuM*, *btuR/cobA*) were conserved in genomes of Clade I but were missing or degenerated in all of the Clade II lineages, except for *Ca.* R. phaseoliformis and *Ca.* R. southwardae. Conversely, the gene for the cobalamin-independent version of this enzyme (*metE*) along with its transcriptional activator (*metR*) were exclusively found in the Clade II symbiont genomes. Notably, almost all genes for *de novo* cobalamin biosynthesis were absent from the investigated symbiont genomes, with the exception of cobyrinic acid A,C-diamide synthase (*cbiA*) (all genomes), adenosylcobalamin/alpha-ribazole phosphatase (*cobC*) (not in *Ca.* R. *magnifica*) and the high affinity cobalamin transporter BtuB (*Ca*. V. gigas, *Ca.* V. marissinica, *Ca.* R. southwardae, *Ca.* R. phaseoliformis).

Nitrate reductase

An operon coding for the membrane-bound nitrate-reductase complex NarGHIJ was conserved in all Clade I symbiont genomes, but not in those of Clade II, which appeared to contain non-functional remnants of this operon. Conversely, the Clade II symbionts encode the cytoplasmic assimilatory nitrate reductase NasA, which was degenerated in Clade I.

Cysteine dioxygenase, cysteine sulfinic acid decarboxylase and isocitrate lyase

The gene coding for cysteine dioxygenase type I (*cdo*), which catalyzes the conversion of L-cysteine to cysteine sulfinic acid, was conserved in all Clade I lineages, but was absent or degenerated in most Clade II symbiont genomes (with the exception of *Ca.* R. phaseoliformis and *Ca.* R. pliocardia). CDO occurs in an operon with a pyridoxal phosphate dependent enzyme of the aspartate aminotransferase superfamily, which likely has cysteine sulfinic acid decarboxylase activity based on sequence homology with corresponding genes of *Bathymodiolus* mussel

endosymbionts (GenBank: SEH86284; recognition motif: $W_1aa_{19}S_2aaC_3$). By contrast, only the Clade II symbionts encode genes for isocitrate lyase (*icl*), a key enzyme of the glyoxylate cycle.

Transcription, translation and post-translational modification

All vesicomyid symbiont genomes contained an operon for a Class Ia ribonucleotide reductase (*nrdAB*), but only the Clade II lineages appeared to also encode the gene for its transcriptional repressor (*nrdR*). In addition, we found genes for several enzymes involved in protein modification and response to cellular stress in the Clade II genomes that were absent in Clade I. For instance, all Clade II lineages contained genes for the GTP-binding protein HflX (exception: *Ca*. R. pliocardia), and the peptide methionine sulfoxide reductase MsrB, which play a role in dissociation of translationally arrested ribosomes [59], and protein repair after oxidative damage, respectively. Likewise, most Clade II lineages encoded genes for GidB and other methyltransferases, which are involved in RNA modification.

Cell wall and membrane biosynthesis

The two symbiont clades differed in several genes that are involved in biogenesis of the cellular envelope. Although we found complete pathways for the production of the common membrane lipid phosphatidylethanolamine in the genomes of all vesicomyid symbiont lineages, genes for diacylglycerol kinase (*dgkA*), which is necessary for phospholipid recycling, was only present in the Clade II symbionts. Similarly, all Clade II symbionts encoded a 1,6-anhydro-N-acetylmuramate kinase (AnmK) and an outer membrane lipoprotein (SlyB), which are important for cell wall recycling and integrity, respectively. The Clade II lineages also contained a small-conductance mechanosensitive channel involved in osmoregulation (MscS), a lipopolysaccharide (LPS) export system protein (LptA) involved in LPS-translocation across the periplasm, and an N-acetyl-anhydromuramyl-L-alanine amidase (AmpD) involved in cell wall degradation. Homologs of these genes were either completely missing or pseudogenized in the Clade I symbionts. Both symbiont clades possessed genes for peptidoglycan biosynthesis, although MurD, MraY and MurG enzyme functionalities might be impaired or altered by the presence of internal stop codons in the case of *Ca.* V. diagonalis and *Ca.* V. extenta.

Transport across membrane

Multiple components of a type I secretion system (lapC, lapB, lapE, and the secreted agglutinin RTX) were found in all of the Clade II symbionts except for *Ca.* R. pliocardia. By contrast, this locus was missing in Clade I. The Clade II symbionts also encoded a putative hydrogenase/urease accessory protein (HupE), which is thought to be a nickel or cobalt transporter [60]. Although *hupE* is often associated with operons coding for [NiFe] hydrogenases, we did not find genes encoding hydrogenase subunits in any of the symbiont genomes. However, a gene encoding a nickel-dependent glyoxalase I (*gloA*) was present in the genomes of most Clade II symbionts.

DNA repair and recombination

In agreement with Kuwahara *et al.* [61] and Shimamura *et al.* [62], genes of the nucleotide excision repair pathway, *uvrA*, *uvrD*, *uvrD* paralog and *mfd*, were conserved in most symbiont genomes, while *uvrB* and *uvrC* were degenerated in all Clade I lineages. Within Clade II, *uvrA*, *uvrB*, *uvrD* paralog and *mfd* were present in all lineages, whereas *uvrC* was lost in *Ca.* R. pliocardia, and *uvrD* was lost in *Ca.* R. phaseoliformis. Many Clade II lineages contained genes for repair of alkylated DNA (*alkD*) and strand breaks (*radA*), while homologs of these genes were absent from all Clade I symbiont genomes. Furthermore, we found that essential genes involved in SOS response to DNA damage, *recA*, *recOR*, and *recX*, were lost in Clade I and *Ca.* R. *magnifica*. In the other Clade II lineages, these genes were conserved with the exception of *recO*, which was degenerated in *Ca.* R. phaseoliformis. Likewise, the gene coding for RuvC, an essential component of the last step of the recF and recBCD pathways for homologous recombination [63], as well as the genes coding for the XerCD recombinase system and the DNA recombination protein RmuC were lost in virtually all Clade I lineages, but conserved in most of the Clade II symbionts.

Mobile elements and defense against pathogens

The genomes of all vesicomyid symbionts are notably sparse in genes related to anti-viral defense and transposition. Phage-related genes except for a putative phage tape measure protein were completely missing in Clade I, while a few transposases, integrases and other phage-derived proteins were found in some of the Clade II lineages, in particular *Ca.* R. southwardae and *Ca.* R. phaseoliformis. In addition, remnants of type I restriction-modification systems (HsdMRS) and

mRNA-degrading toxin-antitoxin systems (e.g., MazEF) were present in the genomes of *Ca.* R. fausta, *Ca.* R. pacifica, *Ca.* R. rectimargo, *Ca.* R. phaseoliformis and *Ca.* R. southwardae, but lost in all other symbiont genomes. *Candidatus* R. southwardae and *Ca.* R. phaseoliformis further contained degenerated operons for the 5-methylcytosine-specific restriction endonuclease McrBC. In addition, putatively defunct versions of Cascade complex genes that were previously part of a CRISPR/Cas system were found in *Ca.* R. pliocardia (*cas2*) and *Ca.* R. southwardae (*cas1*, *cas3*).

References

1. Johnson SB, Krylova EM, Audzijonyte A, Sahling H, Vrijenhoek RC. Phylogeny and origins of chemosynthetic vesicomyid clams. Syst Biodivers. 2017; 15(4): 346–360.
2. Ozawa G, Shimamura S, Takaki Y, Yokobori SI, Ohara Y, Takishita K, *et al.* Updated mitochondrial phylogeny of Pteriomorph and Heterodont Bivalvia, including deep-sea chemosymbiotic *Bathymodiolus* mussels, vesicomyid clams and the thyasirid clam *Conchocele* cf. *bisecta*. Mar Genomics. 2017; 31: 43–52.
3. Ip JC-H, Xu T, Sun J, Li R, Chen C, Lan Y, *et al.* Host-endosymbiont genome integration in a deep-sea chemosymbiotic clam. Mol Biol Evol. 2021; 38(2): 502–518.
4. Yang M, Gong L, Sui J, Li X. The complete mitochondrial genome of *Calyptogena marissinica* (Heterodonta: Veneroida: Vesicomyidae): Insight into the deep-sea adaptive evolution of vesicomyids. PLoS ONE 2019; 14(9): e0217952.
5. Russell SL, Pepper-Tunick E, Svedberg J, Byrne A, Ruelas Castillo J, Vollmers C, *et al.* Horizontal transmission and recombination maintain forever young bacterial symbiont genomes. PLoS Genet. 2020; 16(8): e1008935.
6. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, *et al.* GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017; 45(W1): W6–11.
7. Newton ILG, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, *et al.* The *Calyptogena magnifica* chemoautotrophic symbiont genome. Science. 2007; 315(5814): 998–1000.
8. Liu H, Cai S, Zhang H, Vrijenhoek RC. Complete mitochondrial genome of hydrothermal vent clam *Calyptogena magnifica*. Mitochondrial DNA Part A. 2016; 27(6): 4333–4335.
9. Kuwahara H, Yoshida T, Takaki Y, Shimamura S, Nishi S, Harada M, *et al.* Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. Curr Biol. 2007; 17(10): 881–886.
10. Peek AS, Gustafson RG, Lutz RA, Vrijenhoek RC. Evolutionary relationships of deep-sea hydrothermal vent and cold-water seep clams (Bivalvia: Vesicomyidae): results from the mitochondrial cytochrome oxidase subunit I. Mar Biol. 1997; 130(2): 151–161.
11. Shah V, Morris RM. Genome sequence of "*Candidatus* Thioglobus autotrophica" strain EF1, a chemoautotroph from the SUP05 clade of marine Gammaproteobacteria. Genome Announc. 2015; 3(5): e01156–15.
12. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114–2120.
14. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821–829.
15. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19(5): 455–477.

16. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012; 28(12): 1647–1649.
17. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res. 2013; 41(13): e129.
18. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2010; 27(4): 578–579.
19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012; 9(4): 357–359.
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16): 2078–2079.
21. Laslett D, Canbäck B. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008; 24(2): 172–175.
22. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, *et al.* The RAST Server: rapid annotations using subsystems technology. BMC Genomics 2008; 9: 75.
23. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013; 29: 1072–1075.
24. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020; 36(6): 1925–1927.
25. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018; 9: 5114.
26. Lemoine F, Lespinet O, Labedan B. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evol Biol. 2007; 7(1): 237.
27. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019; 20: 238.
28. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30: 772–780.
29. Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010; 5: e9490.
30. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421.
31. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016; 32(18): 2847–2849.
32. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. Stajich JE, editor. PLoS ONE. 2010; 5(6): e11147.
33. Tesler G. GRIMM: genome rearrangements web server. Bioinformatics. 2002; 18(3): 492–493.
34. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012; 61(3): 539–542.
35. Warren DL, Geneva AJ, Lanfear R. RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. Mol Biol Evol. 2017; 34(4): 1016–1020.

36. Larget BR, Kotha SK, Dewey CN, Ané C. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics. 2010; 26(22): 2910–2911.
37. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination detection. Bioinformatics. 2006; 22(24): 3096–3098.
38. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25(11): 1422–1423.
39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5): 1792–1797.
40. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994; 11(5): 725–736.
41. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. Less Is More: An adaptive Branch-Site Random Effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 2015 May 1;32(5):1342–1353.
42. Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, *et al.* Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinformatics. 2012; 13(1): 43.
43. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Mol Biol Evol. 2015; 32(3): 820–832.
44. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014; 42(Database issue): D206–214.
45. Fisher SRA. Confidence limits for a cross-product ratio. Aust J Stat. 1962; 4(1): 41.
46. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, *et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Mol Biol Evol. 2013; 30(5): 1196–1205.
47. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, *et al.* Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012; 8: e1002764.
48. Spielman SJ, Weaver S, Shank SD, Magalis BR, Li M, *et al.* Evolution of viral genomes: Interplay between selection, recombination, and other forces. In: Anisimova M (ed). Evolutionary Genomics. Methods in Molecular Biology. (Humana, New York, NY, 2019) pp 427–468.
49. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, *et al.* PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Res. 2014; 42(W1): W337–W343.
50. Dubilier N, Bergin C, Lott C. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nat Rev Microbiol. 2008; 6: 725–740.
51. Nakagawa S, Takai K. Deep-sea vent chemoautotrophs: diversity, biochemistry and ecological significance. FEMS Microbiol Ecol. 2008; 65: 1–14.
52. Klatt JM, Polerecky L. Assessment of the stoichiometry and efficiency of CO_2 fixation coupled to reduced sulfur oxidation. Front Microbiol. 2015; 6: 484.
53. Kato K, Tanaka R, Sano S, Tanaka A, Hosaka H. Identification of a gene essential for protoporphyrinogen IX oxidase activity in the cyanobacterium *Synechocystis* sp. PCC6803. PNAS 2010; 107(38): 16649–16654.

54. Zhang Y, Mandava CS, Cao W, Li X, Zhang D, *et al.* HflX is a ribosome-splitting factor rescuing stalled ribosomes under stress conditions. Nat Struct Mol Biol. 2015; 22(11): 906–913.
55. Eitinger T, Suhr J, Moore L, Smith JAC. Secondary transporters for nickel and cobalt ions: theme and variations. Biometals 2005; 18(4): 399–405.
56. Kuwahara H, Takaki Y, Shimamura S, Yoshida T, Maeda T, *et al.* Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of *Calyptogena* clams. BMC Evol Biol. 2011; 11: 285.
57. Shimamura S, Kaneko T, Ozawa G, Matsumoto MN, Koshiishi T, *et al.* Loss of genes related to Nucleotide Excision Repair (NER) and implications for reductive genome evolution in symbionts of deep-sea vesicomyid clams. PLoS ONE 2017; 12: e0171274.
58. Connolly B, Parsons CA, Benson FE, Dunderdale HJ, Sharples GJ, *et al.* Resolution of Holliday junctions in vitro requires the *Escherichia coli ruvC* gene product. PNAS 1991; 88(14): 6063–6067.

Supplementary Figures

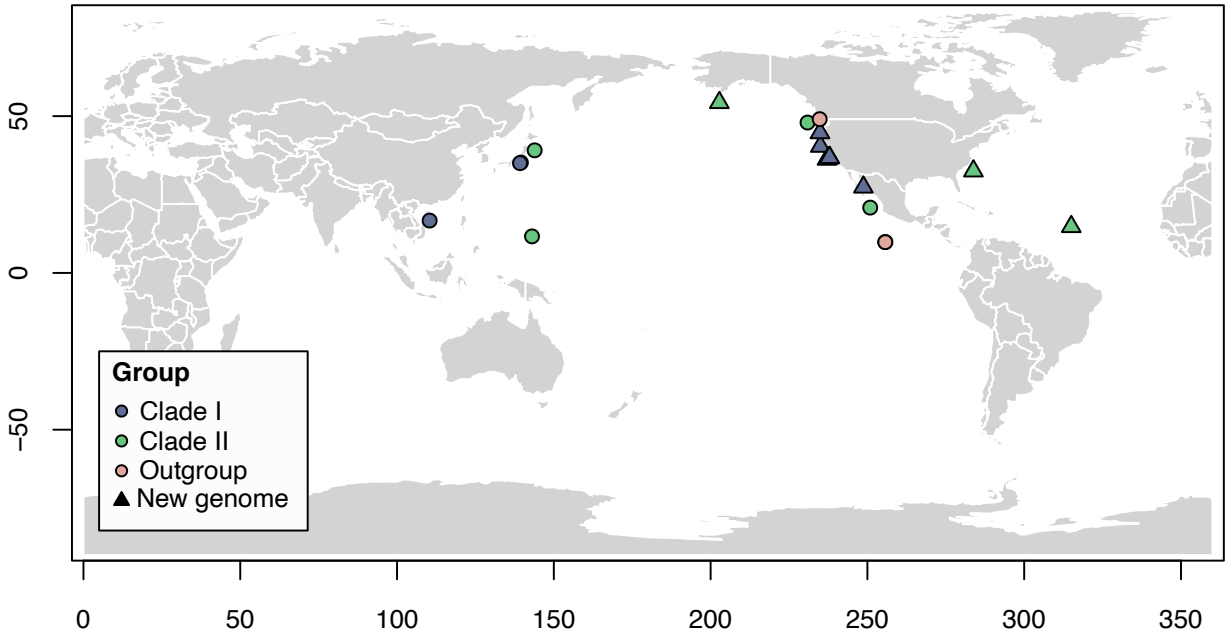

Figure S2.1 Global distribution of bacterial species compared in this study.

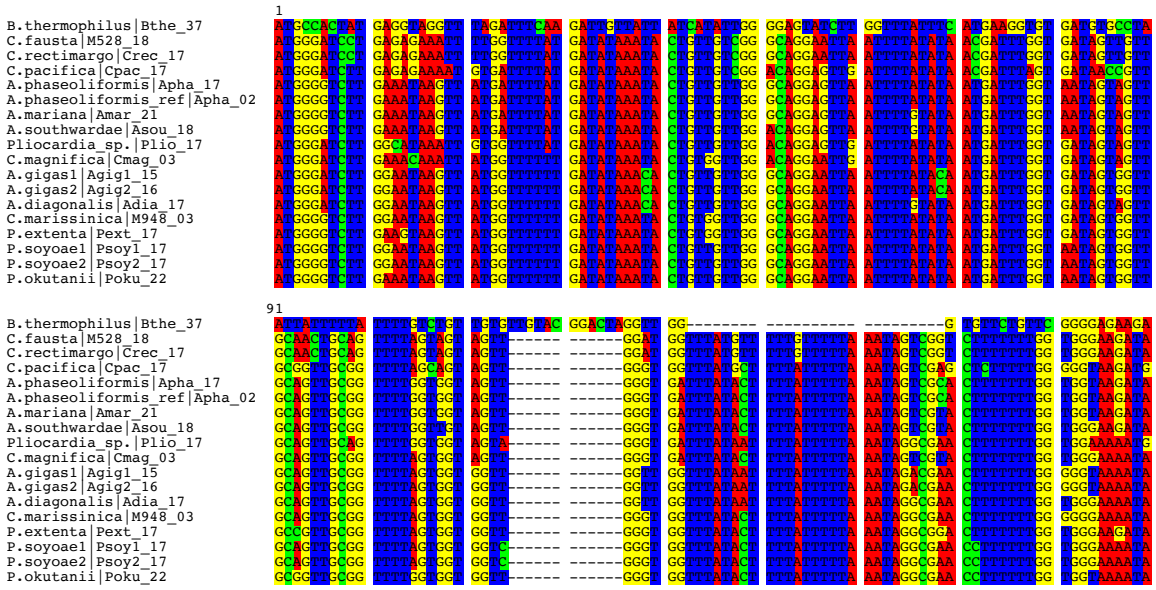

Figure S2.2 Multiple sequence alignments for the mitochondrial cox2 gene.

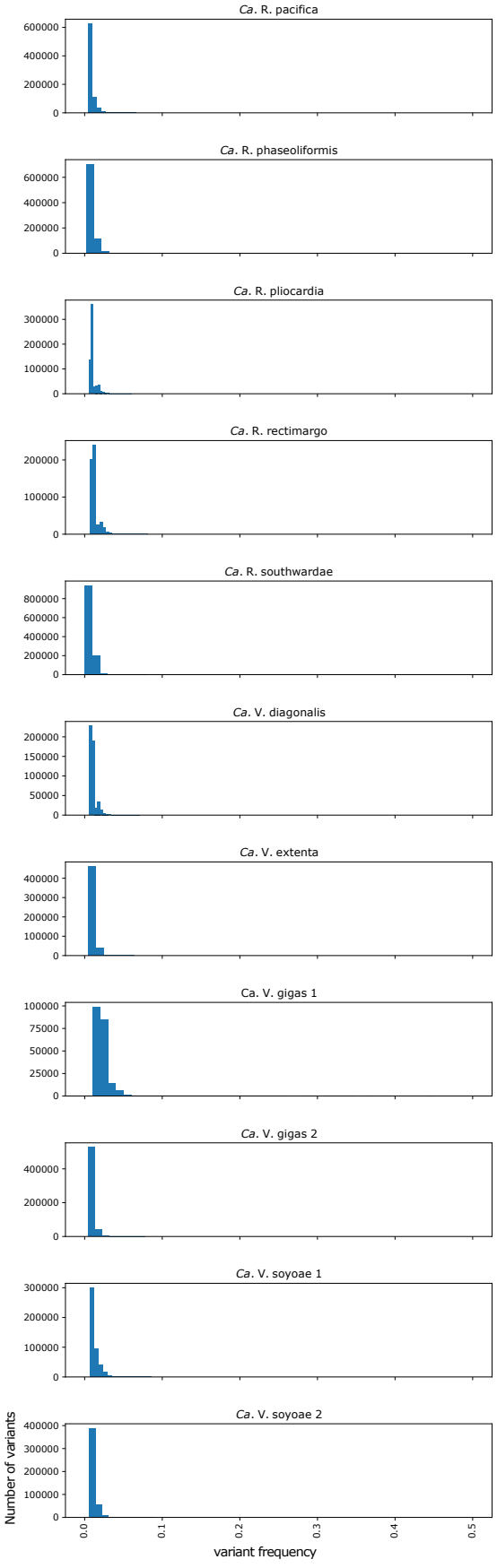

Figure S2.3 Variant frequency distributions for the intra-host symbiont populations sequenced in this study (20 bin histograms). Single nucleotide polymorphism frequencies were computed from the raw symbiont genome coverage for each species. Reads were mapped to the reference with BOWTIE2 using the --very-sensitive-local parameter. Note that the high abundance of low frequency SNPs is typical of monoclonal bacterial populations.

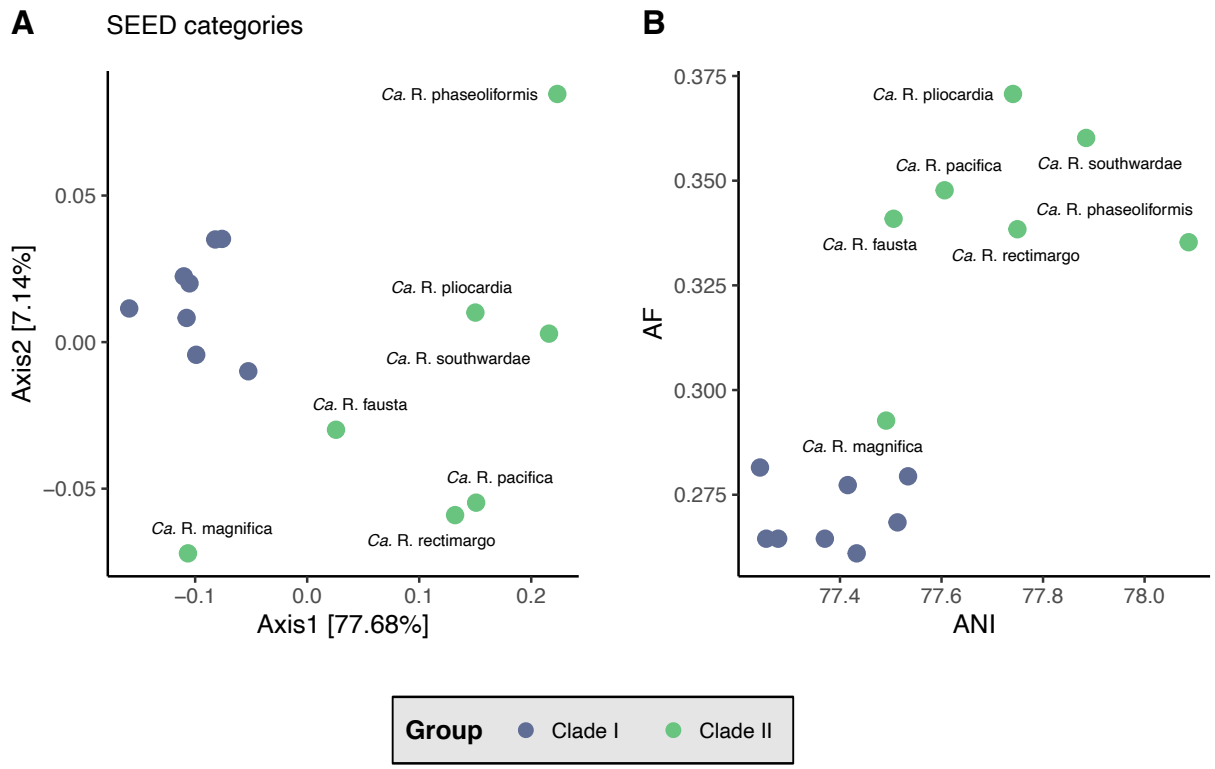

Figure S2.4 Discrimination of symbiont genomes based on A) functional characteristics (SEED categories) and B) relatedness indices. The two clades segregate largely into two groups, with symbiont genomes of Clade I being more homogenous than those of Clade II.

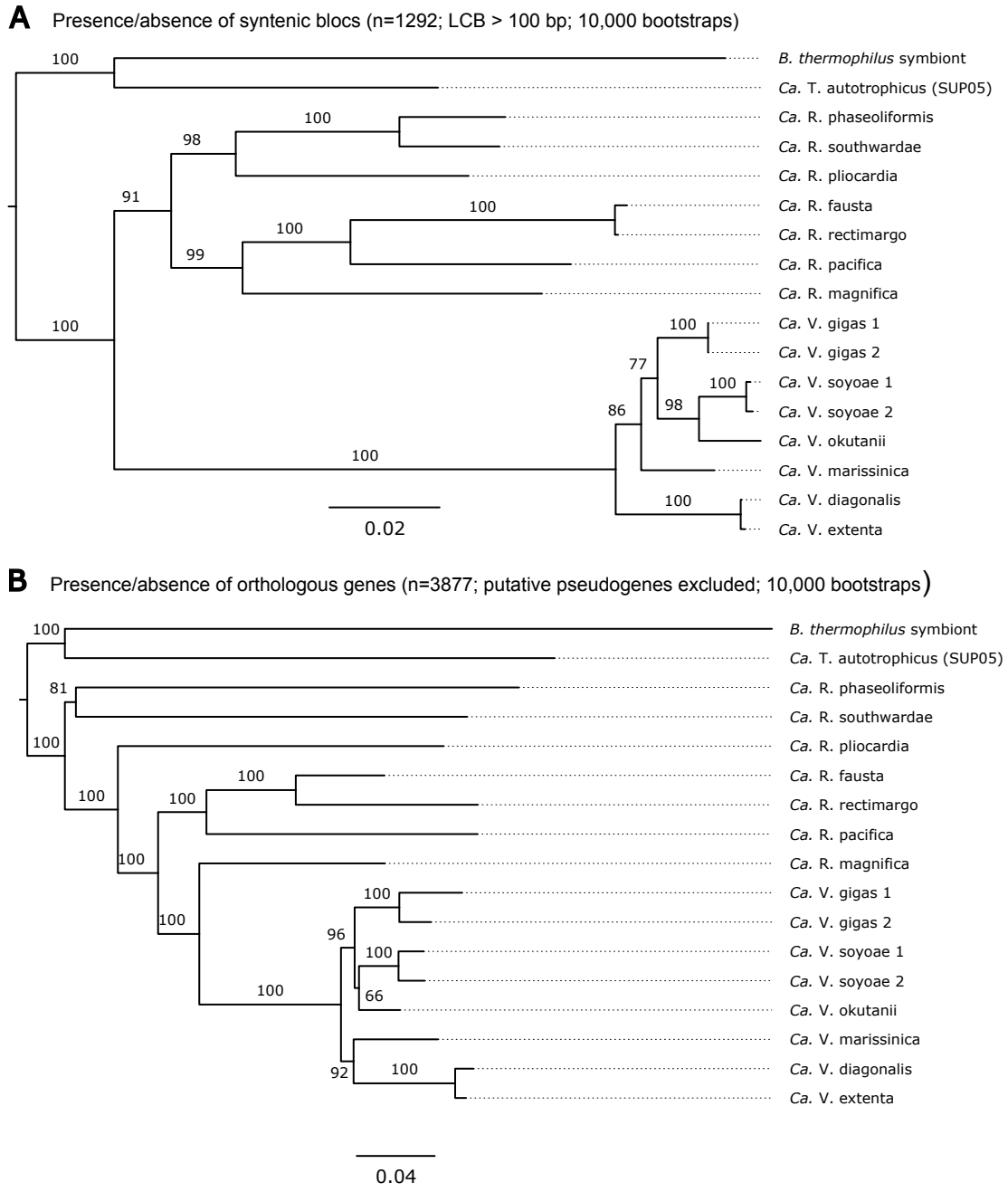

Figure S2.5 Jaccard distance-based neighbor-joining trees established from A) the presence/absence of syntenic blocs (LCBs > 100bp) and B) the presence/absence of positionally orthologous genes. Numbers above branches are bootstrap support values.

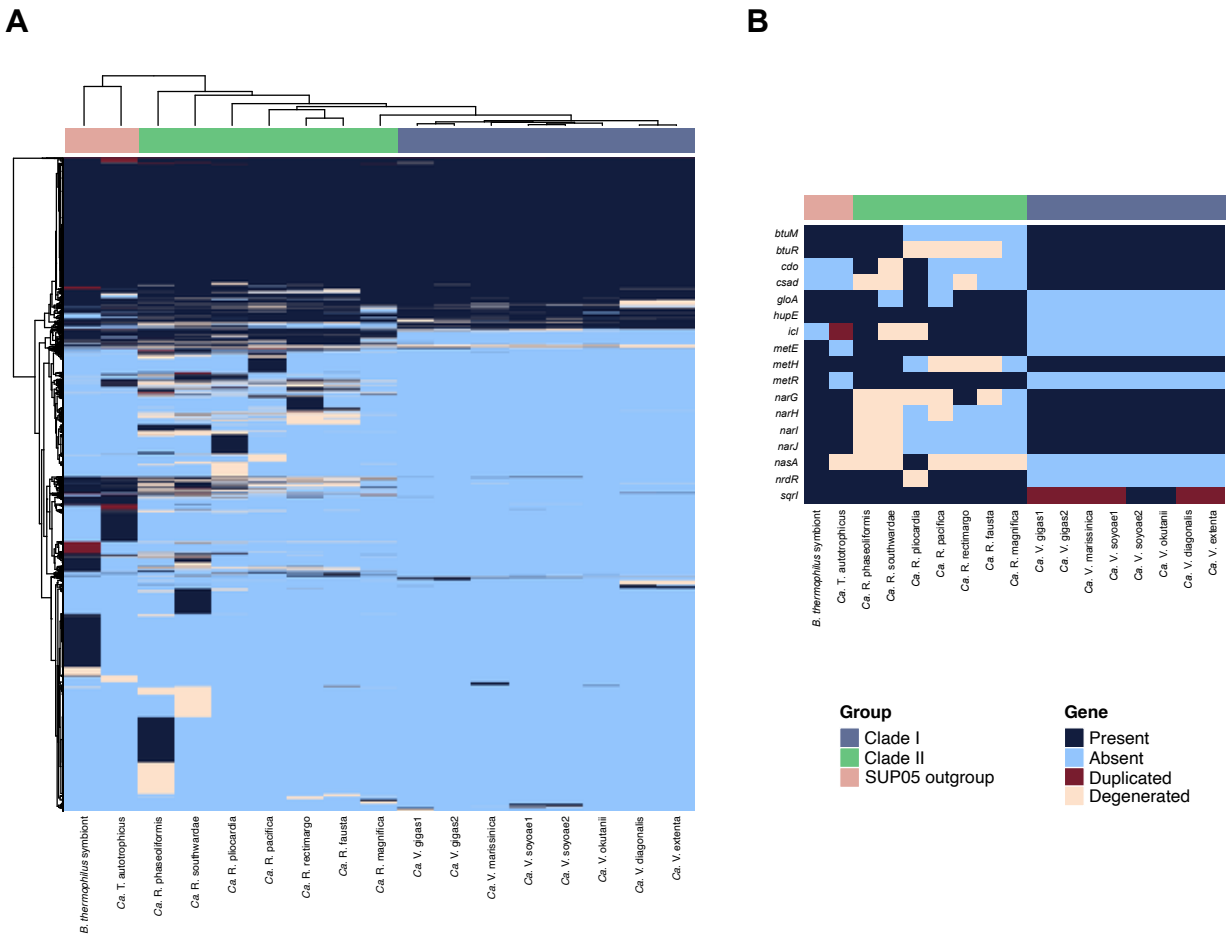

Figure S2.6 A) Heatmap of gene presence/absence, duplication and pseudogenization patterns in symbiont and outgroup genomes based on Manhattan distances and complete clustering. Clade I, Clade II and their relatives from three separate groups based on these genomic characteristics, although *Ca*. R. *magnifica* assumes an intermediate position between symbiont clades. The presence of pseudogenes is more pronounced in Clade II compared to Clade I, in agreement with the less advanced state of genome reduction in this symbiont group. Gene duplications are almost completely absent in the symbiont genomes. **B) Overview of gene presence/absence, duplication and pseudogenization patterns for genes that were differentially preserved between the two symbiont clades.**

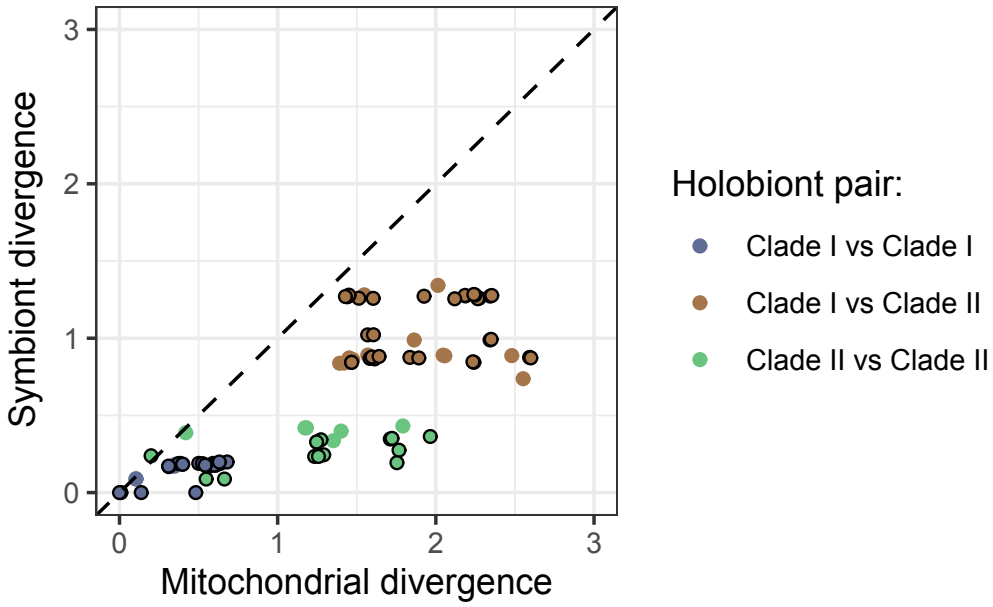

Figure S2.7 Relationship between symbiont and mitochondrial divergence. For each holobiont species, host and symbiont divergences are expressed as genome-wide pairwise synonymous substitutions rates (dS) in their respective genomes. dS values were estimated from the concatenated alignments of 13 mitochondrial and 555 symbiont protein coding genes. Putative pseudogenes and non-core protein coding genes were excluded from the analyses. ○ indicates mitochondrial and symbiont genomes isolated from a single individual.

Figure S2.8 Genome inversions between *Ca.* R *magnifica* and other bacteria. Optimal inversion scenarios were computed by GRIMM from the permutation maps exported from the whole-genome alignments.

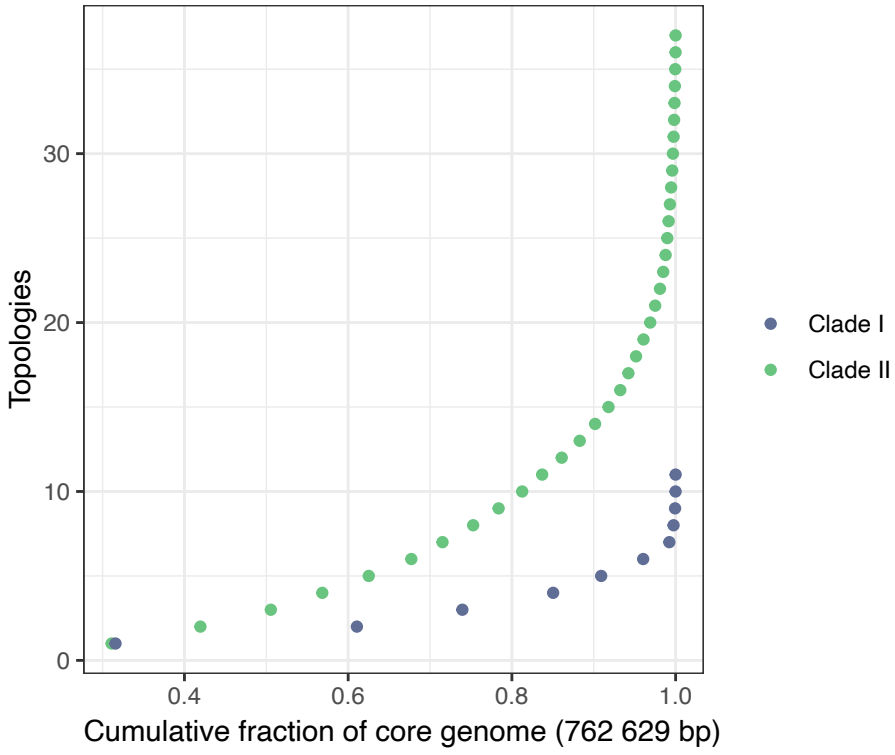

Figure S2.9 Cumulative genome representation by tree topologies as estimated by BUCKY.

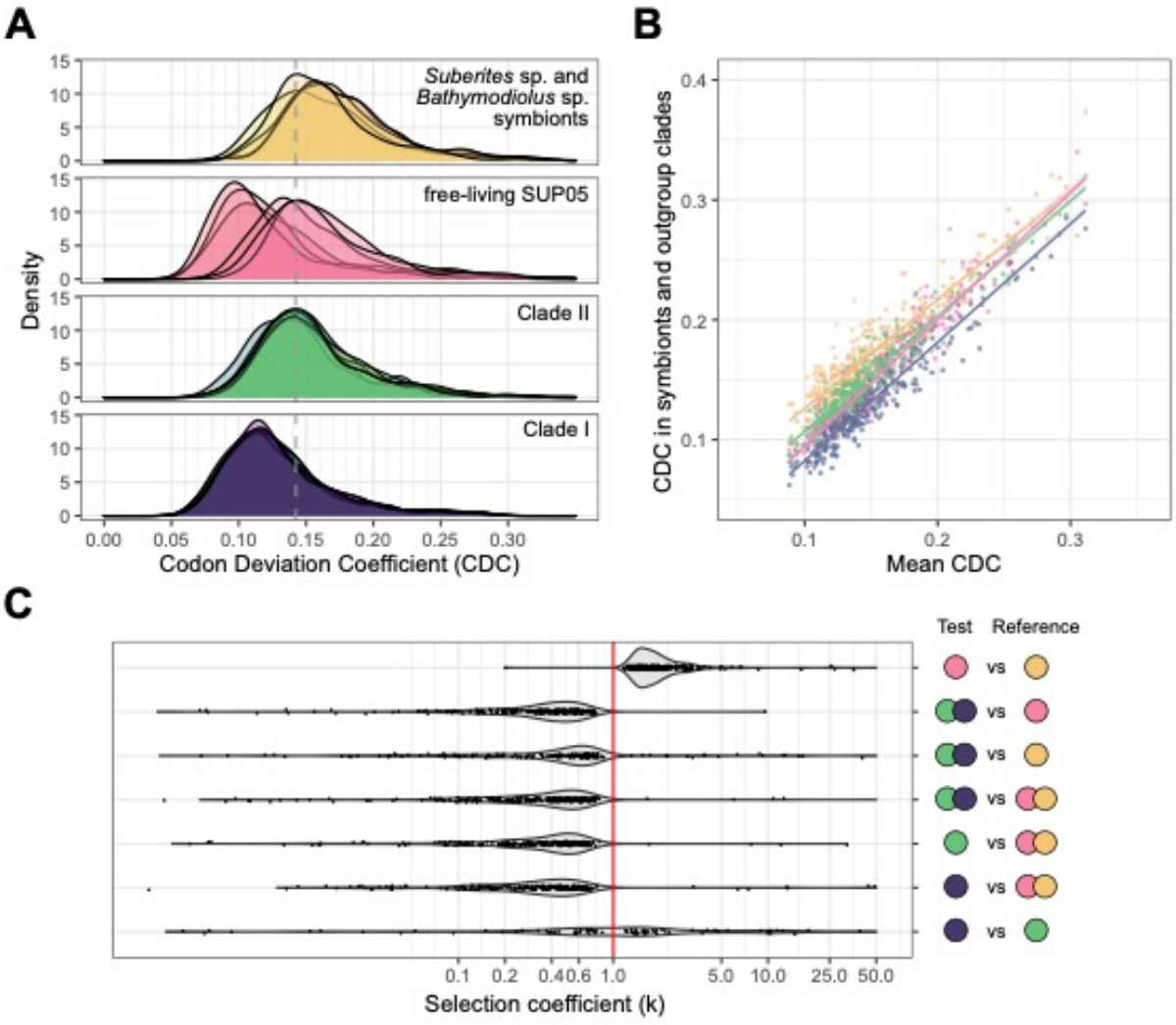

Figure S2.10 Codon usage bias and relative selection intensity in clam symbionts and bacterial relatives based on 336 core genes. A) Codon Deviation Coefficient (CDC) spectra for each genome within the outgroups: Suberites sp. and *Bathymodiolus* sp. symbionts and free-living SUP05. The dotted line shows the Clade II mean CDC. **B)** Correlation between the group-specific and dataset-specific CDC averages. CDC values vary from 0 (no bias) to 1 (maximum bias). **C)** Log-scaled selection parameter (k) spectra of core genes for which a significant change in selection was detected. CDC values were significantly lower in Clade I than Clade II, and CDC and k values were significantly lower in both symbiont clades than either of the outgroups (paired Wilcoxon-Mann Whitney test p-value < 0.01).

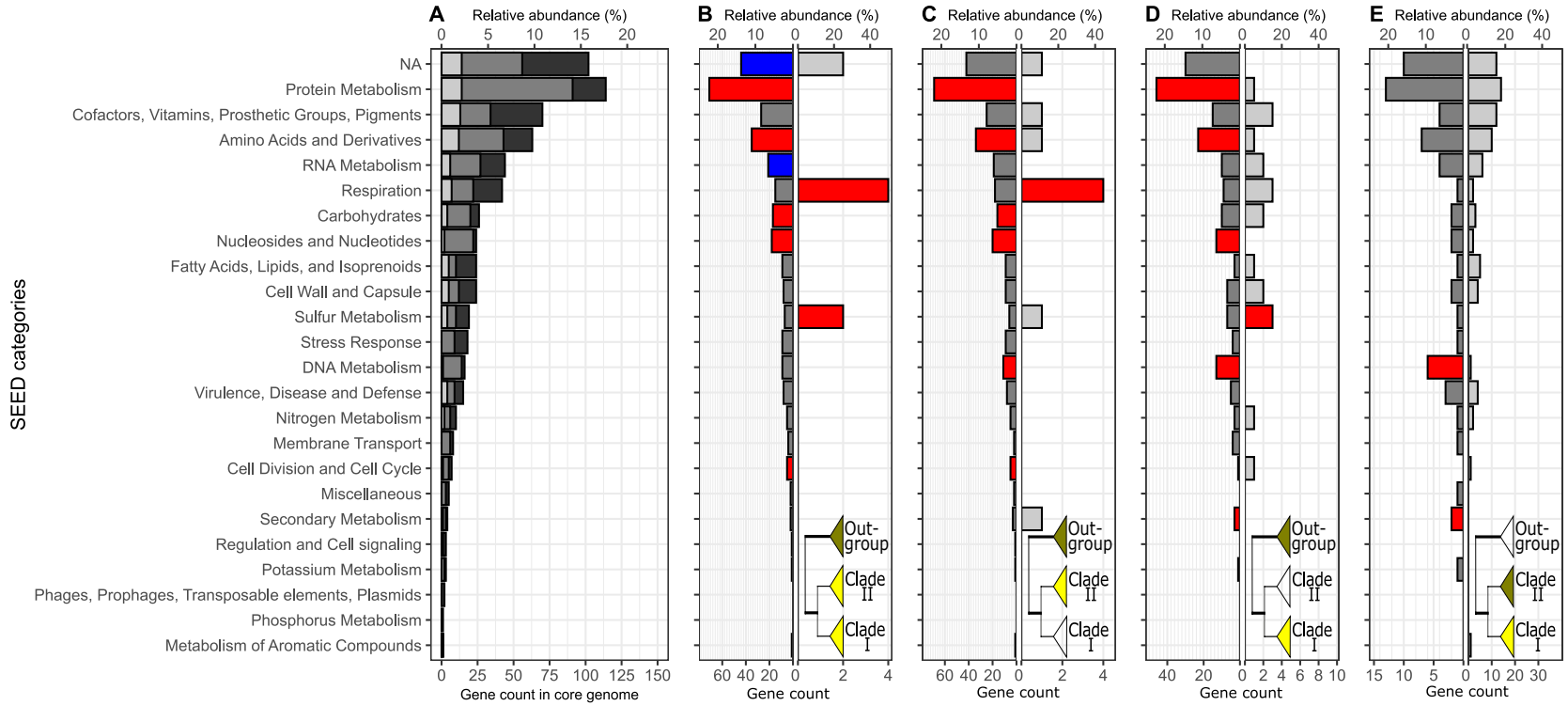

Figure S2.11 SEED category distribution of core genes under relaxed or intensified selection. Insets show the test (bright yellow) and reference (dark yellow) branches for each analysis. **A)** Distribution of all non-recombining core genes (dark grey, 555 loci) and loci under relaxed (grey, 346 loci) or intensified selection (light grey, 83 loci) within all symbiont clades. **B)** Genes with significant change in selection intensity in the symbionts compared to the outgroup. **C)** Genes with significant change in selection intensity in Clade II compared to the outgroup. **D)** Genes with significant change in selection intensity in Clade I compared to the outgroup. **E)** Genes with significant change in selection intensity in Clade I compared to the Clade II. Note that genes may be present in multiple functional categories. SEED categories significantly overrepresented (in red) and underrepresented (in blue) compared to the core genome are highlighted. Refer to text for further breakdown of these categories. NA: no functional annotation.

Supplementary Tables

Table S2.1 Sample site information for vesicomyid host-symbiont associations and bacterial relatives with free-living phase.

Host species	Symbiont	Locality	Latitude	Longitude	Depth (m)	Habitat	Dive #[a]	Year
Abyssogena mariana	*Ca.* R. mariana[b]	Mariana Trench	11.6569	143.0490	5633	Seep	n.a.	2013
Abyssogena phaseoliformis	*Ca.* R. phaseoliformis	Aleutian Trench[d]	54.3050	-157.2133	3550	Seep	–[c]	1996
Abyssogena phaseoliformis	*Ca.* R. phaseoliformis	Japan Trench	39.1052	39.1052	5347	Seep	n.a.	2009
Abyssogena southwardae	*Ca.* R. southwardae	Logatchev[d]	14.7532	-44.9805	3038	Vent	A: 3133	1997
Calyptogena fausta	*Ca.* R. fausta	High Rise	47.9666	-129.0894	2200	Vent	n.a.	1995
Calyptogena magnifica	*Ca.* R. *magnifica*	East Pacific Rise	9.8480	-104.2935	2507	Vent	n.a.	2004
Calyptogena magnifica	*Ca.* R. *magnifica*	East Pacific Rise	20.8305	-109.1030	2601	Vent	T: 555	2003
Calyptogena pacifica	*Ca.* R. pacifica	Monterey Canyon[d]	36.7739	-122.0488	650	Seep	V: 2555	2004
Calyptogena rectimargo	*Ca.* R. rectimargo	Monterey Canyon[d]	36.6816	-122.1197	1540	Seep	V: 2338	2003
Pliocardia sp.	*Ca.* R. pliocardia	Blake Spur[d]	32.4948	-76.1847	2155	Seep	A: 3710	2001
Archivesica diagonalis	*Ca.* V. diagonalis	Monterey Canyon[d]	36.2274	-122.8796	3456	Seep	T: 488	2002
Archivesica gigas (1)	*Ca.* V. gigas 1	East Gorda Escarpment[d]	40.3580	-125.0210	2094	Vent	T: 351	2001
Archivesica gigas (2)	*Ca.* V. gigas 2	Guaymas Transform Fault[d]	27.3400	-111.2700	1754	Seep	T: 548	2003
Archivesica marissinica	*Ca.* V. marissinica	Haima	16.6943	110.3969	1361	Seep	n.a.	2018
Phreagena extenta	*Ca.* V. extenta	Monterey Canyon[d]	36.6088	-122.4366	2889	Seep	T: 406	2002
Phreagena okutanii	*Ca.* V. *okutanii*	Sagami Bay	35.2000	139.5000	1157	Seep	HD: 305	2004
Phreagena okutanii	*Ca.* V. *okutanii*	Sagami Bay	35.0157	139.2220	852	Seep	n.a	2007
Phreagena soyoae (1)	*Ca.* V. soyoae 1	Oregon Subduction Zone[d]	44.6755	-125.1182	765	Seep	A: 2796	1994
Phreagena soyoae (2)	*Ca.* V. soyoae 2	Monterey Canyon[d]	36.7762	-122.0842	985	Seep	V: 2059	2001
Bathymodiolus thermophilus[b]	*B. thermophilus* symbiont	East Pacific Rise	9.8200	-104.3000	2518	Vent	n.a.	2000
n.a.	*Ca.* T. autotrophicus	Effingham Inlet	49.0369	-125.2080	60	OMZ	n.a.	2013

[a]Submersibles: A = Alvin, HD = Hyper Dolphin, T = Tiburon, V = Ventana

[b]Mitochondrial and/or symbiont genome not assembled

[c]Sampled with TV grab

[d]Sampled in this study

Table S2.2 SEED categories for all genes.

Provided as supplementary file

Table S2.3 Positional core orthologs. Positional orthology was exported from the Mauve genome alignments (function getOrthologList, min identity=35; min coverage=51).

Provided as supplementary file

Table S2.4 Orthogroups for all genes.

Provided as supplementary file

Table S2.5 General information about bacterial and mitochondrial genomes analyzed in this study.

Genome	Accession No.	Genome size / N50 (Mb)	Contigs	GC (%)	Coverage (X)	CDS	Pseudogenes	tRNAs	rRNAs	Reference
Bacterial genomes										
Ca. R. fausta	CP054490	1.188	1	36.70	200	1023	181	36	3	Russel *et al.* 2020
Ca. R. *magnifica*	CP000488	1.160	1	34.03	14	953	23	36	3	Newton *et al.* 2007
Ca. R. pacifica	CP060683	1.184	1	36.58	140	1102	354	35	3	This study
Ca. R. phaseoliformis	JACRUR01	1,527 / 0,37	8	36.93	69	1445	765	36	3	This study
Ca. R. pliocardia	CP060688	1.231	1	36.98	113	1200	443	36	3	This study
Ca. R. rectimargo	CP060684	1.189	1	36.69	91	1143	333	37	3	This study
Ca. R. southwardae	JACRUS01	1,586 / 0,06	39	36.86	90	1414	621	36	3	This study
Ca. V. diagonalis	CP060680	1.024	1	31.10	110	902	103	36	3	This study
Ca. V. extenta	CP060685	1.024	1	31.10	137	896	99	36	3	This study
Ca. V. gigas 1	CP060681	1.034	1	31.37	49	914	94	36	3	This study
Ca. V. gigas 2	CP060682	1.033	1	31.38	153	929	50	36	3	This study
Ca. V. marissinica	CP054877	1.032	1	31.20	570	928	53	36	3	Ip et al 2021
Ca. V. *okutanii*	AP009247	1.022	1	31.59	9	918	19	35	3	Kuwahara *et al.* 2007
Ca. V. soyoae 1	CP060687	1.019	1	31.63	89	947	45	36	3	This study
Ca. V. soyoae 2	CP060686	1.017	1	31.63	110	929	54	36	3	This study
B. thermophilus symbiont	CP024634	2.832	1	39.00	126	1928	139	36	3	Patra *et al.* 2022
Ca. T. autotrophicus	CP010552	1.512	1	39.00	106	1463	43	35	3	Shah *et al.* 2015
Mitochondrial genomes										
Abyssogena mariana	LC126311	0.016	1	30	n.a.	13	n.a.	23	2	Ozawa *et al.* 2017
Abyssogena phaseoliformis	MT947384	0.018	1	31.00	10	13	n.a.	23	2	This study
Abyssogena phaseoliformis	AP014557	0.019	1	30.00	n.a.	13	n.a.	24	2	Ozawa *et al.* 2017
Abyssogena southwardae	MT947385	0.019	1	29.00	15	13	n.a.	24	2	This study
Archivesica diagonalis	MT947381	0.020	1	33.00	8	13	n.a.	22	2	This study
Archivesica gigas	MT947383	0.016	1	35.00	7	13	n.a.	21	2	This study
Archivesica marissinica	MK948426	0.017	1	34.60	18	13	n.a.	22	2	Yang *et al.* 2019
Calyptogena fausta	MT528632	0.017	1	33.50	n.a.	13	n.a.	27	2	Russel *et al.* 2020
Calyptogena magnifica	KR862368	0.020	1	32.00	n.a.	13	n.a.	22	2	Liu *et al.* 2016
Calyptogena pacifica	MT947386	0.020	1	31.00	18	13	n.a.	23	2	This study
Calyptogena rectimargo	MT947387	0.019	1	32.00	22	13	n.a.	25	2	This study
Phreagena extenta	MT947388	0.018	1	33.00	6	13	n.a.	22	2	This study
Phreagena *okutanii*	AP014555	0.016	1	34.00	n.a.	13	n.a.	23	2	Ozawa *et al.* 2017
Phreagena soyoae	MT947390	0.019	1	34.00	25	13	n.a.	23	2	This study
Pliocardia sp.	MT947391	0.019	1	28.00	20	13	n.a.	22	2	This study

Table S2.6 Taxonomic classification of symbiont genomes based on the Genome Taxonomy Database.

Provided as supplementary file

Table S2.7 GC content and Codon Deviation Coefficient (CDC) for core genes.

Provided as supplementary file

Table S2.8 Statistically supported ($p < 0.05$) change in intensity of selection detected by RELAX (555 non-recombining core protein coding genes tested).

Provided as supplementary file

Table S2.9 Statistically significant (Holm-Bonferroni corrected $p < 0.05$) episodic diversifying selection detected in the phylogeny. Positive selection was found in 185 out of 555 non-recombining genes

Provided as supplementary file

Table S2.10 Codon positions under pervasive or episodic positive selection in 17 candidate genes based on Fubar and Meme analyses, respectively. α = synonymous substitution rate at a site; $\beta+$ = non-synonymous substitution rate at a site for the positive/neutral evolution component; Pr = posterior probability (a value $\geqq 0.9$ indicates strong evidence for positive selection); $p+$ = very approximate proportion of branches evolving under positive selection; LRT = likelihood ratio test statistic for episodic diversification; p = p-value for episodic diversification (a value $\leqq 0.1$ indicates positive selection); n.c. = not calculated.

Provided as supplementary file

Annex II. Chapter 3 supplementary materials

Supplementary Data

Provided as supplementary file

Supplementary Tables

Table S3.1 Sample collection information.

Aggregation ID	Collection ID	Date and time of collection (UTC)	Lat	Long	Depth	Region	Flow-regime	Plume avg temp	Base Avg temp
R01L	M0011	2013-06-18 23:48	47° 56.9663 N	129° 05.9164 W	2195	Main Endeavour Field (Grotto)	Low-flow	N.A.	N.A.
R02H	M0016	2013-06-23 02:49	47° 56.9600 N	129° 05.9185 W	2190	Main Endeavour Field (Grotto)	High-flow	N.A.	N.A.
R07B	R1939-07	2016-08-09 19:57	47° 56.9995 N	129° 05.8242 W	2195	Main Endeavour Field (Hulk)	Basalt-hosted	2.8	2.8
R08H	R1939-08	2016-08-09 20:44	47° 56.9996 N	129° 05.8193 W	2197	Main Endeavour Field (Hulk)	High-flow	13.7	37.9
R09L	R1939-09	2016-08-09 21:11	47° 56.9973 N	129° 05.8158 W	2197	Main Endeavour Field (Hulk)	Low-flow	2.9	7.5
R13L	R1941-05	2016-08-11 18:10	47° 57.7791 N	129° 05.4910 W	2182	Endeavour (Clam-Bed)	High-flow	6	30
R14B	R1941-07	2016-08-11 19:40	47° 57.7834 N	129° 05.4999 W	2180	Endeavour (Clam-Bed)	Basalt-hosted	2.5	3.9
R15H	R1942-03	2016-08-12 18:21	48° 25.8171 N	128° 40.9278 W	2433	Middle valley	High-flow	20.7	24.5
R16L	R1943-08	2016-08-13 22:42	48° 27.3174 N	128° 42.5152 W	2410	Middle valley	Low-flow	2.9	21.1

Table S3.2 Genetic markers and PCR conditions used in this study.

Locus id (product)	Scaffold accession number	Locus-tag in reference genome assembly LDXT01	Start-end position of the amplicon	SNP positions in scaffold[a]	F: Forward primer R: Reverse primer Amplicon length in LDXT01 (not including CS)	PCR conditions for initial amplification[b]
V4 (16S ribosomal RNA)	KQ557152 LDXT01000000	Ga0074115_1403	2673-2964		F: **CS1**[*] + GTGYCAGCMGCCGCGGTAA R: **CS2**[†] + GGACTACNNGGGTDTCTAAT 291 bp	Denaturation: 94°C, 30s Annealing: 57°C, 30s Extension: 68°C, 30s 25-35 cycles
lpxA (acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase)	KQ557115 LDXT01000000	Ga0074115_101122	121173-121528	121194; 121508	F: **CS1** + CCGCTAGGCCAGAGATGAAG R: **CS2** + CGAGGGCTGCGTGATCG 356 bp	
pleD (diguanylate cyclase (GGDEF) domain)	KQ557117 LDXT01000000	Ga0074115_10695	112275-112783	112296; 112760	F: **CS1** + GAGGGCGTCCAACTGCTTTT R: **CS2** + TCGATACCGTGCATCTGCC 509 bp	
tufB (translation elongation factor 1A)	KQ557137 LDXT01000000	Ga0074115_12214	16103-16402	16133; 16324	F: **CS1** + ATCACCTTCCAGCGCCTTC R: **CS2** + CCCGGCCATGCTGACTAT 300 bp	
CRISPR	KQ557120 LDXT01000000	KQ557120 REGION: 48218..48978	48149-49364		F: **CS1** + TCGGCCAAAAAGATCGGTAGA R: **CS2** + GTTCGGCCTGTACCCGGAG 1216 bp	Denaturation: 95°C, 30s Annealing: 53°C, 30s Extension: 68°C, 2min 30 cycles

[a] Perez and Juniper (32)
[b] A second PCR with the following condition was performed to add sample-specific barcodes: Denaturation, 94°C, 30s; Annealing 60°C, 30s; Extension: 68°C, 45s or 2min (for CRISPR)
[*] CS1 overhang: ACACTGACGACATGGTTCTACA
[†] CS2 overhang: TACGGTAGCAGAGACTTGGTCT

Annex III. Chapter 4 supplementary materials

Supplementary Methods

Nuclear libraries decontamination

DNA and RNA samples were sent for sequencing to Novogene (Beijing, China) for sequencing on the Illumina NovaSeq platform. Paired-end libraries of 350bp insert size were constructed with the NEBNext® DNA Library Prep Kit. DNA samples were also sequenced on the Oxford Technologies Nanopore (ONT) PromethION platform.

In order to facilitate downstream processes and reduce contamination, raw reads were first binned to discriminate nuclear sequences from these of the mitochondria and known endosymbionts (Figure S4.1). We used bbmap tool bbsplit.sh to bin Illumina data using reference mitochondrial (a mitogenome provided by Dr. Yannan Sun (pers.comm.) for *P. echinospica*, NC_024653 for *R. piscesae*, and OL802212 for *P. palmiformis*) and symbiont genomes (assembly provided by Dr. Yannan Sun (pers.comm) for *P. echinospica*, and LDXT01 for *R. piscesae*). Similarly, genomic ONT reads were mapped onto the mitochondrial and symbiont references with minimap2 (v2.17) (Li 2018) and extracted with bbmap (Bushnell 2014) filterbyname.sh. The sequencing output of each library bin is presented in Figure S4.1and Table S4.1.

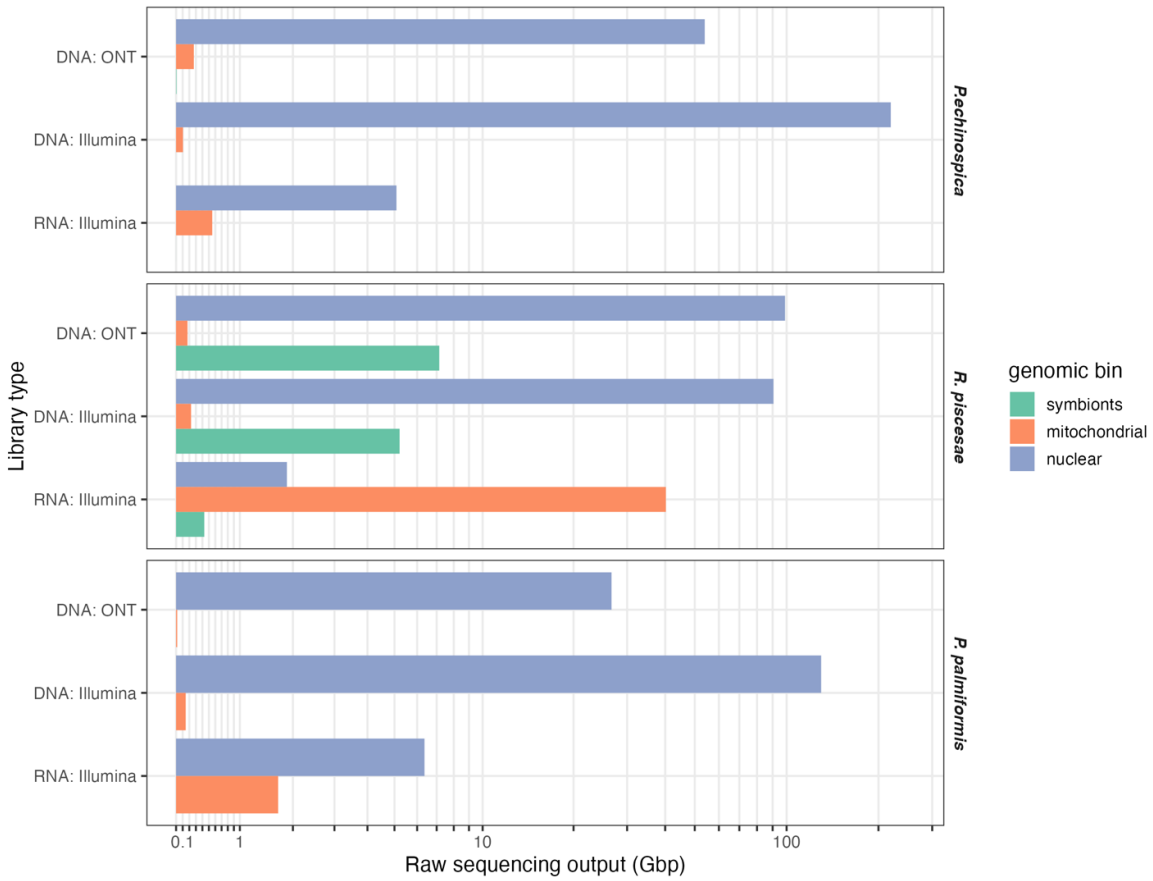

Figure S4.1 Sequencing output for each library type and genomic bin. DNA data for *P. echinospica* and *R. piscesae* are from the vestimentum tissues whereas DNA data from *P. palmiformis* are from body-wall tissues. The transcriptomic data of *P. echinospica* are from the same individual and tissue as DNA data whereas transcriptomic data of *R. piscesae* and *P. palmiformis* come from different tissues and individuals (*R. piscesae*: gills, trophosome and epiderm of BioSample SAMN26521182, and vestimentum and trophosome of BioSample SAMN26521183; *P. palmiformis*: whole body of BioSample SAMN14351933).

Table S4.1 Sequencing output in Gbp. DNA data for *P. echinospica* and *R. piscesae* are from the individual vestimentum tissues whereas DNA data from *P. palmiformis* are from body-wall tissues. The transcriptomic data of *P. echinospica* are from the same individual and tissue as DNA data whereas transcriptomic data of *R. piscesae* and *P. palmiformis* come from different tissues and individuals (*R. piscesae*: gills, trophosome and epiderm of BioSample SAMN26521182, and vestimentum and trophosome of BioSample SAMN26521183; *P. palmiformis*: whole body of BioSample SAMN14351933).

	Raw reads			Trimmed/corrected reads			
	DNA: ONT	DNA: Illumina	RNA: Illumina	DNA: ONT	DNA: Illumina	RNA: Illumina	Estimated genome size
P. echinospica							
symbionts	0.01	0.00	0.00	0.27	0.00	0.00	***0.0041***
mitochondrial	0.27	0.10	0.55	0.01	0.10	0.53	***0.000015***
nuclear	53.92	219.47	5.08	53.00	203.92	4.80	***1.1***
R. piscesae							
symbionts	7.14	5.21	0.43	7.18	5.15	0.41	***0.0036***
mitochondrial	0.17	0.23	40.20	0.19	0.23	38.80	***0.000015***
nuclear	98.84	90.56	1.87	99.47	87.06	1.68	***0.6***
P. palmiformis							
mitochondrial	0.02	0.15	1.70	0.02	0.15	1.57	***0.000016***
nuclear	26.66	129.77	6.36	26.86	129.36	5.62	***0.6***

Whole-genome bisulfite sequencing

Whole genome bisulfite sequencing libraries were prepared by Novogene (Beijing, China) for *R. piscesae* from the same DNA extraction used for Nanopore sequencing. A total of 5.2 µg of genomic DNA mixed with 26 ng of lambda DNA (unmethylated control) was fragmented to 200–300 bp by sonification using a Covaris S220 sonicator. End-repair and A-tailing were applied to the DNA fragments prior to cytosine conversion with two rounds of bisulfite treatment using the EZ DNA Methylation-Gold Kit (Zymo Research). The resulting paired-end 150 bp bisulfite library was sequenced on the Illumina NovaSeq 6000 platform to a total of 174.5 million paired-end reads (52.38 Gbp; nuclear coverage >85X). In silico, reads were trimmed with trimmomatic v0.30 (Bolger *et al.* 2014) (parameters: leading = 10, trailing = 10, minlen = 50) and processed with bismark v0.22.3 (Krueger and Andrews 2011) to compute per base-pair methylation frequencies. CpGs with less than 10x coverage (representing 15% of all CpGs) were removed.

Methylation detection from ONT data

CpG methylation was called from ONT reads following the Nanopolish (v0.13.3) pipeline with default parameters (Simpson *et al.* 2017). A certain amount of reads were lost at each step of the Nanopolish pipeline because some reads failed to map against the reference genome or raw squiggles failed to be realigned to their respective mapped reads (Figure S4.2). Due to these loss and the lower sequencing depth of the P. palmiformis Nanopore libraries, the final methylation call coverage was much lower in that species than in *R. piscesae* or P. echinospica. In P. palmiformis, 50% of the CpGs called had a coverage of 4 or lower (C50 = 4) whereas in *R. piscesae* and P. echinospica, C50 values were 53 and 21, respectively (Figure S4.3).

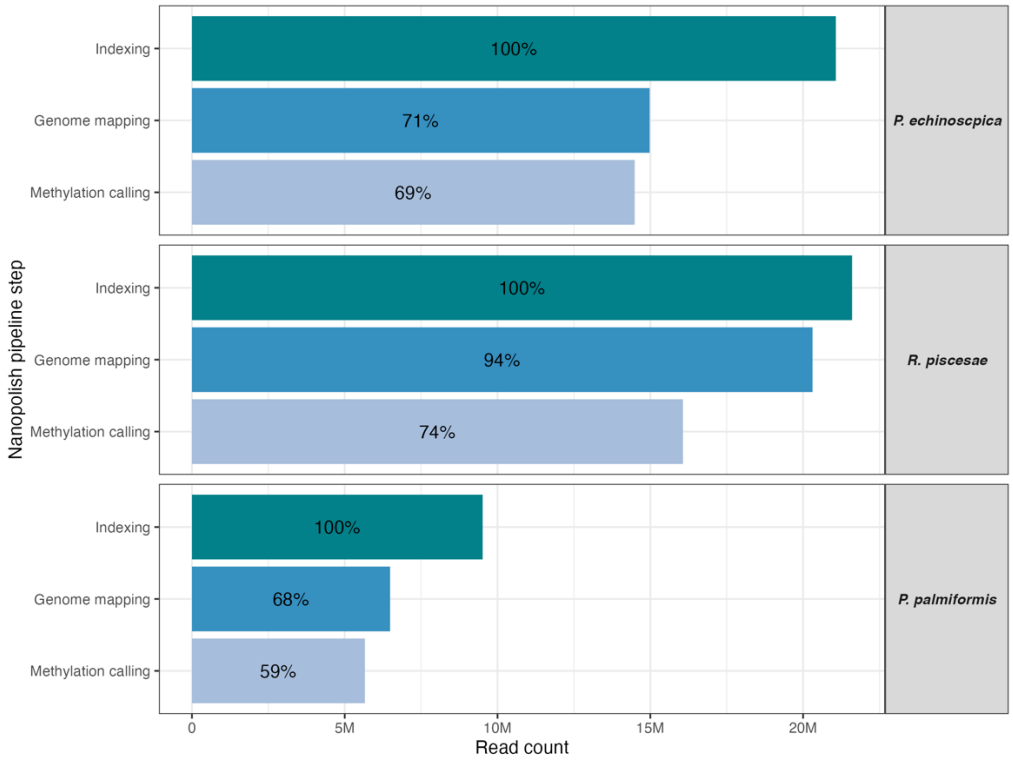

Figure S4.2 Count of unique reads remaining after each step of the Nanopolish pipeline.

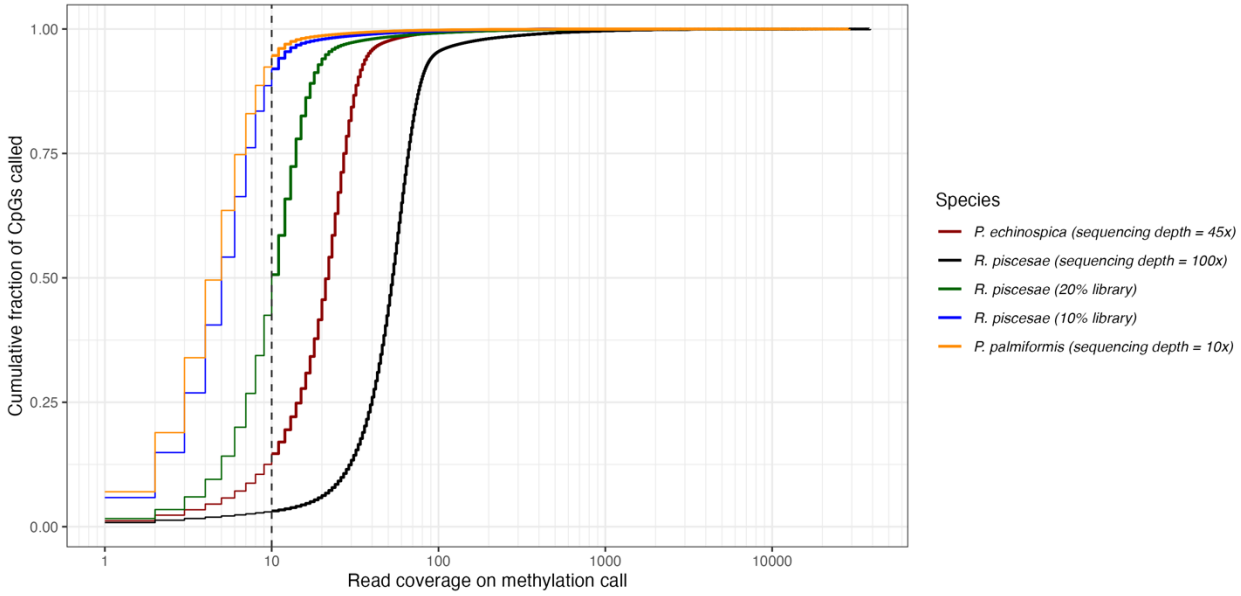

Figure S4.3 Cumulative distribution of Nanopore CpG methylation call coverage. The dotted line represent the recommended minimum coverage threshold.

Evaluation of ONT methylation detection accuracy

ONT sequencing appear to be ideal tools for quick and economical DNA methylation detection because unlike the current gold standard method of Whole-Genome Bisulfite Sequencing (WGBS), which requires a chemical or enzymatic treatment of the DNA prior to sequencing (Krueger *et al.* 2012), third-generation sequencing can detect base modification directly through variation in the electric "squiggle" signal (with Oxford Nanopore data [Simpson *et al.* 2017]). DNA methylation information thus comes "for free" with any chromosome-level genome sequencing project which typically requires long reads to improve assembly contiguity. With this technologies however, high uncertainty about the accuracy of DNA methylation calls remains, especially on non-model and atypical genomes. Hence, we first assessed the precision of DNA methylation calls derived from Oxford Nanopore sequencing by comparing whole-genome DNA methylation estimates of the same DNA extraction (of the vestimentum of one *R. piscesae* individual) obtained through Nanopore sequencing (sequencing depth >100X) and WGBS (sequencing depth >85X).

CpG motifs uniquely called with Bismark or Nanopolish were extracted with bedtools subtract (Quinlan and Hall 2010). We observed 97% overlap between whole-genome DNA methylation calls obtained through Nanopore sequencing (sequencing depth >100X; 90% CpGs called) and WGBS (sequencing depth >85X; 73% CpGs called). The per-CpG motif methylation rates detected by both methods were significantly correlated (Pearson correlation p-value <0.001, r2=0.93, Figure S4.4). Nanopore calls slightly overestimated the methylation frequencies of weakly methylated CpGs and underestimated those of strongly methylated CpGs (Wilcox tests on methylation frequencies <20% and >80% p-value < 0.001) when compared with those with WGBS. Furthermore, the incidence of methylated CpGs tend to be underestimated by Nanopore compared to WGBS in highly methylated regions and inversely in lowly methylated regions (Figure S4.5). The correlation between the two methods was the highest when methylation frequencies were averaged over genomic windows of 1 Kbp to 100 Kbp; roughly corresponding to gene sizes (FiguresFigure S4.4 and Figure S4.6).The correlation between the two methods was the highest when methylation frequencies were averaged over genomic windows of 1 to 100 Kbp; roughly corresponding to gene sizes (Figure S4.4). Applying minimum coverage thresholds also improved the overall correlation between WGBS and Nanopore particularity at small and large genomic scales (Figure S4.4). Furthermore, we tested the effect of sequencing effort on the accuracy of

Nanopore methylation frequency estimation by reducing in-silico the sequencing libraries depth of our mapped library to 20X and 10X with sambamba view (Tarasov *et al.* 2015). While the per-CpG accuracy of methylation frequency measure decreased with sequencing depth, high sequencing efforts appeared to be unnecessary to achieve acceptable estimates at gene-level scales. Indeed, when down-sampling our Nanopore library to 20X and 10X coverage, which respectively represent the mid to lower-end range of sampling effort for genome scaffolding purposes (Koren *et al.* 2017), the 1 to 100 Kbp average methylation frequency estimates derived from Nanopore methylation calls still matched those of WGBS closely.

At the level of genes, the accuracy of Nanopore methylation calls was over 98% (Table S4.2). Specificity was low because many weakly methylated genes were not called with WGBS (Table S4.2, and Figure S4.5). Nanopore's false positive detection rate compared to WGBS was greatly reduced by applying a minimum mean methylation frequency thresholds for calling methylated genes. (Table S4.3)

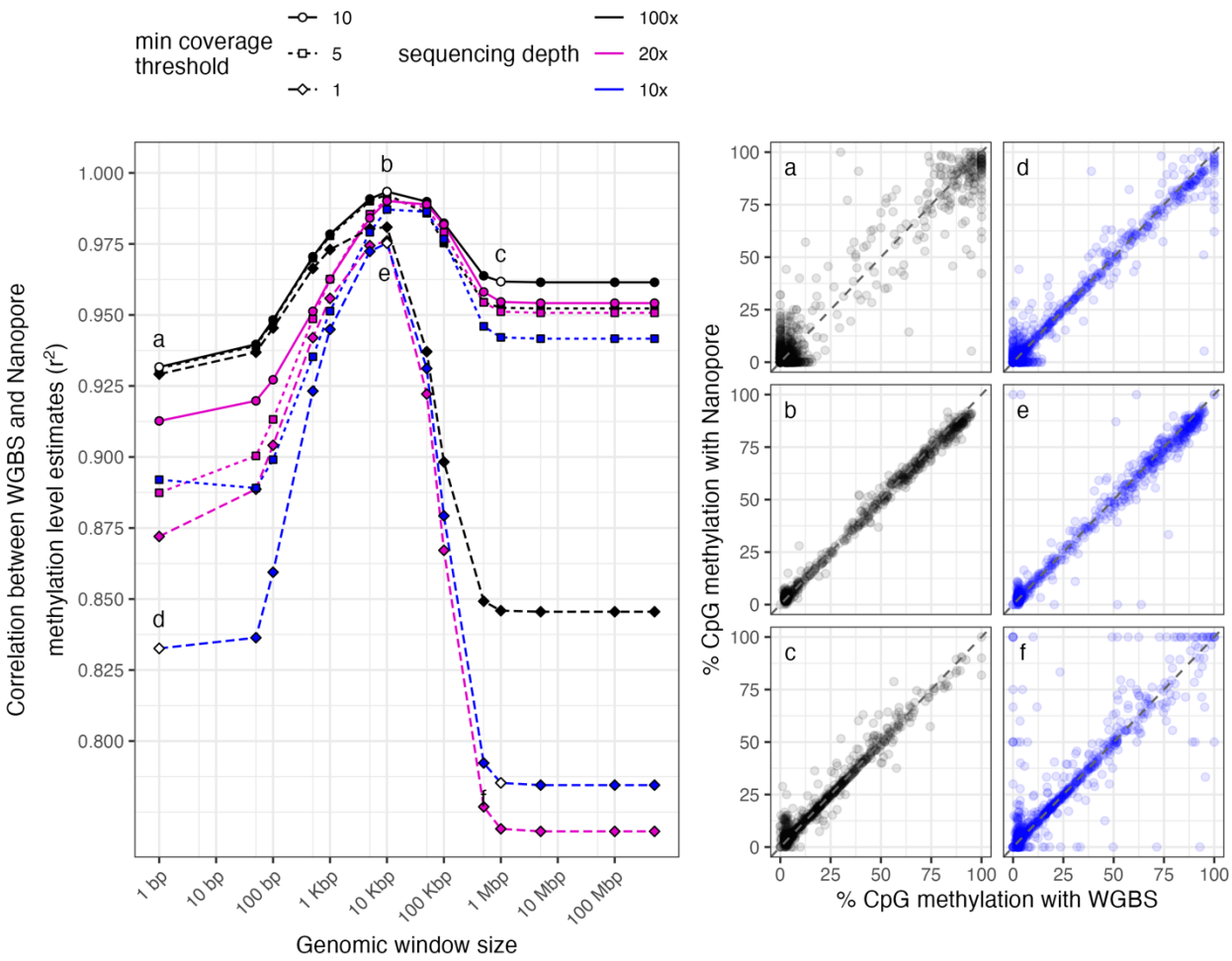

Figure S4.4 Evaluation of Nanopore-derived methylation calls. The correlation between Nanopore and WGBS methylation level estimates is the highest at gene-level scales and is improved by increasing sampling effort and removing CpG sites with low call coverage. Left panel: Squared Pearson correlation coefficient (r2) for Nanopore and WGBS methylation frequencies averaged over increasing genomic windows. Right panel: Mean methylation frequencies of a random subsample of 1000 genomic windows of size 1 bp (**a, d**), 10 Kbp (**b, e**), and 1 Mbp (**c, f**) at 100X (**a-c**) and 10X (**d-f**) sequencing depth; letters in the right panel correspond to these in the left panel.

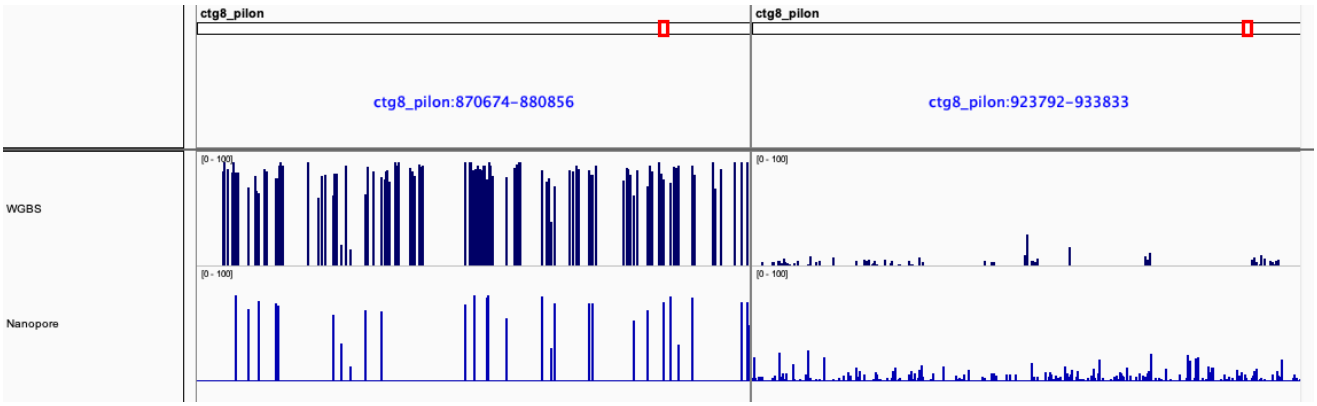

Figure S4.5 WGBS and Nanopore methylation estimates in regions with contrasting methylation. Genome browser snapshot of WGBS and Nanopore-derived methylation levels in *Ridgeia piscesae* (NCBI accession: JAODUO010000008.1) in 10 Kbp regions of high (left) and low (right) methylation. Methylation estimates supported by less than 10 reads are not represented (minimum methylation call coverage threshold = 10). Note that compared to WGBS, Nanopore underestimates methylation in highly methylated region and underestimates it in lowly methylated regions.

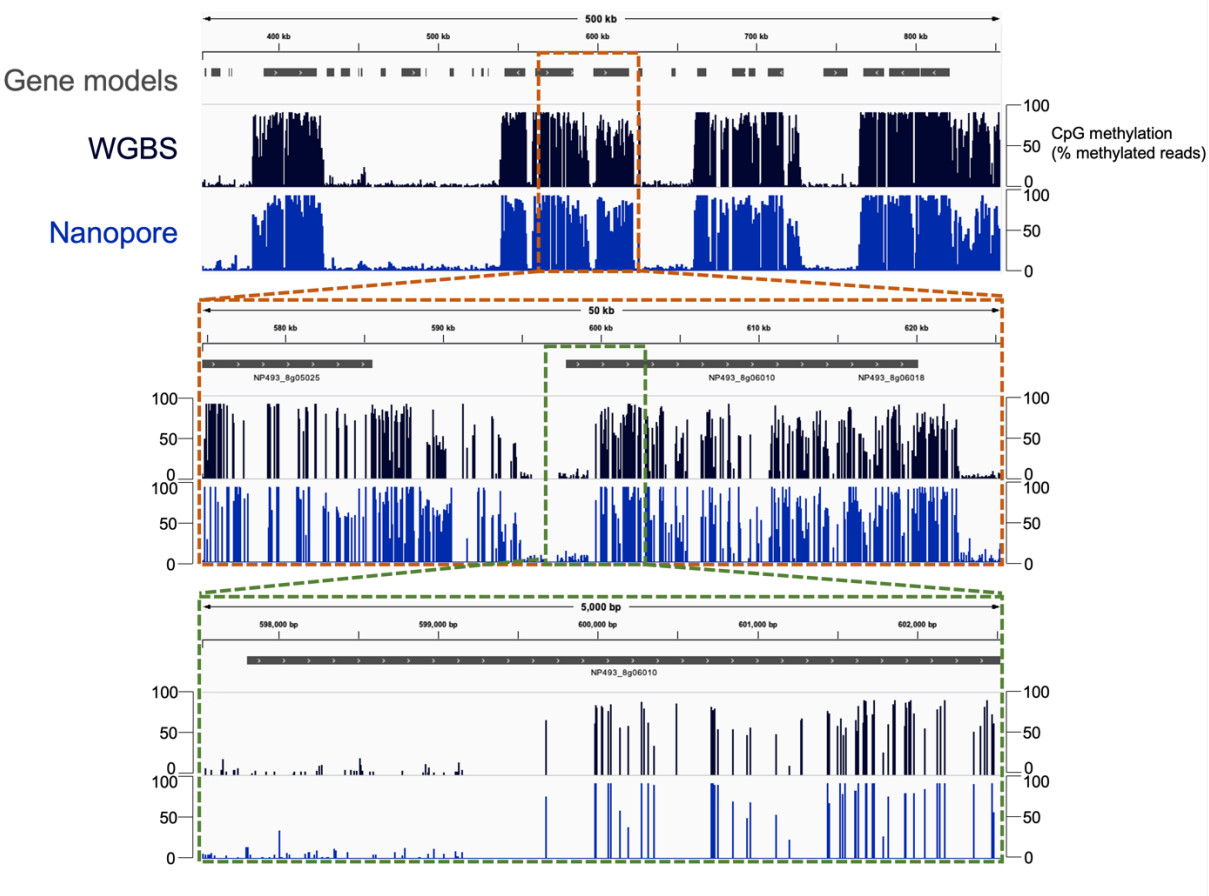

Figure S4.6 WGBS and Nanopore methylation estimates at different genomic scales. Genome browser snapshot of WGBS and Nanopore-derived methylation levels in *Ridgeia piscesae* (NCBI accession: JAODUO010000008.1) at 500Kbp (top), 50Kbp (middle) and 5Kbp (bottom). Methylation estimates supported by less than 10 reads are not represented (minimum methylation call coverage threshold = 10).

Table S4.2 Contingency table for methylated *R. piscesae*'s genes detected with WGBS and Nanopore using a low detection threshold (gene mean methylation frequency > 0). Total gene models = 31703; total gene models with methylation estimates = 29549. Per-CpG minimum methylation call coverage = 10.

	WGBS		
	FALSE	TRUE	NA
Nanopore			
FALSE	19	74	52
TRUE	260	29196	856
NA	25	502	719
Sensitivity	*0.9975*		
Specificity	*0.0681*		
Accuracy	*0.9887*		

Table S4.3 Contingency table for methylated *R. piscesae*'s genes detected with WGBS and Nanopore using a high detection threshold (gene mean methylation frequency > 0.5). Total gene models = 31703; total gene models with methylation estimates = 29549. Per-CpG minimum methylation call coverage = 10.

	WGBS		
	FALSE	TRUE	NA
Nanopore			
FALSE	17341	177	766
TRUE	271	11760	142
NA	270	257	719
Sensitivity	*0.9852*		
Specificity	*0.9846*		
Accuracy	*0.9848*		

To sum up, our results showed that this technology was a powerful tool to uncover invertebrates methylomics landscapes. As long as they are averaged over 1 to 100Kbp genomic windows, Nanopore-derived 5mCpG estimates are highly reliable even at the low sequencing depth typical of genome assembly projects. Only a handful of genome-wide invertebrate methylomes are published on NCBI's Gene Expression Omnibus (GEO) database but reference genomes for over 250 invertebrate species published on NCBI as of September 2022 were generated by using Nanopore sequences. Methylomes can also be recovered from whole-genome shotgun PacBio data which have been used in the genome assembly of more than 700 invertebrate species (Figure S4.7). We argue that reinvestigating these available data to characterize DNA methylation would greatly improve our understanding of the roles and evolutionary history of DNA methylation in invertebrates.

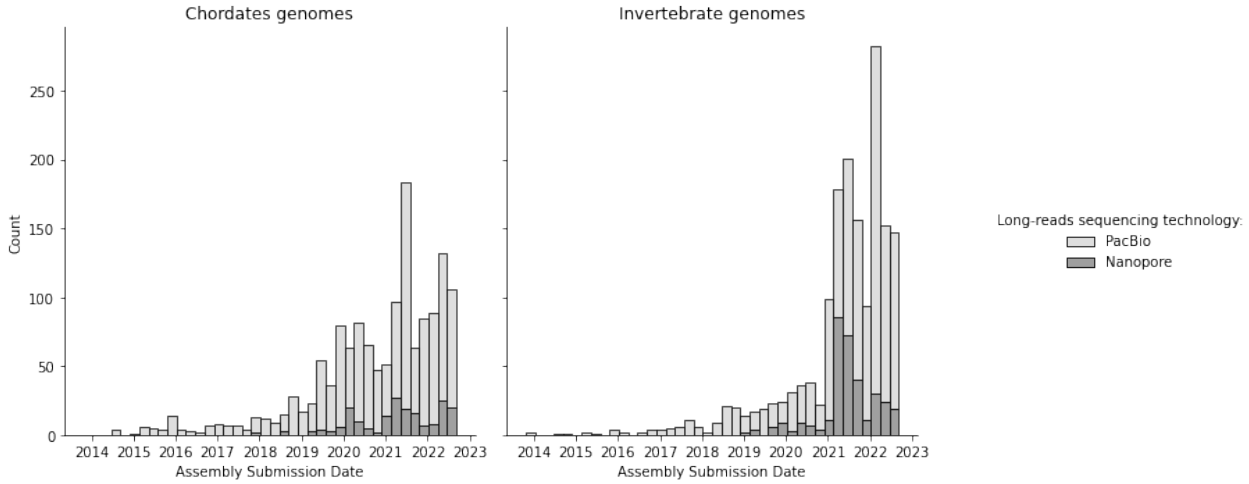

Figure S4.7 Nanopore data are increasingly being used in genome assembly projects. Sequencing platform used for 2928 whole-genome sequencing projects which used third generation sequencing technology. Data were retreived from NCBI's genome database in September 2022.

Genome assembly and polishing

To improve the overall quality of the sequencing libraries prior assembly, raw reads were trimmed or corrected. Illumina reads were trimmed using trimmomatic v0.30 (Bolger *et al.* 2014) and the following parameters (leading = 10, trailing = 10, slidingwindow=4:18, minlen = 50). FMLRC2 (v0.1.4) (Wang *et al.* 2018) was used to correct ONT reads using the trimmed Illumina libraries (Table S4.1).

Genome assemblies for R. piscesae and P. palmiformis were then produced with six different pipelines which can be grouped into three main strategies:

- Short read (Illumina) assembly followed by contig scaffolding using long (ONT) reads,
- Long read assembly followed by sequence polishing using short reads, and
- Hybrid assembly using short and long read libraries.

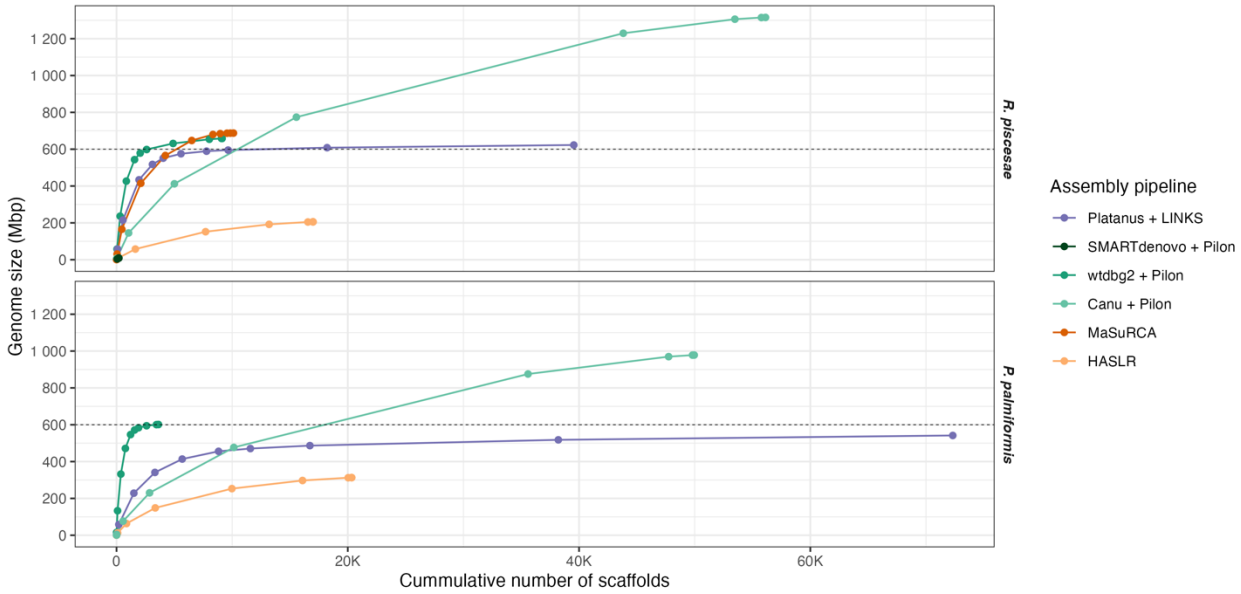

Figure S4.8 Assembly size fraction. The horizontal dotted line represent the species estimated genome sizes.

Table S4.4 and Figure S4.8 Assembly size fraction. The horizontal dotted line represent the species estimated genome sizes.present the results of these assemblies. The best assembly contiguity and completeness were obtained with wtdbg2 (v2.5) (Ruan and Li 2020) followed by one round of sequence polishing with Pilon (v1.23) (Walker *et al.* 2014).

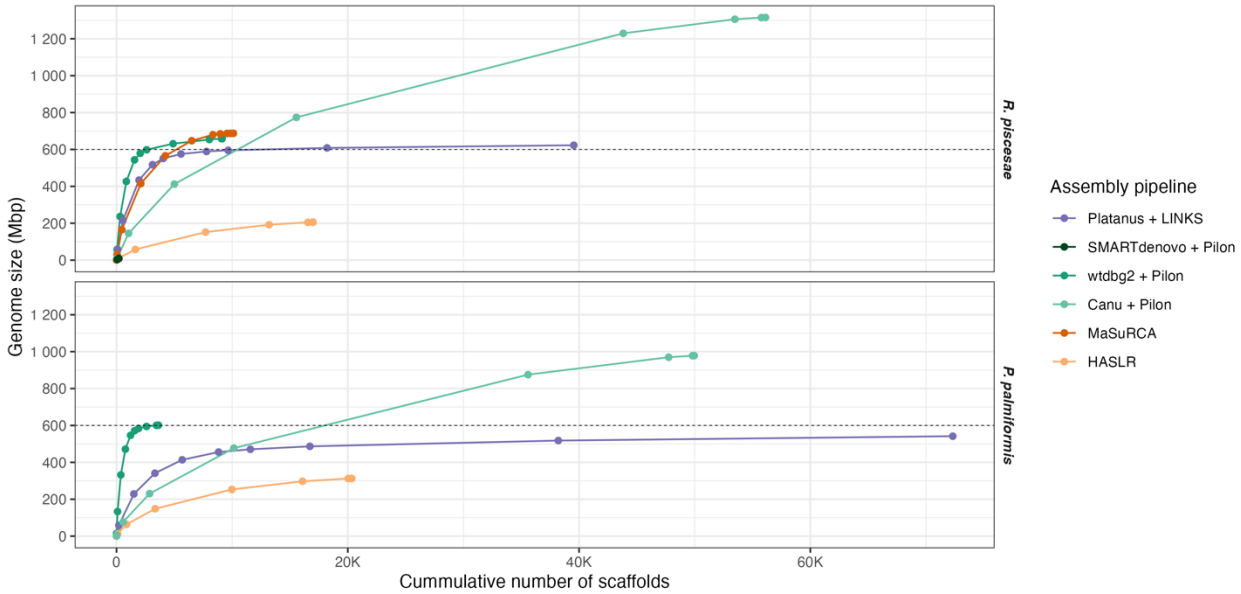

Figure S4.8 Assembly size fraction. The horizontal dotted line represent the species estimated genome sizes.

Table S4.4 Statistics for different assembly pipelines tested in this study. Scaffolds < 500bp were not taken into account.

Assembly pipeline	*Ridgeia piscesae*				*Paralvinella palmiformis*			
	Tot. size (Mbp)	Contig max. len. (Mbp)	# of scaffolds	% genome > 50kbp	Tot. size (Mbp)	Contig max. len. (Mbp)	# of scaffolds	% genome > 50kbp
Short-read + scaffolding								
Platanus + LINKS	1257	1.224	4284603	41%	1145	0.714	3331631	34%
Long-read + polishing								
SMARTdenovo + Pilon	9	0.092	195	43%	n.a.	n.a.	n.a.	n.a.
wtdbg2 + Pilon	657	1.182	9207	88%	601	3.770	3622	95%
Canu + Pilon	1315	0.041	56123	31%	978	0.839	49957	24%
Hybrid								
MaSuRCA	676	1.092	9867	82%	n.a.	n.a.	n.a.	n.a.
HASLR	205	0.103	16993	4%	313	0.550	20338	20%

Genome annotation

Repeat masking

Species-specific repeat libraries were created for *R. piscesae* and *P. palmiformis* with RepeatModeler (v2.0.1) (Tarailo-Graovac and Chen 2009) and merged to the universal repeat databases RepBase27.02 (Bao *et al.* 2015) and dfam3.3 (Storer *et al.* 2021). For *R. piscesae*, the repeat library of its cousin species *P. echinospica* was also added. Using these newly compiled databases, repeats were masked in *R. piscesae* and *P. palmiformis* genomes with RepeatMasker (v4.1.1) (Tarailo-Graovac and Chen 2009). Putative transposable element proteins from a database included in RepeatMasker were also identified in the genomes and masked with RepeatRunner (Smith *et al.* 2007). Intact LTR retrotransposons (i.e. containing at least one protein domain, a pair of each for the target site duplication and long terminal repeats) were identified using the pipeline described in Aroh and Halanych (2021). Briefly, putative LTR elements with and without TGCA motifs were predicted de-novo with LTRharvest genometools (v1.5.10; parameters: -minlentltr 100, -maxlenltr 7000, -mintsd 4, -maxtsd 6, -similar 85, -vic 10, -seed 20, [-motif TGCA, -motifmis 1]) (Ellinghaus, Kurtz, and Willhoeft 2008) and LTR_Finder (v1.07;
parameters= -D 15000, -d 1000, -l 100, -L 7000, -p 20, -C, -M 0.85) (Xu and Wang 2007). LTR_retriever (v2.8.5) was used to filter false positive, estimate insertion times (neutral mutation rate (μ) = 1.3×10^{-8}), and annotate the LTRs. Additional annotations were made by Hidden Markov Model (HMM) profile searches against the REXDB-Metazoan database (Neumann *et al.* 2019) with TE_sorter (Zhang *et al.* 2022).

Transcriptome assemblies

De novo and genome guided reference transcriptome assemblies were constructed with Trinity (v2.11.0) (Grabherr *et al.* 2011) for *R. piscesae* and *P. palmiformis* using the nuclear bins of RNA-Seq libraries of four tissues (gills, vestimentum, epiderm, trophosome) and the whole body, respectively (Figure S4.9). Contigs with no expression evidence (TPM=0) as quantified with Salmon (v1.3.0) (Patro *et al.* 2017) were filtered out. To produce normalized transcriptomes the assemblies were further filtered to keep only the longest isoform of each transcript and clustered according to a 95% identity criteria with CD-HIT (v4.8.1) (Fu *et al.* 2012). The normalized transcriptomes were ultimately translated into amino acid sequences with transdecoder (v5.5.0; https://github.com/TransDecoder/TransDecoder). De novo transcriptomes were also reconstructed from two RNA-Seq libraries of *Paralvinella grasslei* (Stiller *et al.* 2020): accessions SRR3665377 (BioSample SAMN05223276) and SRR3665378 (Biosample SAMN05223277).

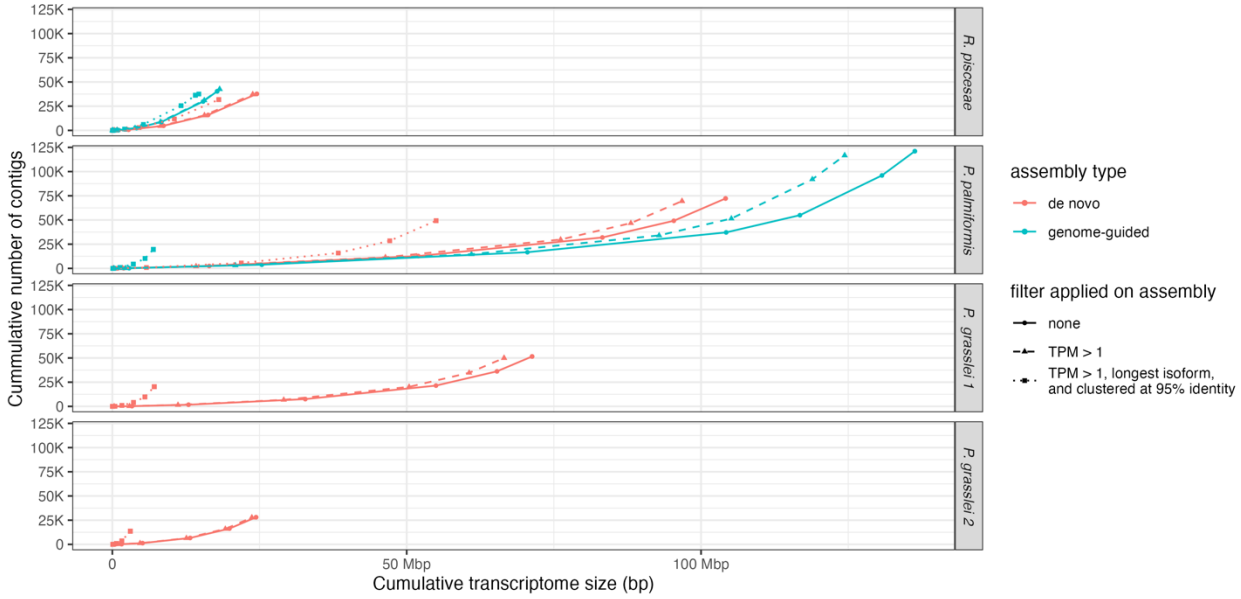

Figure S4.9 Transcriptomes assembled in this study.

Training of Augustus ab-initio gene predictor

Augustus (v 3.4.0) (Hoff and Stanke 2018) was used for ab-initio gene prediction. We used Braker (Hoff *et al.* 2019) with the -etp mode (Augustus is trained with transcriptomic and proteomic evidence) to produce species-specific trainings of the Augustus algorithm for *R. piscesae* and *P. palmiformis*. The raw RNA-Seq data, *de novo* and genome guided transcriptome assemblies were provided as transcriptomic evidence for training. For *Ridgeia*, we used additional data from two publicly available Roche 454 sequencing libraries of *R. piscesae* cDNA (Biosample accessions SAMN00726688 and SAMN00153321) because contamination of our RNA-Seq libraries by ribosomal RNA led to low sequencing depth of nucleus-derived RNA and thus poor transcriptome completeness (busco score <30%, see Table S4.6). For the proteomic evidence, we used protein sequences of closely-related species that best matched the R. *piscesae* and P. palmiformis genomes (>80% coverage and 85% identity as identified with exonerate [v2.4.0] [Slater and Birney 2005] and blastp [v2.12.0] [Camacho *et al.* 2009] through Maker2 [Holt and Yandell 2011]).

Final gene models

The final annotations for the two genomes were obtained with maker2 (Holt and Yandell 2011) using gene models from 1) trained Augustus ab-initio predictors 2) the species respective transcriptomes, and 3) the publicly available translated transcriptomes and proteomes of closely-related species: *Riftia pachyptila* (de Oliveira *et al.* 2022; Hinzke *et al.* 2019; Li *et al.* 2017) Parascarpia echinospica (Sun *et al.* 2021), *Lamellibrachia luymesi* (Li *et al.* 2017, 2019), and *Escarpia spicata* (Li *et al.* 2017) for *R. piscesae*; *Alvinella pompejana* (Gagnière *et al.* 2010), *Paralvinella hessleri* (Hao *et al.* unpublished] and *Paralvinella grasslei* (Stiller *et al.* 2020) for *P. palmiformis*. Only transcripts and proteins which aligned to the reference with a coverage > 80 % and percent identity > 85 % were used for Augustus training. Direct gene model inference used more relaxed mapping criteria: coverage > 50, percent identity > 40. Gene models poorly supported by transcriptomic or proteomic evidence (eAED=1) were removed unless their best hit against NCBI's non-redundant database matched an Annelid sequence [see Functional annotations]. Rescued gene models represented 1.8 % of the 1258 and 0.9 % of the 689 poorly-supported gene models (eAED=1) in *Ridgeia* and *Paralvinella*, respectively. Ultimately, 31703 and 24682 gene models were identified for *R. piscesae* and *P. palmiformis*, respectively. More than 85% of these

gene models had eAED < 0.5 indicating high support from transcriptomic and proteomic evidence (Figure S4.10).

Table S4.5 Transcriptomic and proteomic evidence used for genome annotation.

	Species	Tot. sequences	Busco score[1]	Reference	Biosample (Bioproject)
R07B-5 genome annotation					
transcriptome[a]	*Ridgeia piscesae*	37120	C:23.5%[S:17.3%,D:6.2%],F:18.7%	this study	SAMN26521182
transcriptome[b]	*Ridgeia piscesae*	42586	C:14.2%[S:13.4%,D:0.8%],F:23.9%	this study	SAMN26521182
transcriptome[c]	*Ridgeia piscesae*	515	C:1.7%[S:1.7%,D:0.0%],F:0.5%	Xu 2007 (unpublished)	SAMN00153321
proteome[ad]	*Ridgeia piscesae*	33069	C:36.8%[S:32.6%,D:4.2%],F:24.4%	Li et al 2017[*]	SAMN00726688
proteome[ad]	*Lamellibrachia luymesi*	40745	C:94.0%[S:91.5%,D:2.5%],F:4.5	Li et al 2017,2019[*]	SAMN05013708 (PRJNA322163)
proteome[ad]	*Escarpia spicata*	5836	C:25.9%[S:21.9%,D:4.0%],F:15.6%	Li et al 2017[*]	SAMN05013691 (PRJNA322163)
proteome[ad]	*Riftia pachyptila*	36261	C:47.6%[S:42.9%,D:4.7%],F:27.6%	Li et al 2017[*]	SAMN00726684, SAMN00726685
proteome[ad]	*Riftia pachyptila*	67092	C:96.8%[S:87.4%,D:9.4%],F:1.5%	Hinzke et al 2019[*]	
proteome[bd]	*Riftia pachyptila*	40165	C:94.2%[S:60.8%,D:33.4%],F:3.8%	de Oliveira et al 2021[†]	SRS9777048-SRS9777056 (PRJNA754493)
proteome[e]	*Parascarpia echinospica*	22641	C:81.4%[S:79.4%,D:2.0%],F:7.0%	Sun et al 2021[‡]	(PRJNA472657)
P08H-3 genome annotation					
transcriptome[a]	*Paralvinella palmiformis*	69527	C:95.8%[S:57.9%,D:37.9%],F:2.8%	Stiller et al 2020[§]	SAMN14351933 (PRJNA611902)
transcriptome[b]	*Paralvinella palmiformis*	116679	C:94.2%[S:52.6%,D:41.6%],F:3.1%	Stiller et al 2020[§]	SAMN14351933 (PRJNA611902)
proteome[bdf]	*Paralvinella palmiformis*	19602	C:87.0%[S:84.4%,D:2.6%],F:3.8%	Stiller et al 2020[§]	SAMN14351933 (PRJNA611902)
proteome[cd]	*Alvinella pompejana*	218454	C:53.9%[S:12.7%,D:41.2%],F:26.4%	Ganiere et al 2010	SAMN00169595, SAMN00167295
proteome[ad]	*Paralvinella hessleri*	22289	C:91.2%[S:81.2%,D:10.0%],F:5.3%	Hao et al (unpublished)[*]	
proteome[adf]	*Paralvinella grasslei*	20331	C:76.4%[S:75.4%,D:1.0%],F:12.5%	Stiller et al 2020[§]	SAMN05223276 (PRJNA325100)
proteome[adf]	*Paralvinella grasslei*	13657	C:35.0%[S:34.8%,D:0.2%],F:21.6%	Stiller et al 2020[§]	SAMN05223277 (PRJNA325100)

[1] busco v5.2.2 metazoa_odb10 database, C: Complete, S: Single-copy, D: Duplicated, F: Fragmented
[a] RNA-Seq library: de novo assembly, [b] RNA-Seq library: genome guided assembly, [c] cDNA library, [d] translated transcriptome,
[e] translated gene models, [f] isoforms removed [*] assembly provided by author via personnal communication [†] assembly available at https://phaidra.univie.ac.at/o:1220863
[‡] assembly available at https://figshare.com/articles/online_resource/Genome_assembly_and_annotation/15050478 [§] re-assembeled in this study

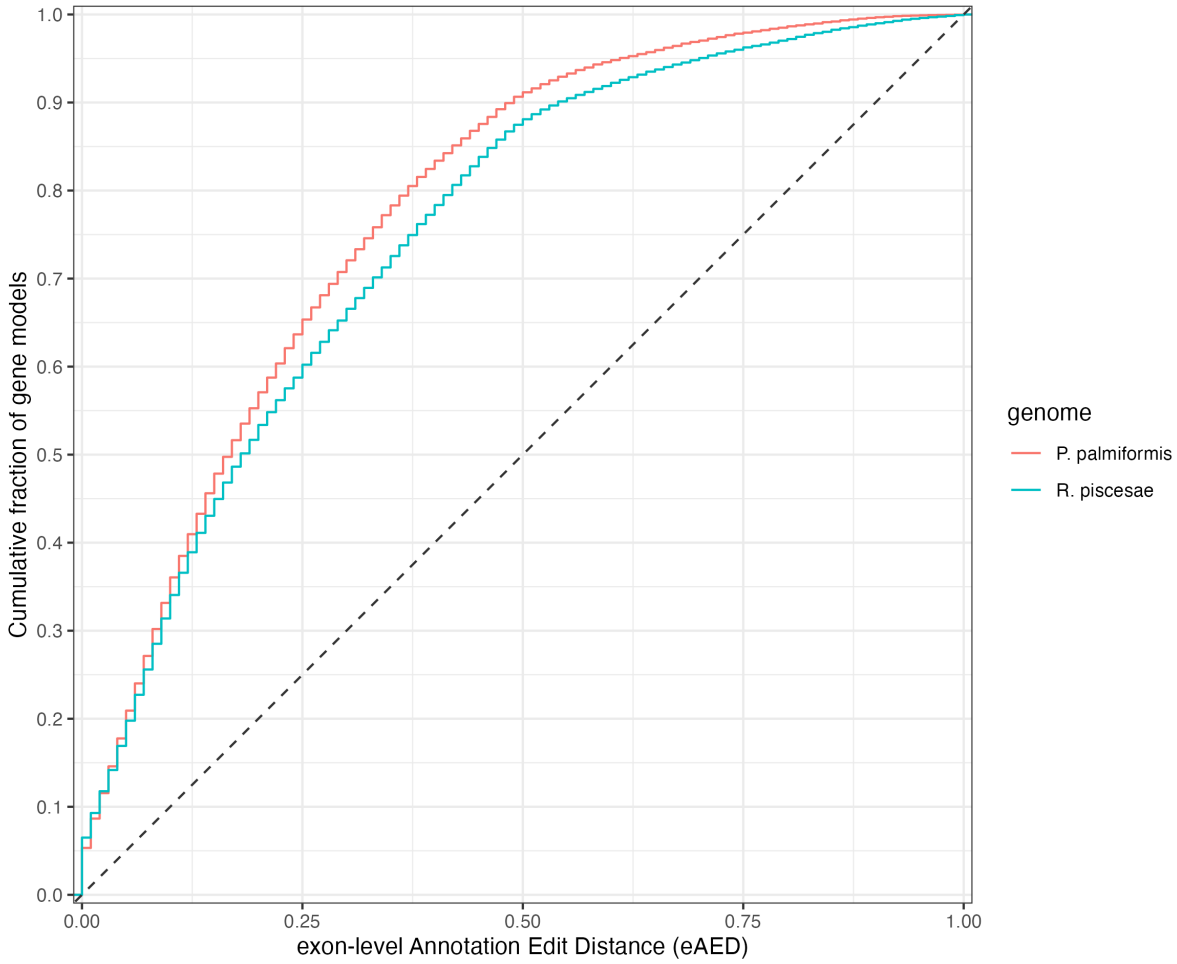

Figure S4.10 Concordance between final gene models and annotation evidences.

Table S4.6 Final gene models BUSCO searches against the metazoan database (metazoa_odb10).

	P. echinospica		*R. piscesae*		*P. palmiformis*	
	count	%	count	%	count	%
Complete BUSCOs (C)	776	81%	838	88%	842	88%
Complete and single-copy BUSCOs (S)	757	79%	821	86%	833	87%
Complete and duplicated BUSCOs (D)	19	2%	17	2%	9	1%
Fragmented BUSCOs (F)	67	7%	62	6%	72	8%
Missing BUSCOs (M)	111	12%	54	6%	40	4%
Total BUSCO groups searched	954	100%	954	100%	954	100%

Functional annotations

The proteins issued from the final gene models were blasted (blast+ v12.12.0) against NCBI non-redundant (nr) protein database (v2020-04-22; max_evalue = 1E-6, max_hsps = 5) and scanned for known functional domains with InterProScan (v5.55-88.0, default parameters) (Zdobnov and

Apweiler 2001). We used Blast2Go (Conesa *et al.* 2005) to merge the results of the two analyses and obtain GO annotations (go database v2022-03-01, default parameters). The euKaryotic Orthologous Groups (KOG) annotations were obtained by functional domain search with rpsblast (v2.2.15, evalue cutoff = 0.00001) against NCBI KOG database (v02-02-2011) through the annotation server webMGA (Wu *et al.* 2011). The webMGA server was also used to screen our gene models against TIGRFAM (v10.0) (Haft *et al.* 2003) and Pfam (v24.0) (Finn *et al.* 2010) databases with hmmscan (v3.0, evalue cutoff = 0.00001). We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (method = blast bidirectional hit; reference GENES data set included all available invertebrate genomes, *Homo sapiens*, *Mus muscus*, *Rattus norvegicus* and *Danio rerio*[2]). Figure S4.11 shows a summary of these functional annotations. Finally, the gene models of *R. piscesae* were further searched against these of *R. pachyptila* and *P. echinospica* with blastp (Figure S4.12).

[2] genome references: hro, lgi, pcan, crg, myi, obi, pvm, isc, tut, dpte, cscu, ptep, cel, cbr, bmy, loa, nai, tsp, smm, lak, nve, epa, adf, amil, pdam, spis, dgt, hmg, tad, aqu, dme, sko, aplc, spu, cin, bfo, dre, hsa, mmu, rno

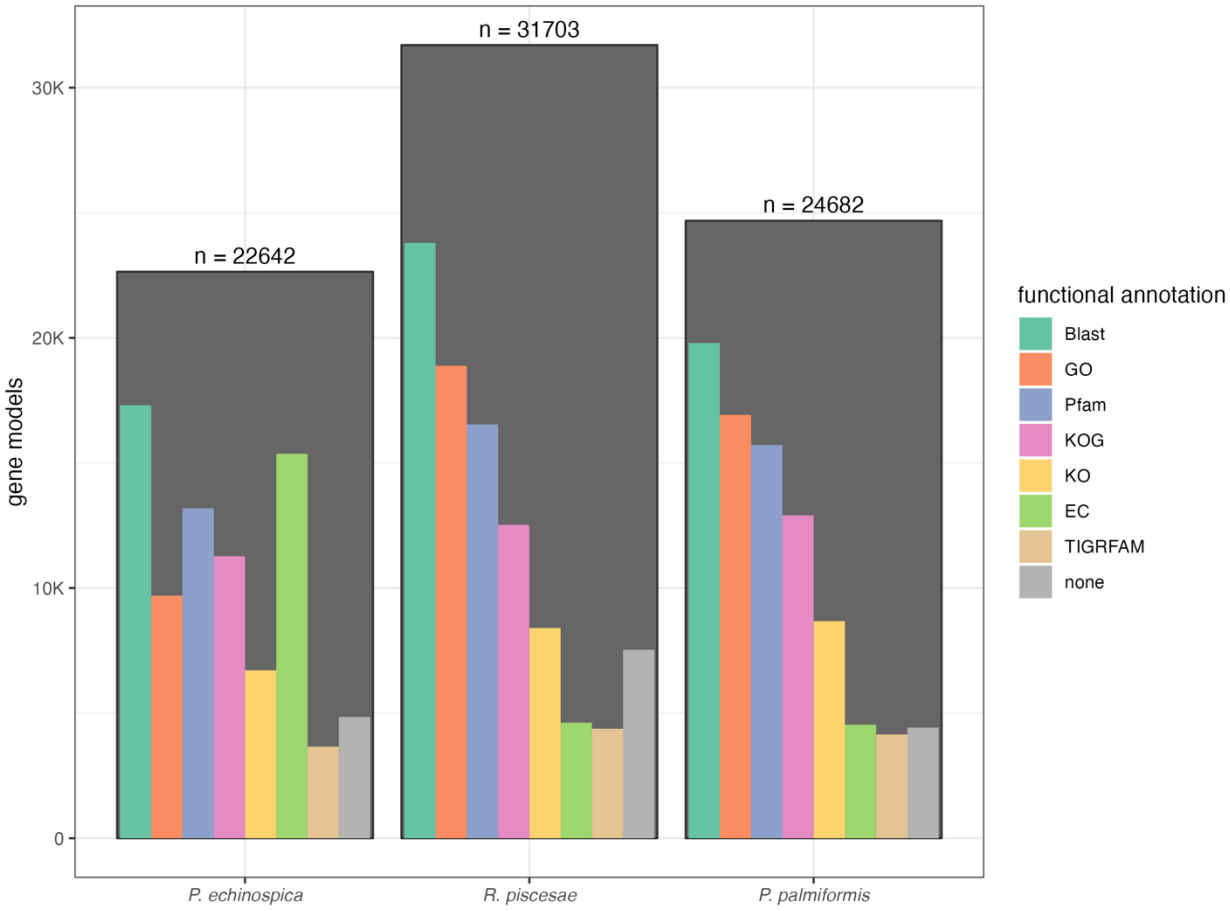

Figure S4.11 Summary of functional annotations. Blast: non-redundant nucleotide collection; GO: the gene ontology knowledgebase; Pfam: protein families and domains database; KOG: the eukaryotic orthologous groups of proteins database; KO: the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology database; EC: the database of enzyme commission number; TIGRFAM: database of protein family definitions.

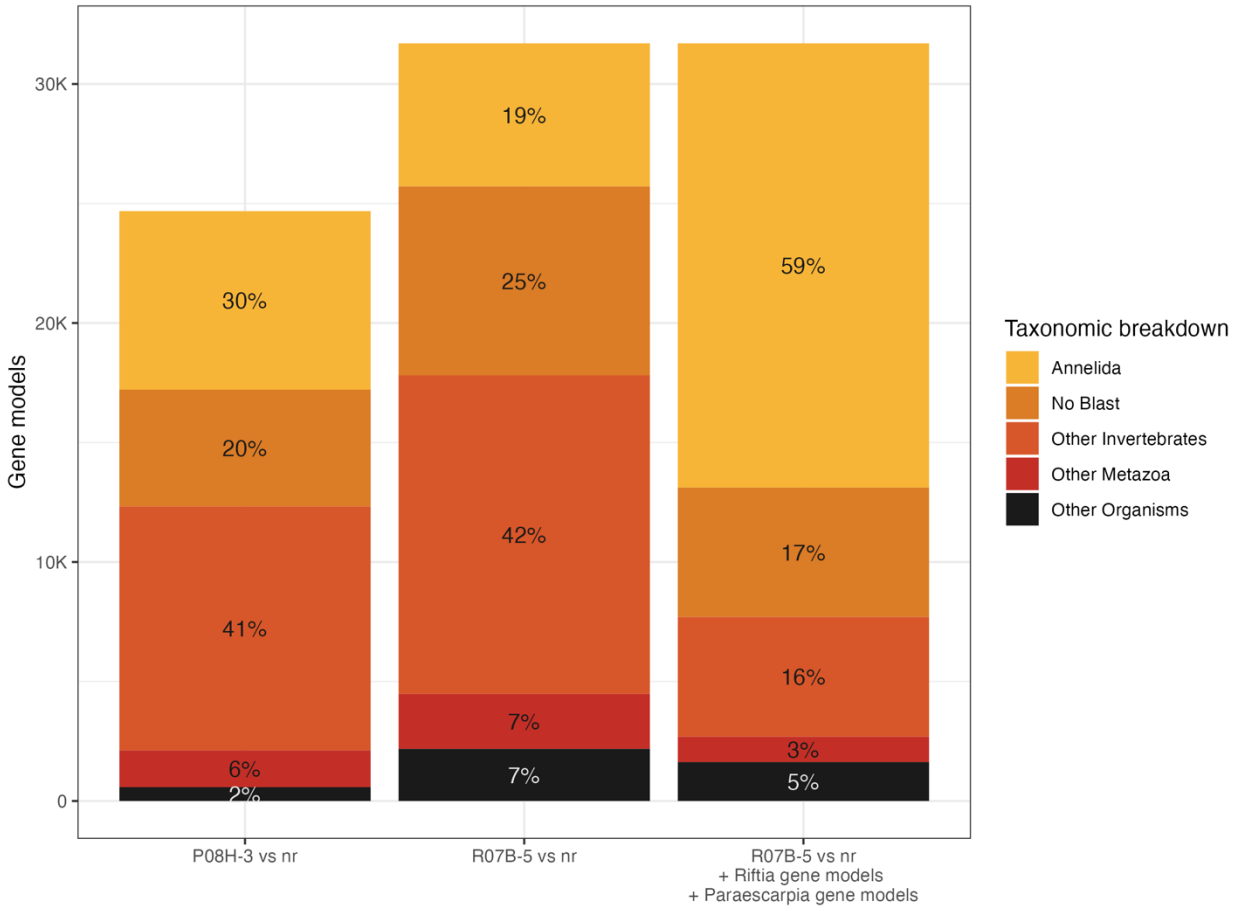

Figure S4.12 Taxonomy of gene models best blastp hits. P08H-3: *P. palmiformis* assembly, R07B-5: *R. piscesae* assembly.

Phylogenetic and selective inferences on the MAT gene

Gene sequences for the MAT protein were recovered from the transcriptomes of polychaete species of the order Terebellida (Stiller *et al.* 2020) and Sabellida (Tilic *et al.* 2020), and the genomes of recently sequenced siboglinidae: *O. alvinae*, *O. frankpressi* (Moggioli *et al.* 2023), and *R. pachyptila* (de Oliveira *et al.* 2022). Additional sequences of outgroup species were used: *D. melanogaster* (Larsson and Rasmuson-Lestander 1994), *P. caudatus* (Martín-Durán and Hejnol 2015), *O. fusiformis* (Moggioli *et al.* 2023), *D. gyrociliatus* (Martín-Durán *et al.* 2021), *H. robusta* (Simakov *et al.* 2013), *P. dumerili* (Schenk *et al.* 2016), *C. teleta* (Simakov *et al.* 2013), and *S. benedicti* (Zakas *et al.* 2022). The general time-reversible (GTR) nucleotide substitution model with four substitution rate classes was used in all phylogenetic analyses. Across-sites rates were set to be gamma-distributed with a proportion of invariable sites. BI was performed in MrBayes

v3.2.7 (Ronquist *et al.* 2012) running 4 chains for 5,000,000 generations with a 100,000 generations burn-in and sampling every 5,000 generations. These parameters were optimized using RWTY v1.0.2 (Warren, Geneva, and Lanfear 2017) to ensure convergence and to avoid autocorrelation in the tree sampling. The ML tree was generated with PhyML (Guindon *et al.* 2010) and the following parameters: –model GTR –nclasses 4 –search SPR –n_rand_starts 10. Branch confidence was estimated with 100 bootstraps. Relaxed selection in Mat-b compared to Mat-a was tested in 1) vestimentiferans, 2) alvinellids and amphraetids, and 3) terebellids with RELAX (Wertheim *et al.* 2015) implemented in Datamonkey (Weaver *et al.* 2018). Episodic diversifying selection was detected with aBSRel (M. D. Smith *et al.* 2015) across all Canalipalpata taxa.

Methylome annotations

Gff files were generated for the genomic features of interest: genes, introns, exons, LTR-retrotransposons, repeats, and uncharacterized intergenic regions. Methylation coverage information, CpG counts, and transcriptomic expression level was added with bedtools (Quinlan and Hall 2010) map. Likewise, bedtools map and bedtools makewindows were used annotate bed files of genes and LTR-retrotranposons 10Kbp upstream and downstream regions (sliding window size=1Kbp, step size=200bp). Genome-wide methylation frequency spectra were obtained by mapping genomic features onto the per-CpG methylation information with bedtools map (Quinlan and Hall 2010). For any feature, methylation coverage was defined as the fraction of called CpGs for which methylation was detected and methylation depth as the fraction of reads for which methylation was detected amongst methylated CpGs. The mean methylation frequency (methylation coverage x methylation depth) of a feature is referred to as methylation level. Finally, highly methylated features were defined as those with methylation coverage > 75%
and methylation depth > 50% as these two thresholds conservatively capture a high methylation density cluster (Figure S4.13).

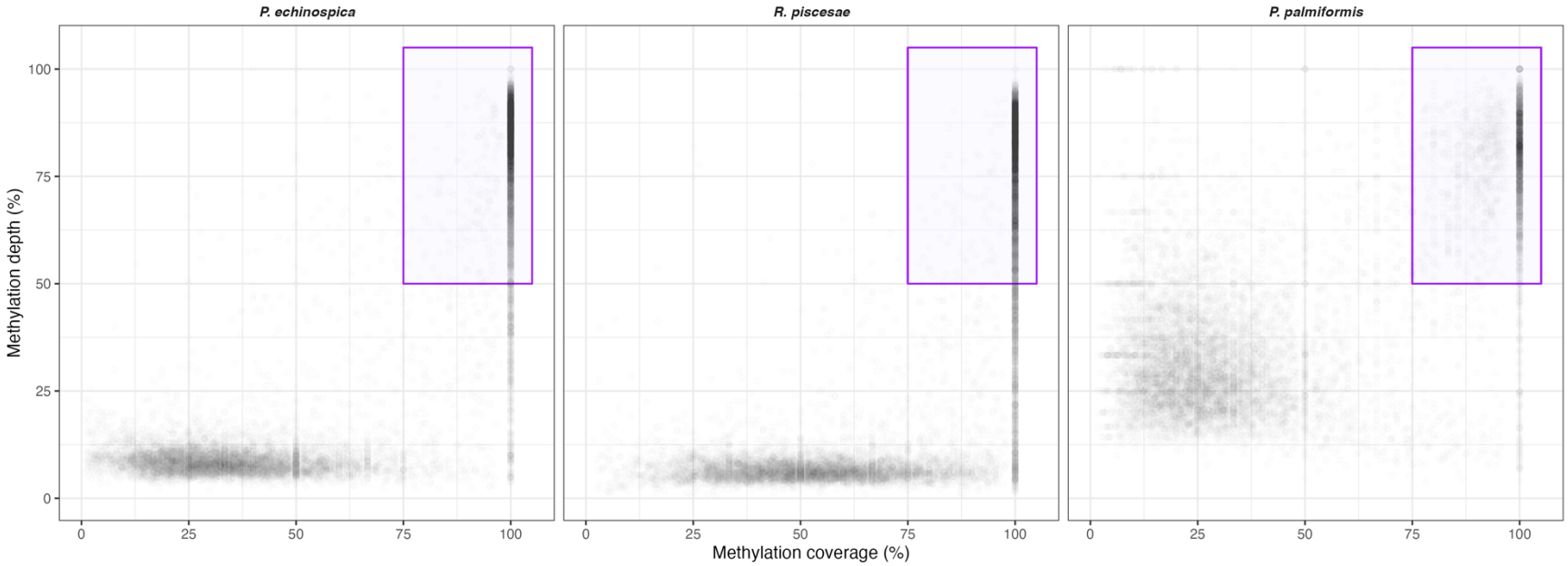

Figure S4.13 Methylation coverage and depth of genome features. Distribution of methylation coverage (fraction of methylated CpGs) and methylation depth (fraction of methylated reads amongst methylated CpGs) for 10,000 randomly selected genomic windows of 1Kbp. CpGs covered by less than 10 reads were only called for methylation in *P. palmiformis*. The purple rectangle highlights genomic regions which are highly methylated (i.e. methylation coverage > 75% and methylation depth > 50%).

References

Aroh, O., and Halanych, K.M. 2021. Genome-wide characterization of LTR retrotransposons in the non-model deep-sea annelid *Lamellibrachia luymesi*. BMC Genomics 22(1): 466. doi:10.1186/s12864-021-07749-1.

Bao, W., Kojima, K.K., and Kohany, O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1): 11. doi:10.1186/s13100-015-0041-9.

Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120. doi:10.1093/bioinformatics/btu170.

Bushnell, B. 2014. BBMap: A fast, accurate, splice-aware aligner. Available from https://escholarship.org/uc/item/1h3515gn [accessed 6 December 2021].

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1): 421. doi:10.1186/1471-2105-10-421.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18): 3674–3676. doi:10.1093/bioinformatics/bti610.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., and Bateman, A. 2010. The Pfam protein families database. Nucleic Acids Res 38(Database issue): D211-222. doi:10.1093/nar/gkp985.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23): 3150–3152. doi:10.1093/bioinformatics/bts565.

Gagnière, N., Jollivet, D., Boutet, I., Brélivet, Y., Busso, D., Da Silva, C., Gaill, F., Higuet, D., Hourdez, S., Knoops, B., Lallier, F., Leize-Wagner, E., Mary, J., Moras, D., Perrodou, E., Rees, J.-F., Segurens, B., Shillito, B., Tanguy, A., Thierry, J.-C., Weissenbach, J., Wincker, P., Zal, F., Poch, O., and Lecompte, O. 2010. Insights into metazoan evolution from *Alvinella pompejana* cDNAs. BMC Genomics 11: 634. doi:10.1186/1471-2164-11-634.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 29(7): 644–652. doi:10.1038/nbt.1883.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59(3): 307–321. doi:10.1093/sysbio/syq010.

Haft, D.H., Selengut, J.D., and White, O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31(1): 371–373. doi:10.1093/nar/gkg128.

Hinzke, T., Kleiner, M., Breusing, C., Felbeck, H., Häsler, R., Sievert, S.M., Schlüter, R., Rosenstiel, P., Reusch, T.B.H., Schweder, T., and Markert, S. 2019. Host-microbe interactions in the chemosynthetic *Riftia pachyptila* symbiosis. MBio 10(6): 20.

Hoff, K., Lomsadze, A., Borodovsky, M., and Stanke, M. 2019. Whole-genome annotation with BRAKER. Methods Mol Biol 1962: 65–95. doi:10.1007/978-1-4939-9173-0_5.

Hoff, K.J., and Stanke, M. 2018. Predicting Genes in Single Genomes with AUGUSTUS. Current Protocols in Bioinformatics: e57. doi:10.1002/cpbi.57.

Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 12(1): 491. doi:10.1186/1471-2105-12-491.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27(5): 722–736. doi:10.1101/gr.215087.116.

Krueger, F., and Andrews, S.R. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27(11): 1571–1572. doi:10.1093/bioinformatics/btr167.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. 2012. DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9(2): 145–151. doi:10.1038/nmeth.1828.

Larsson, J., and Rasmuson-Lestander, A. 1994. Molecular cloning of the S-adenosylmethionine synthetase gene in *Drosophila melanogaster*. FEBS Lett 342(3): 329–333. doi:10.1016/0014-5793(94)80526-1.

Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34(18): 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, Y., Kocot, K.M., Whelan, N.V., Santos, S.R., Waits, D.S., Thornhill, D.J., and Halanych, K.M. 2017. Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. Zool Scr 46(2): 200–213. doi:10.1111/zsc.12201.

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

Martín-Durán, J.M., and Hejnol, A. 2015. The study of *Priapulus caudatus* reveals conserved molecular patterning underlying different gut morphogenesis in the Ecdysozoa. BMC Biol 13: 29. doi:10.1186/s12915-015-0139-z.

Martín-Durán, J.M., Vellutini, B.C., Marlétaz, F., Cetrangolo, V., Cvetesic, N., Thiel, D., Henriet, S., Grau-Bové, X., Carrillo-Baltodano, A.M., Gu, W., Kerbl, A., Marquez, Y., Bekkouche, N., Chourrout, D., Gómez-Skarmeta, J.L., Irimia, M., Lenhard, B., Worsaae, K., and Hejnol, A. 2021. Conservative route to genome compaction in a miniature annelid. Nat Ecol Evol 5(2): 231–242. doi:10.1038/s41559-020-01327-6.

Moggioli, G., Panossian, B., Sun, Y., Thiel, D., Martín-Zamora, F.M., Tran, M., Clifford, A.M., Goffredi, S.K., Rimskaya-Korsakova, N., Jékely, G., Tresguerres, M., Qian, P.-Y., Qiu, J.-W., Rouse, G.W., Henry, L.M., and Martín-Durán, J.M. 2023. Distinct genomic routes underlie transitions to specialised symbiotic lifestyles in deep-sea annelid worms. Nat Commun 14(1): 2814. doi:10.1038/s41467-023-38521-6.

Neumann, P., Novák, P., Hoštáková, N., and Macas, J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA 10(1): 1. doi:10.1186/s13100-018-0144-1.

de Oliveira, A.L., Mitchell, J., Girguis, P., and Bright, M. 2022. Novel insights on obligate symbiont lifestyle and adaptation to chemosynthetic environment as revealed by the giant tubeworm genome. Mol Biol Evol 39(1): msab347. doi:10.1093/molbev/msab347.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. 2017. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 14(4): 417–419. doi:10.1038/nmeth.4197.

Quinlan, A.R., and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6): 841–842. doi:10.1093/bioinformatics/btq033.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61(3): 539–542. doi:10.1093/sysbio/sys029.

Ruan, J., and Li, H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17(2): 155–158. doi:10.1038/s41592-019-0669-3.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J.A., Shapiro, H., Aerts, A., Otillar, R.P., Terry, A.Y., Boore, J.L., Grigoriev, I.V., Lindberg, D.R., Seaver, E.C., Weisblat, D.A., Putnam, N.H., and Rokhsar, D.S. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493(7433): 526–531. doi:10.1038/nature11696.

Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods 14(4): 407–410. doi:10.1038/nmeth.4184.

Slater, G.S.C., and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6(1): 31. doi:10.1186/1471-2105-6-31.

Smith, C.D., Edgar, R.C., Yandell, M.D., Smith, D.R., Celniker, S.E., Myers, E.W., and Karpen, G.H. 2007. Improved repeat identification and masking in Dipterans. Gene 389(1): 1–9. doi:10.1016/j.gene.2006.09.011.

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol 32(5): 1342–1353. doi:10.1093/molbev/msv022.

Stiller, J., Tilic, E., Rousset, V., Pleijel, F., and Rouse, G.W. 2020. Spaghetti to a tree: a robust phylogeny for Terebelliformia (Annelida) based on transcriptomes, molecular and morphological data. Biology 9(4): 73. doi:10.3390/biology9040073.

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12(1): 2. doi:10.1186/s13100-020-00230-y.

Tarailo-Graovac, M., and Chen, N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics 25(1): 4.10.1-4.10.14. doi: 10.1002/0471250953.bi0410s25.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31(12): 2032–2034. doi:10.1093/bioinformatics/btv098.

Tilic, E., Sayyari, E., Stiller, J., Mirarab, S., and Rouse, G.W. 2020. More is needed—Thousands of loci are required to elucidate the relationships of the ‘flowers of the sea’ (Sabellida, Annelida). Mol Biol Evol 151: 106892. doi:10.1016/j.ympev.2020.106892.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE 9(11): e112963. Public Library of Science. doi:10.1371/journal.pone.0112963.

Wang, J.R., Holt, J., McMillan, L., and Jones, C.D. 2018. FMLRC: Hybrid long read error correction using an FM-index. BMC Bioinformatics 19(1): 50. doi:10.1186/s12859-018-2051-3.

Warren, D.L., Geneva, A.J., and Lanfear, R. 2017. RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. Mol Biol Evol: msw279. doi:10.1093/molbev/msw279.

Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., and Kosakovsky Pond, S.L. 2018. Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol 35(3): 773–777. doi:10.1093/molbev/msx335.

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., and Scheffler, K. 2015. RELAX: Detecting relaxed selection in a phylogenetic framework. Mol Biol Evol 32(3): 820–832. doi:10.1093/molbev/msu400.

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. 2011. WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics 12: 444. doi:10.1186/1471-2164-12-444.

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

Zakas, C., Harry, N.D., Scholl, E.H., and Rockman, M.V. 2022. The genome of the poecilogonous annelid *Streblospio benedicti*. Genome Biology and Evolution 14(2): evac008. doi:10.1093/gbe/evac008.

Zdobnov, E.M., and Apweiler, R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9): 847–848. doi:10.1093/bioinformatics/17.9.847.

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., and Ma, Y. 2022. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Horticulture Research 9: uhac017. doi:10.1093/hr/uhac017.

Supplementary Figures

Figures S4.1 to S4.13 are presented within the Supplementary Methods.

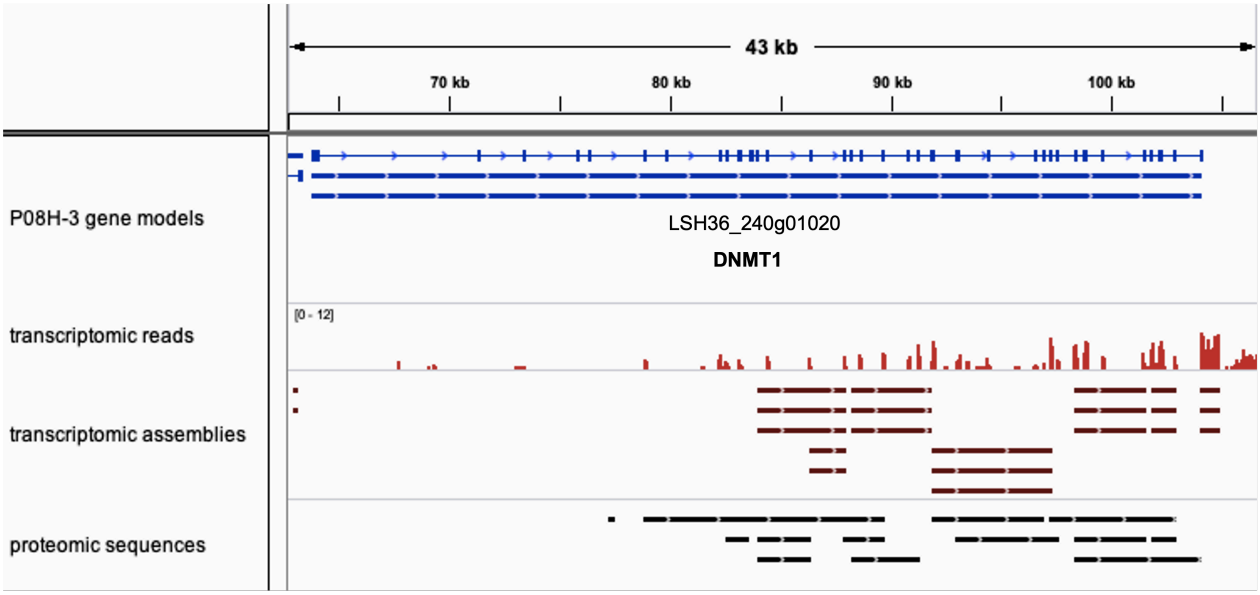

Figure S4.14 Gene model and mapped expression evidence for the DNMT1 gene in *P. palmiformis*. Raw transcriptomic reads from *P. palmiformis* (Stiller *et al.* 2020); de novo and genome-guided transcriptomic sequences from *P. palmiformis* (Stiller *et al.* 2020); proteomic sequences from *P. palmiformis* (Stiller *et al.* 2020), *P. grasslei* (Stiller *et al.* 2020), and *P. hessleri* (Wang *et al.* unpublished). Note the low depth of coverage (max=12) of the transcriptomic reads.

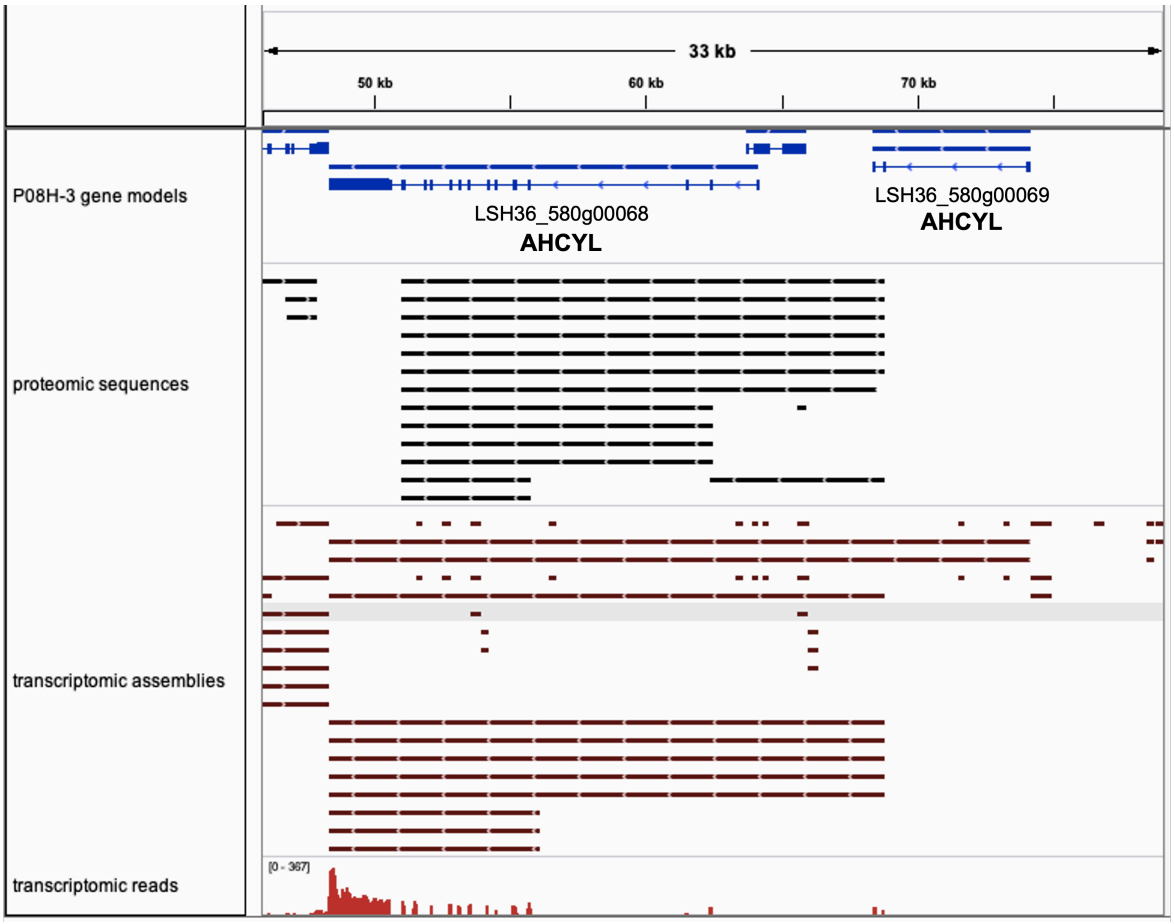

Figure S4.15 Gene model and mapped expression evidence for the AHCYL gene in *P. palmiformis*. Raw transcriptomic reads from *P. palmiformis* (Stiller *et al.* 2020); de novo and genome-guided transcriptomic sequences from *P. palmiformis* (Stiller *et al.* 2020); proteomic sequences from *P. palmiformis* (Stiller *et al.* 2020), *P. grasslei* (Stiller *et al.* 2020), and *P. hessleri* (Wang *et al.* unpublished). Note that despite transcriptomic and proteomic evidence mapping across both gene models MAKER failed to merge the two gene models in our reannotation attempts.

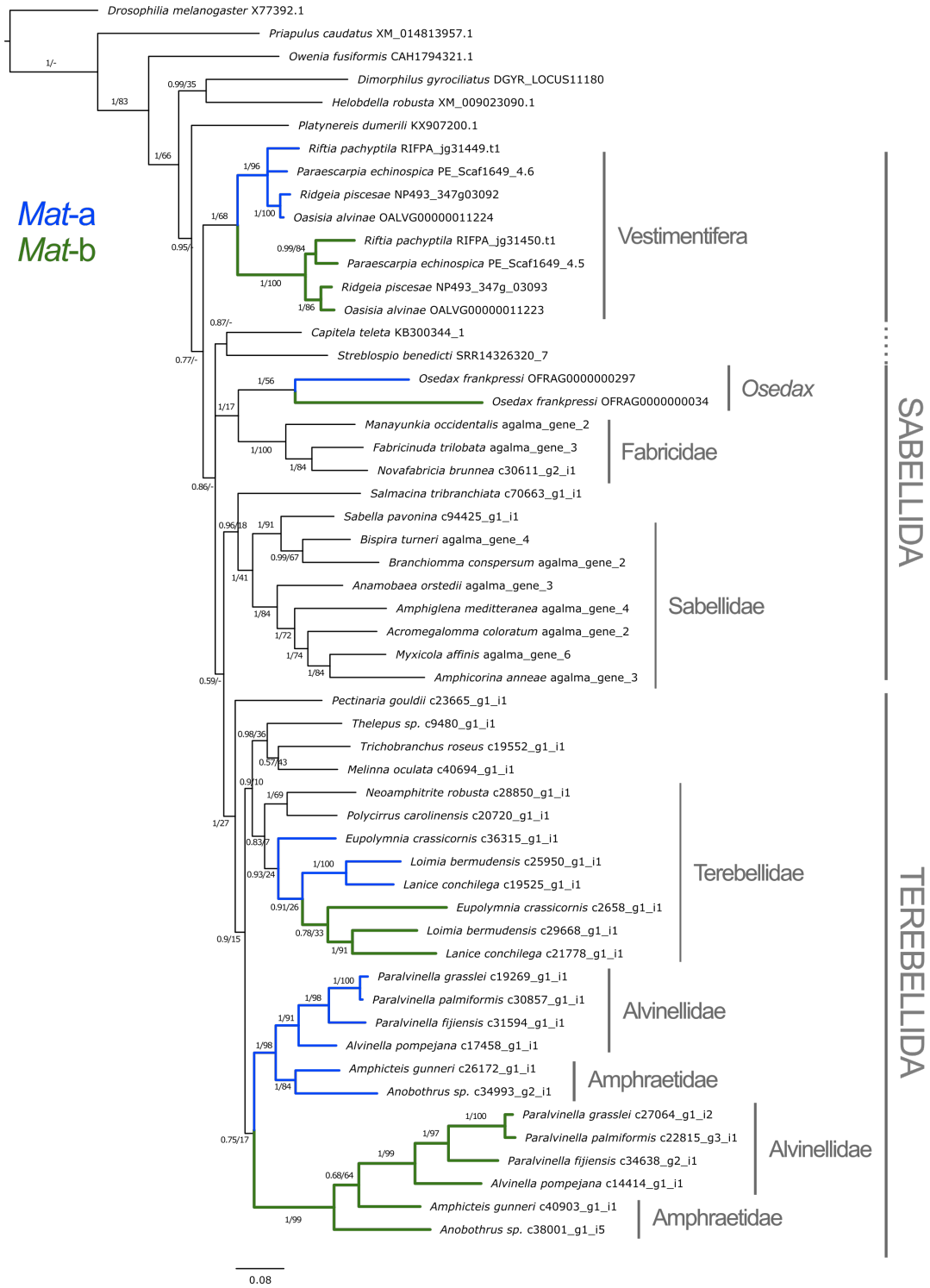

Figure S4.16 Bayesian inference tree for the methionine adenosyltransferases (MAT) in polychaetes. Consensus tree built from trimmed codon alignments (n = 1140bp). Numbers above branches are the posterior probabilities (Bayesian inference) and bootstrap values (maximum likelihood inference). Hyphens indicate nodes not present in the maximum likelihood tree. Leaves are labelled with the species names followed by their unique sequence identifiers.

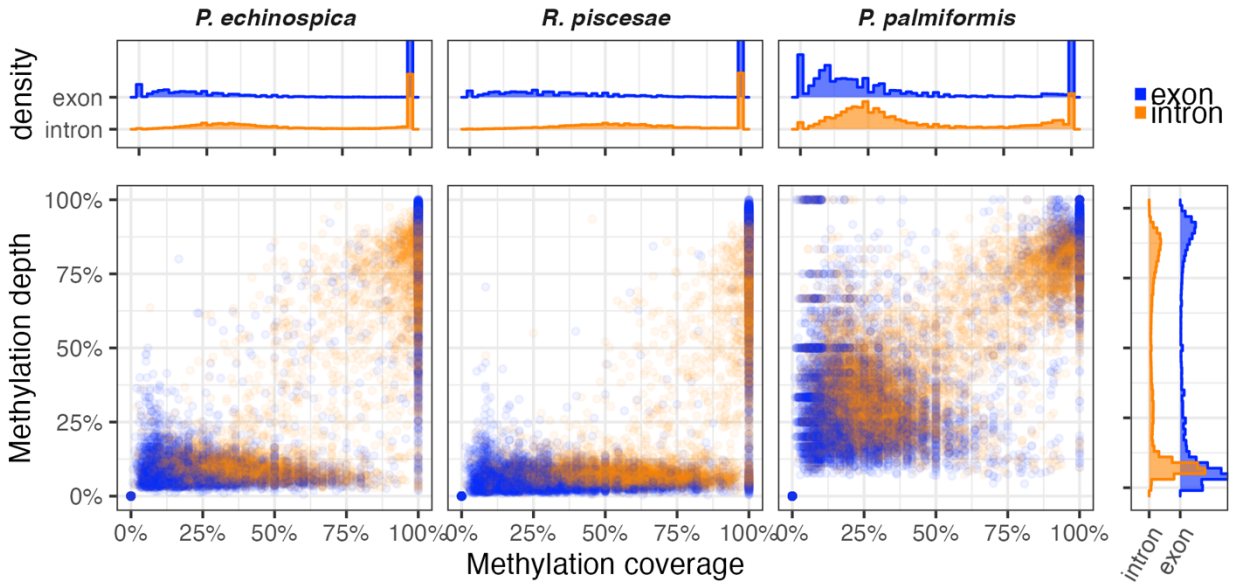

Figure S4.17 More methylation variance is found in exons than in introns. Methylation coverage (fraction of methylated CpGs) and methylation depth (fraction of methylated reads amongst methylated CpGs) in introns and exons. Only introns and exons pairs with each 10 or more CpGs called for methylation were considered (n=27692, 26310, and 11357 for *P. echinospica*, *R. piscesae*, and *P. palmiformis*, respectively). CpGs covered by less than 10 reads were only called for methylation in P. palmiformis. For better visibility, the scatter plots only represent a random subsample of 10000 exon-intron pairs for each species features whereas the marginal densities include all data. Note that methylation coverage and depth variances are greater in exons than in introns (permutations tests with 500 permutations).

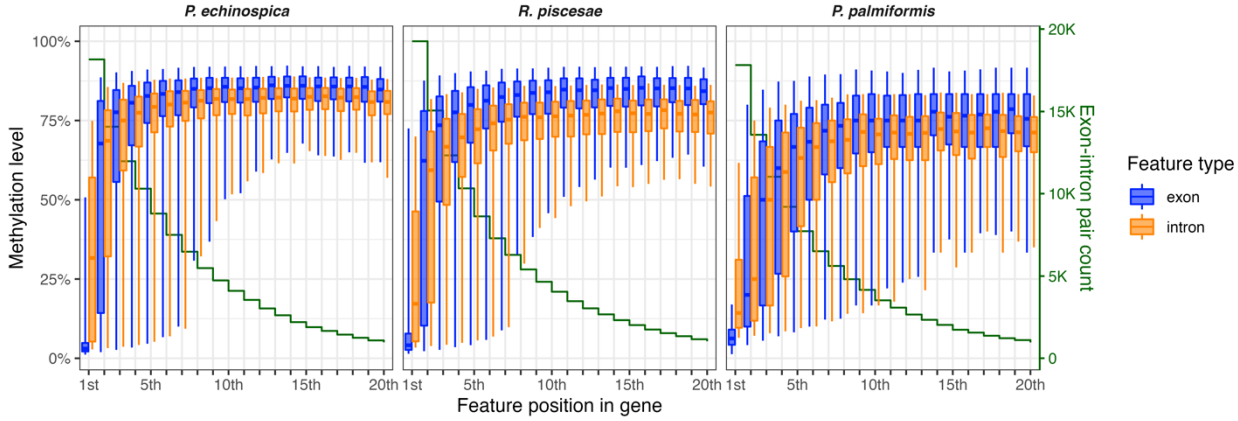

Figure S4.18 Methylation is lower at the beginning of genes and in introns. Distribution of methylation levels along the first 20 exon-introns pairs of genes. The box represent the 20 inter-percentile and the whiskers extend to 50 inter-percentile.

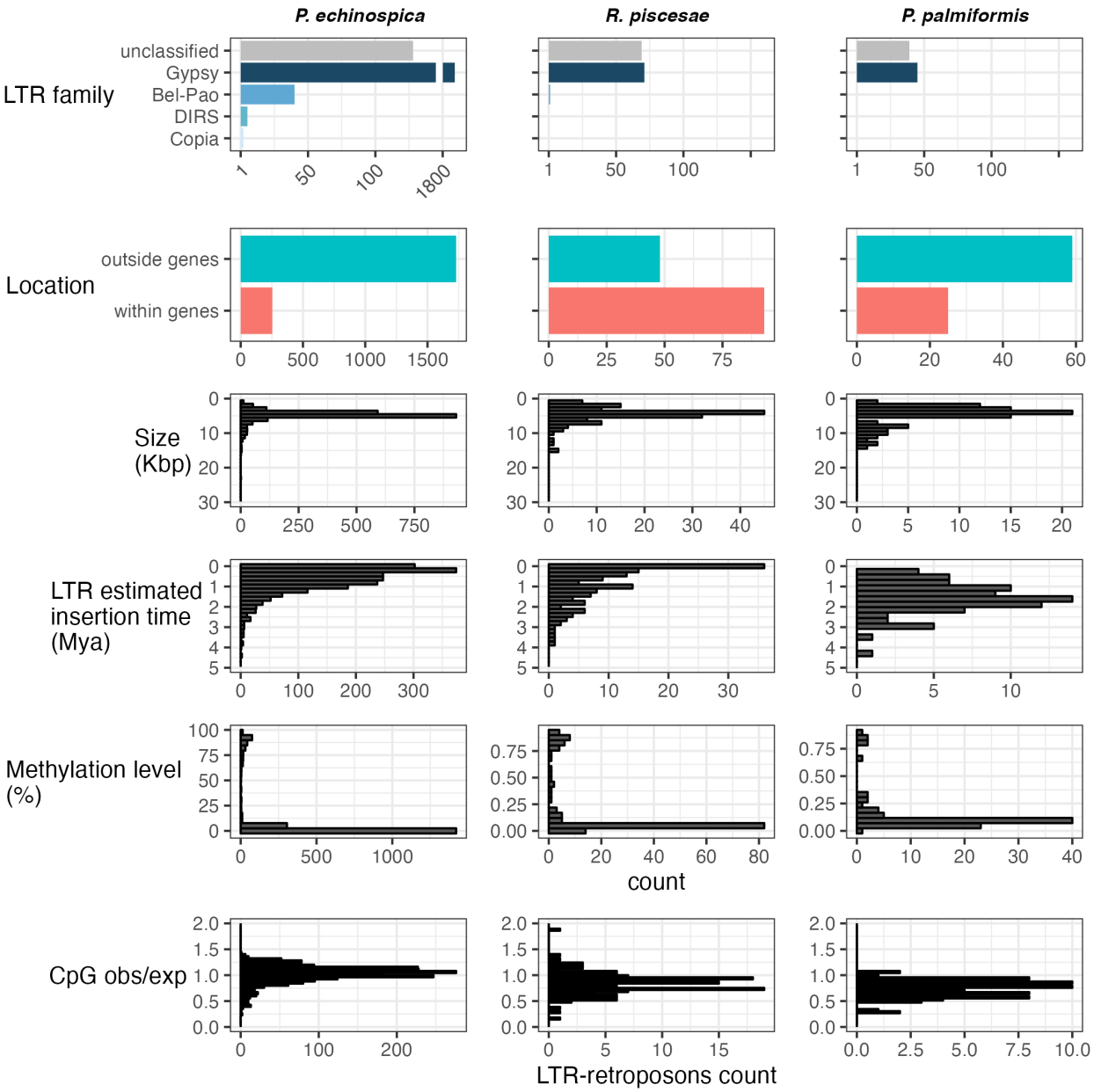

Figure S4.19 Distribution of multiple LTRs characteristics.

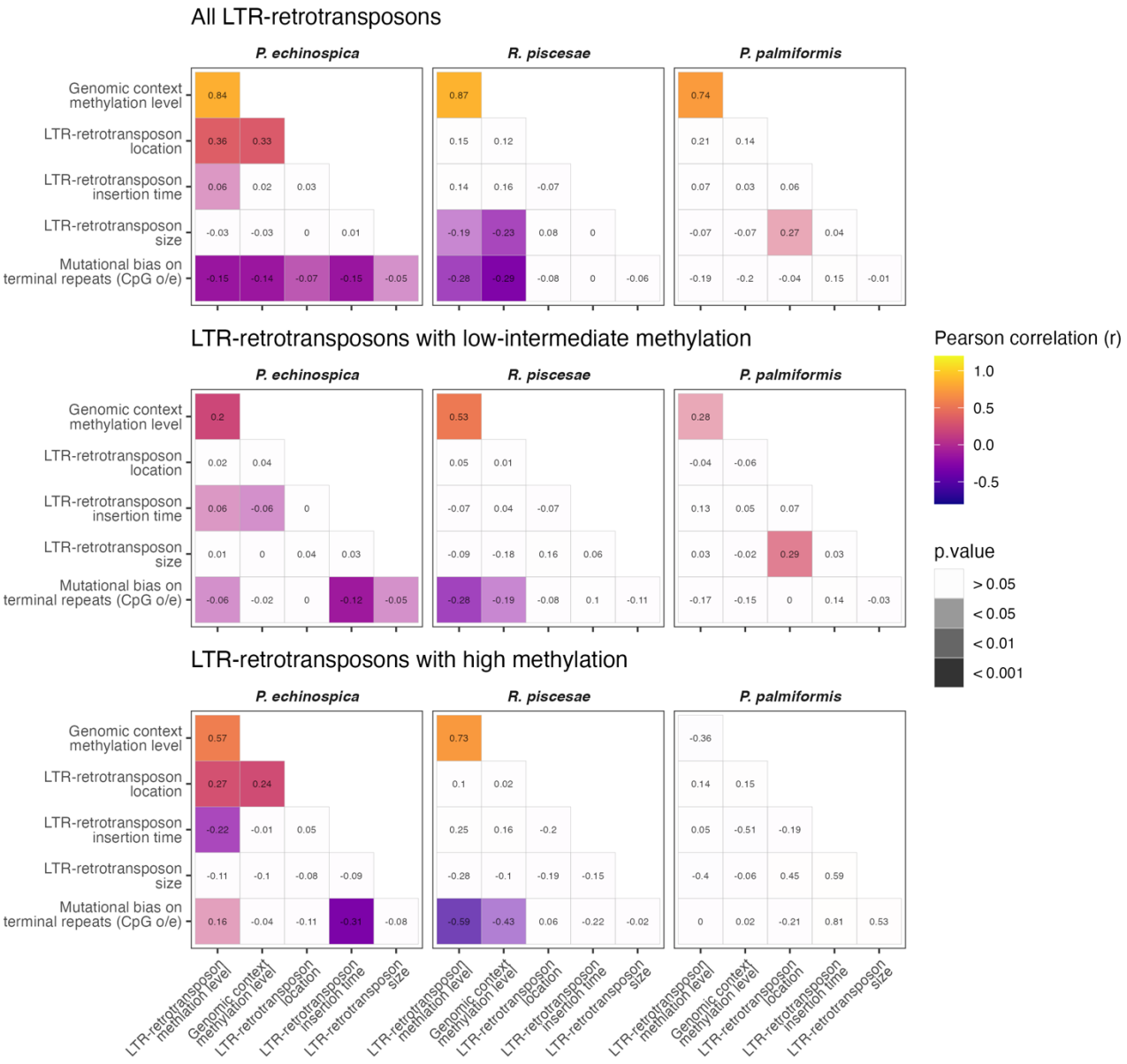

Figure S4.20 Pairwise Pearson correlations between pairs of LTR-retrotransposon characteristics.

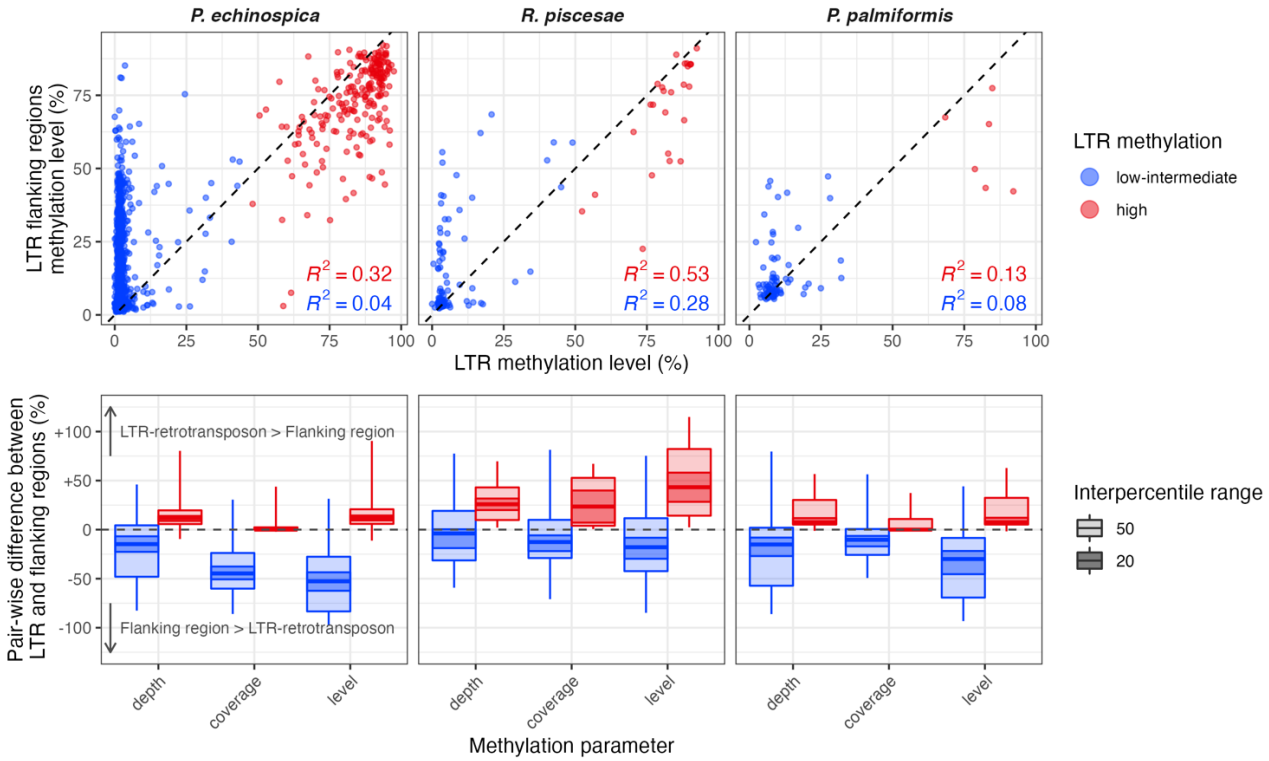

Figure S4.21 Methylation level on LTRs and their respective genomic context. Top: Positive correlation between DNA methylation on LTRs and their respective genomic context (average 4Kbp flanking upstream and downstream regions). **Bottom**: Methylation bias in LTR-retrotransposons compared to their 4Kbp flanking regions. methylation depth = % reads for which methylation was detected; methylation coverage = % called CpGs for which methylation was detected; methylation level = % mean methylation frequency (i.e. depth x coverage).

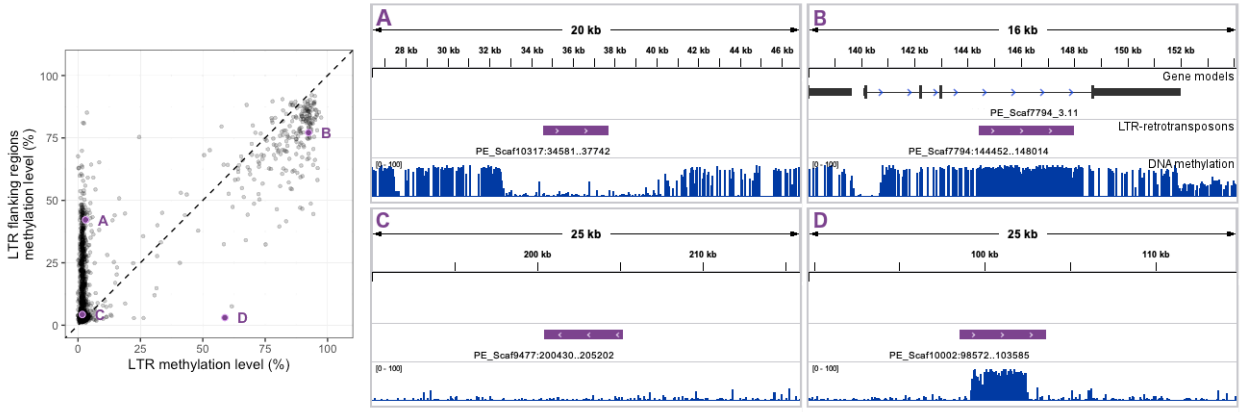

Figure S4.22 Genome browser view of representative LTR-retrotransposons. Left: Distribution of LTR-retrotransposon according to their methylation level and that of their 4Kb flanking regions. **Right**: panels A-C show the genome-browser views of the LTR-retrotransposons annotated A-C in the scatter plot, respectively.

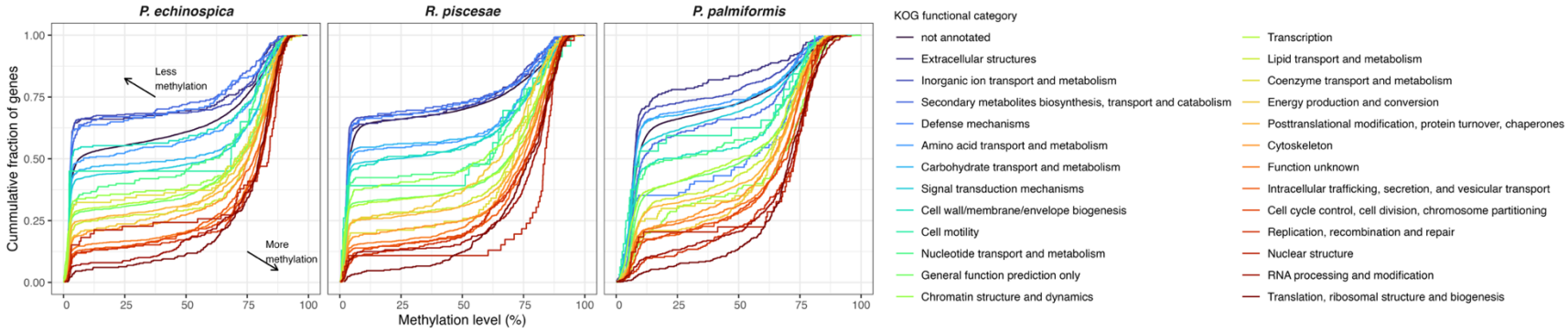

Figure S4.23 ECDF plots of genes mean methylation frequency in KOG functional categories. Only genes with 10 or more called CpGs were considered.

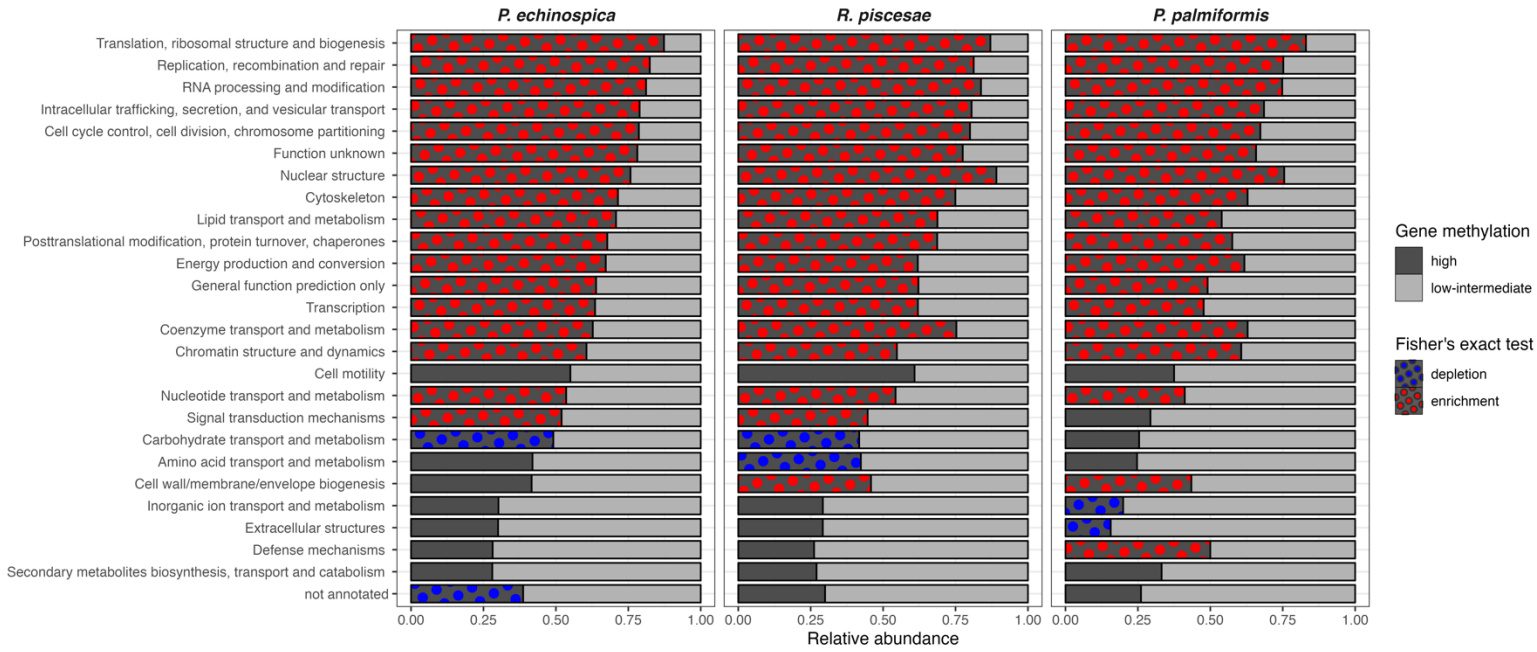

Figure S4.24 Relative abundance of highly methylated genes (methylation coverage >75% and methylation depth >50%) across KOG functional categories. Only genes with 10 CpGs or more were considered. Fisher's exact tests followed by Holms correction were used to test for the enrichment of methylated genes in each functional category compared to the functional distribution of all genes for each genome, respectively.

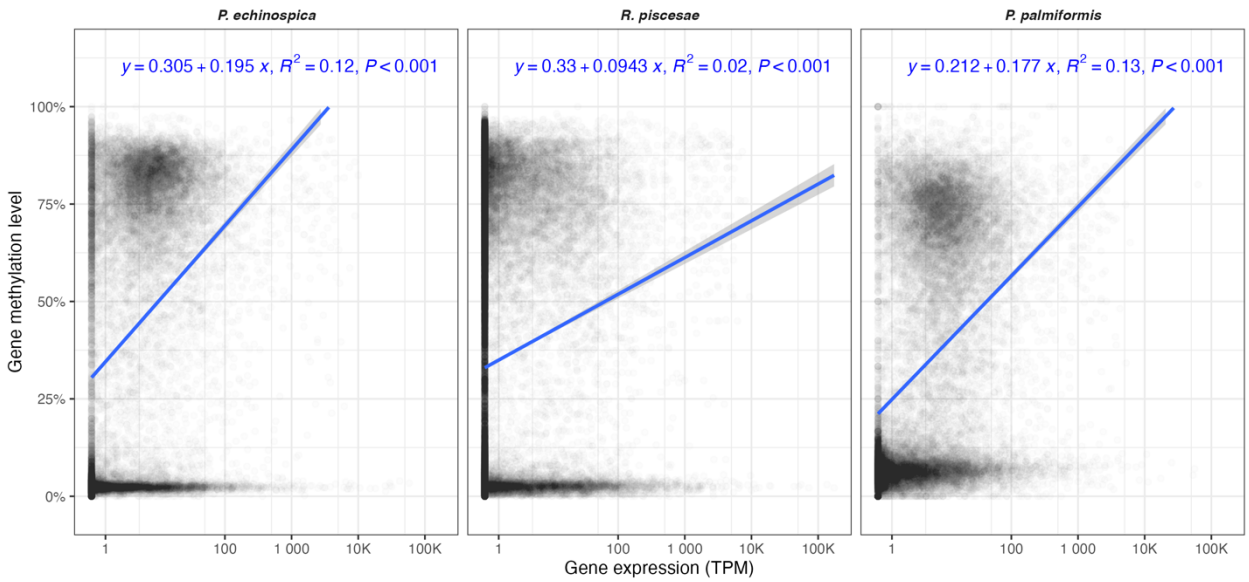

Figure S4.25 Positive correlations between gene methylation and expression in *P. echinospica*. Linear regressions are represented in blue.

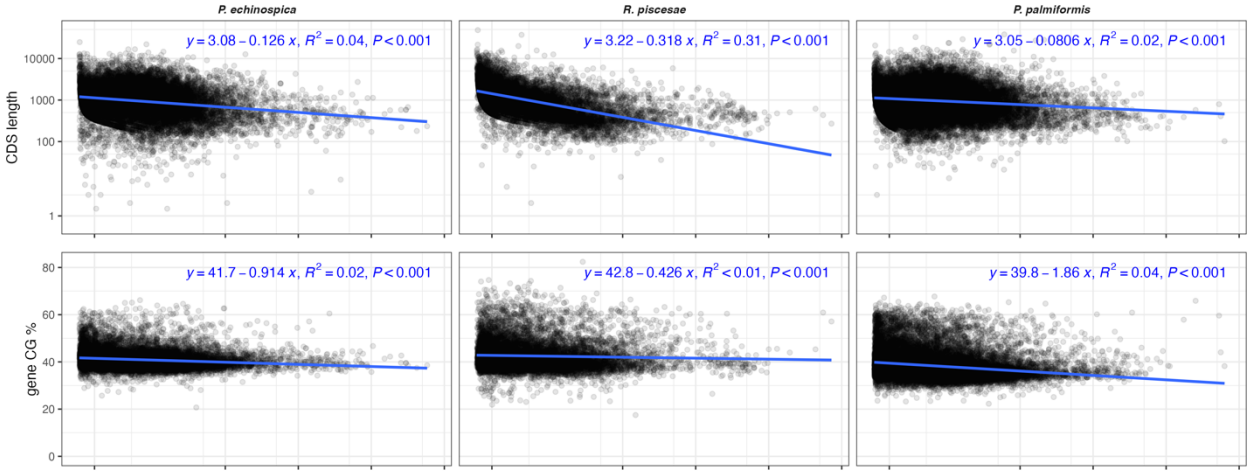

Figure S4.26 Positive correlations between gene characteristics and expression in *P. echinospica*. Linear regressions are represented in blue.

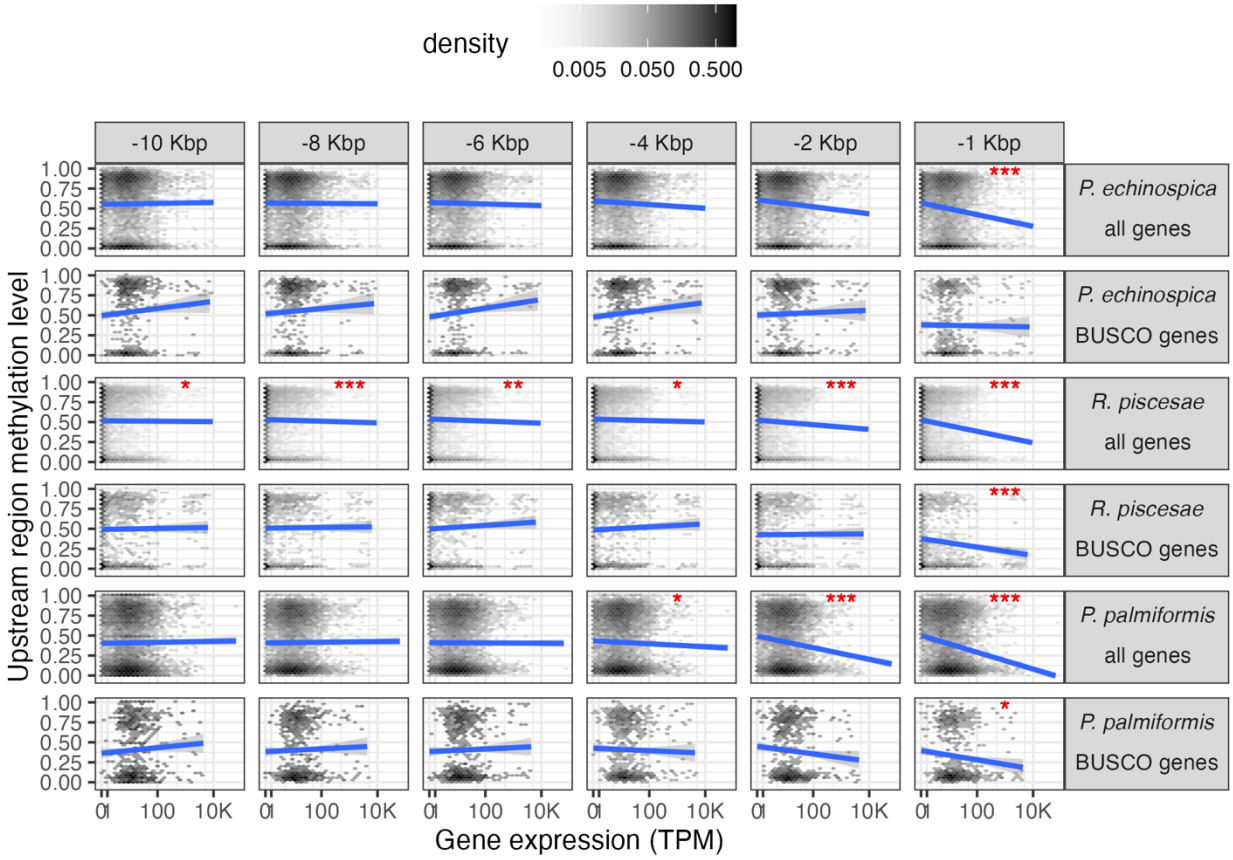

Figure S4.27 Negative correlations between gene expression and methylation level in sliding 1Kbp windows upstream of the transcription start site. Linear regressions are represented in blue but to account for known confounding factors, the significance of the monotonic relation between gene upstream methylation and expression was tested with partial correlation tests (method = Spearman) controlling for gene and CDS GC content and length (see Table S6 and S7 for detailed results). The significance of the partial correlations are indicated by red asterixes. *: pvalue < 0.1; **: pvalue < 0.05; ***: pvalue < 0.01.

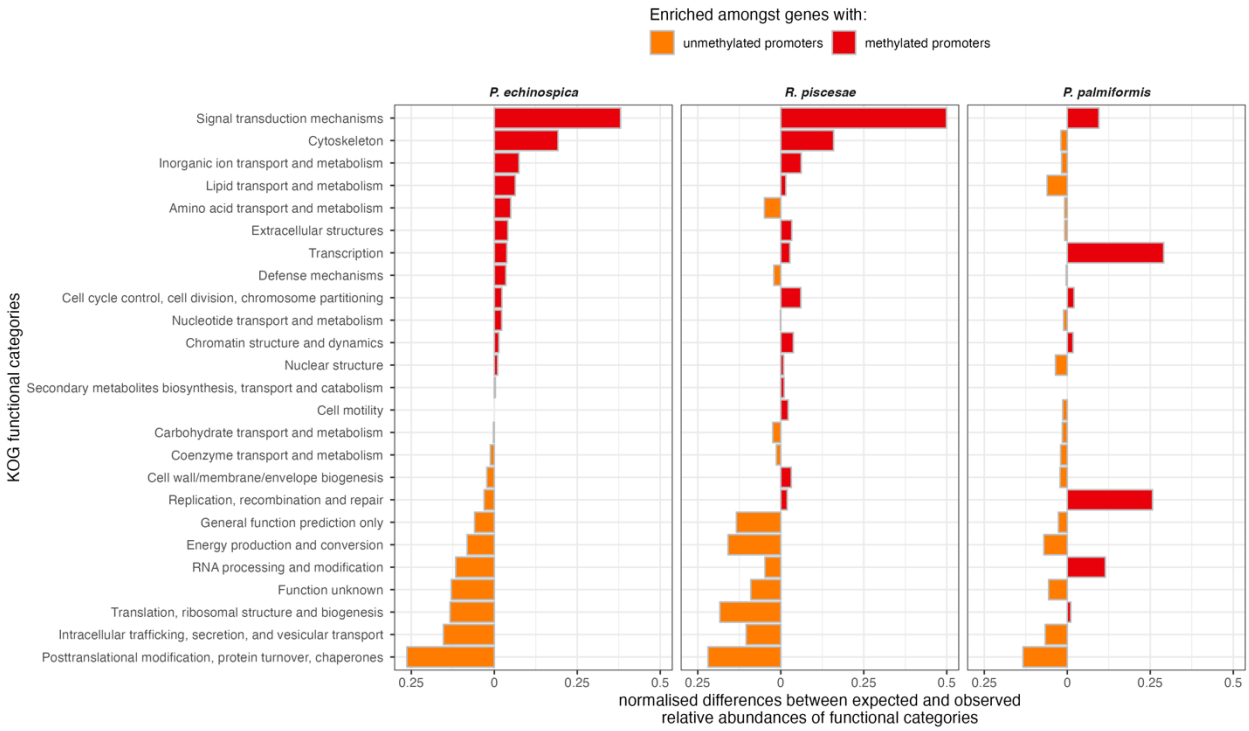

Figure S4.28 KOG functional enrichment of genes with unmethylated (low U high G) and methylated (high UG) promoters as compared to all highly methylated genes in each species. Only significant enrichments (Fisher test pvalue < 0.05) are represented.

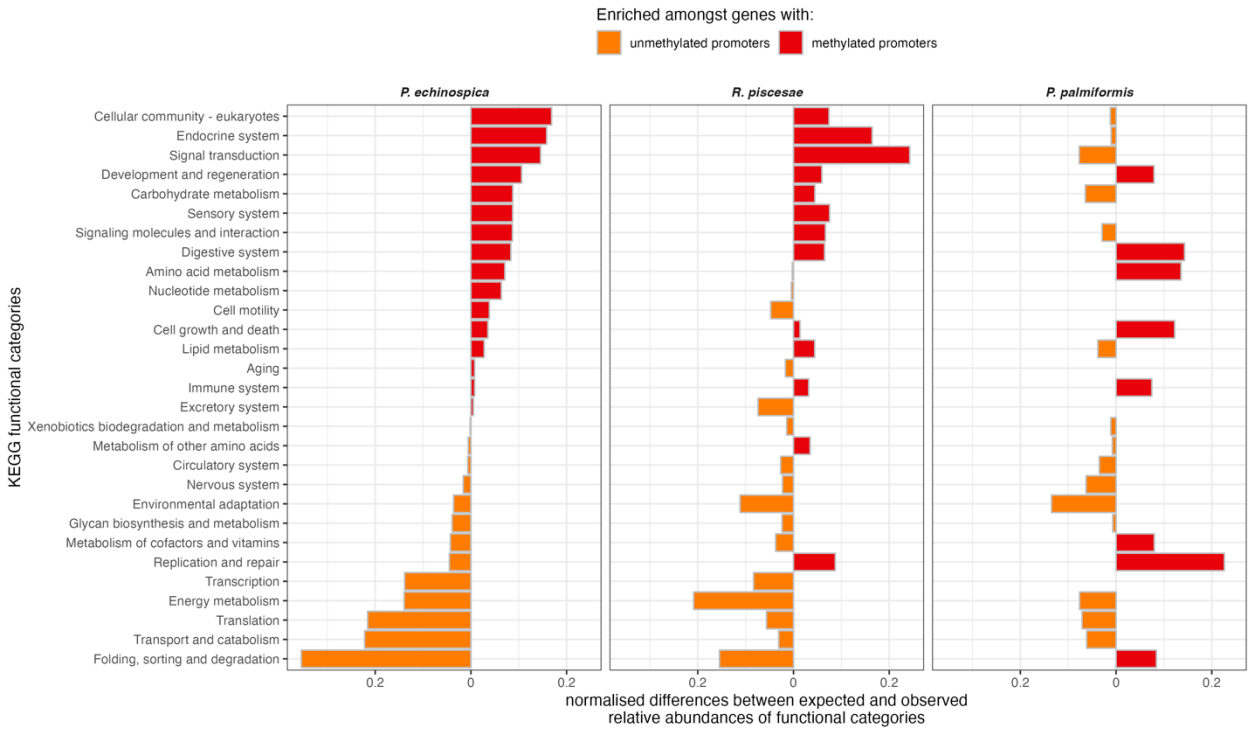

Figure S4.29 KEGG functional enrichment of genes with unmethylated (low U high G) and methylated (high UG) promoters as compared to all highly methylated genes in each species. Only significant enrichments (Fisher test pvalue < 0.05) are represented.

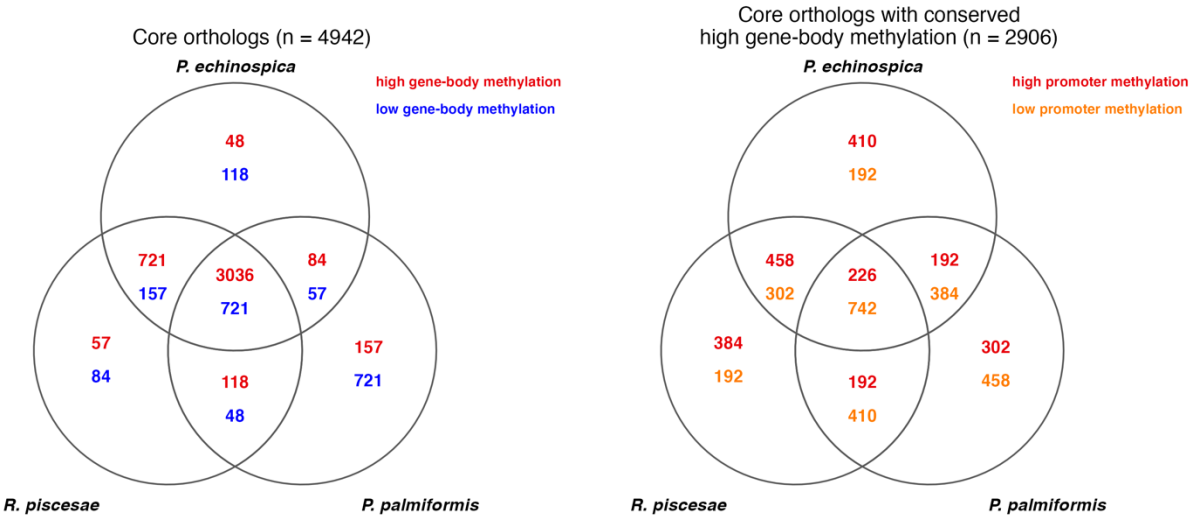

Figure S4.30 Shared gene-body (left) and promoter (right) methylation state of single copy orthologs across the worm species.

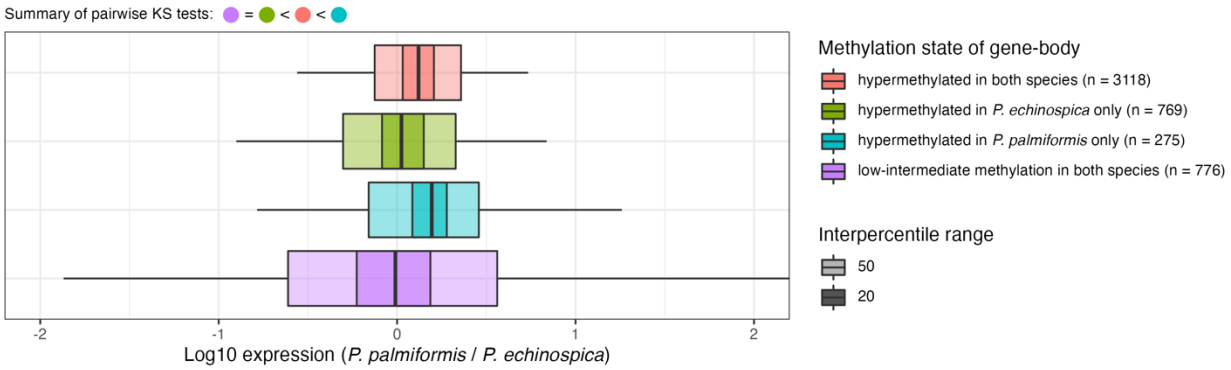

Figure S4.31 Gene-body methylation changes across *P. echinospica* and *P. palmiformis* is accompanied by shifts in their relative expression. The monotonic relations between the distributions of relative expression (*P. palmiformis/P. echinospica*) was tested with pairwise KS tests using a p-value threshold of 0.05.

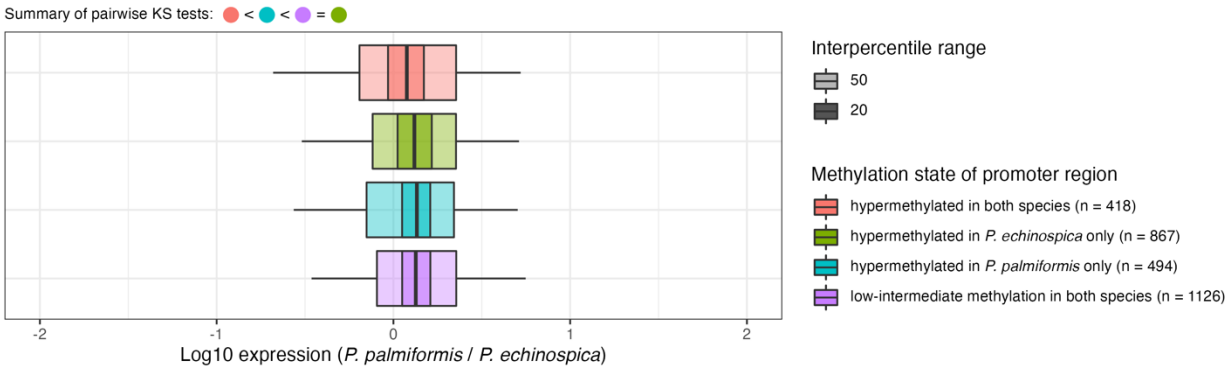

Figure S4.32 Promoter methylation changes across *P. echinospica* and *P. palmiformis* is accompanied by shifts in their relative expression. The monotonic relations between the distributions of relative expression (*P. palmiformis/P. echinospica*) was tested with pairwise KS tests using a p-value threshold of 0.05.

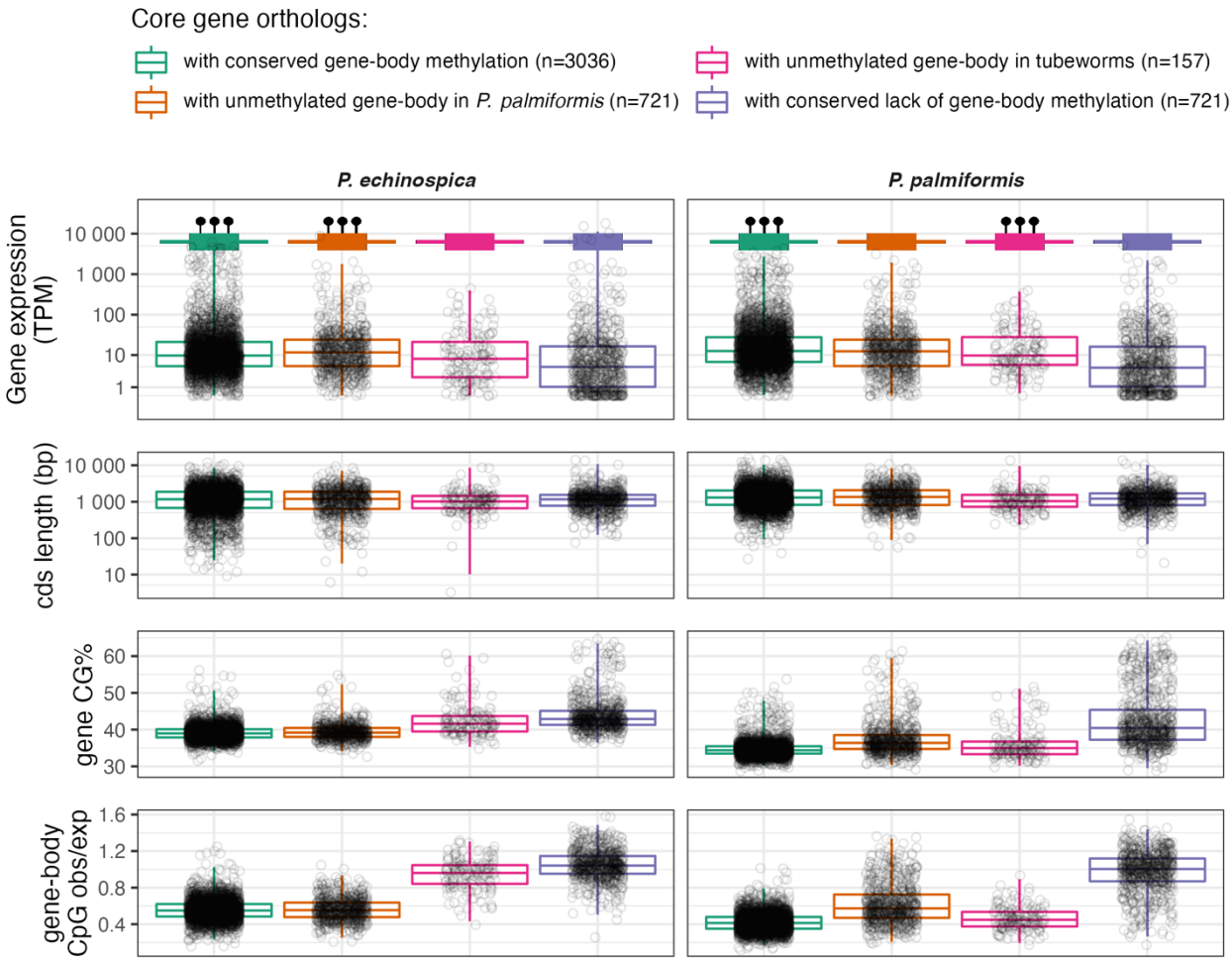

Figure S4.33 Gene-body methylation variation across worm families. Distribution of gene expression, coding sequence length, gene CG content, and gene CpG observed/expected for groups of orthologous genes classified according to their methylation level in the tubeworms and *P. palmiformis*. Gene-bodies were considered unmethylated if they had low-intermediate methylation level (methylation depth ≤ 0.5 or methylation coverage ≤ 0.75).The whiskers extend to 95% of the data while the boxes represent the medians and interquartile ranges. Schematic representations of the gene methylation profile groups are displayed above the gene expression distributions with methylation marks represented by black pins.

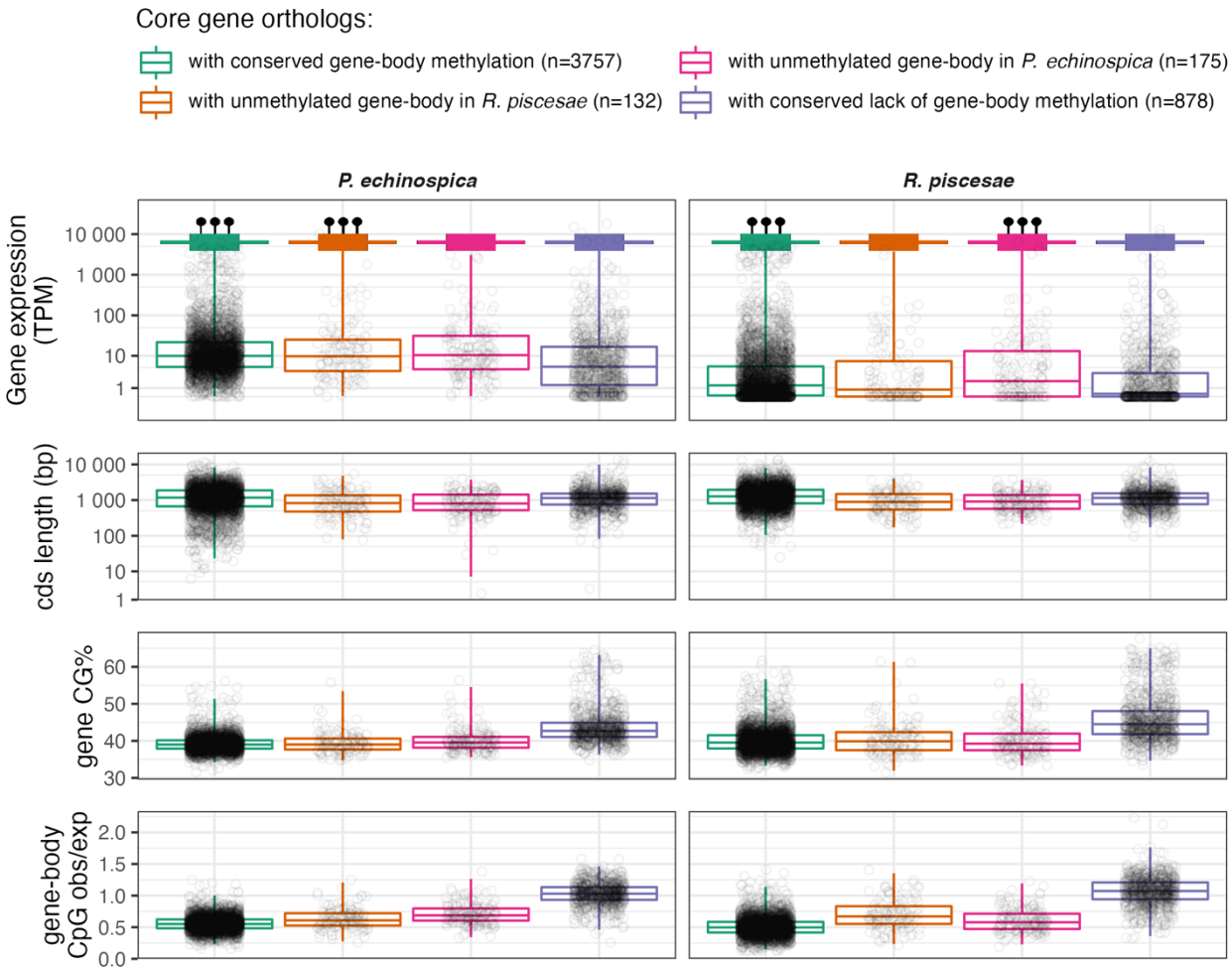

Figure S4.34 Gene-body methylation variation between the two tubeworm species. Distribution of gene expression, coding sequence length, gene CG content, and gene CpG observed/expected for groups of orthologous genes classified according to their methylation level in *P. echinospica* and *R. piscesae*. Gene-bodies were considered unmethylated if they had low-intermediate methylation level (methylation depth ≤ 0.5 or methylation coverage ≤ 0.75).The whiskers extend to 95% of the data while the boxes represent the medians and interquartile ranges. Schematic representations of the gene methylation profile groups are displayed above the gene expression distributions with methylation marks represented by black pins.

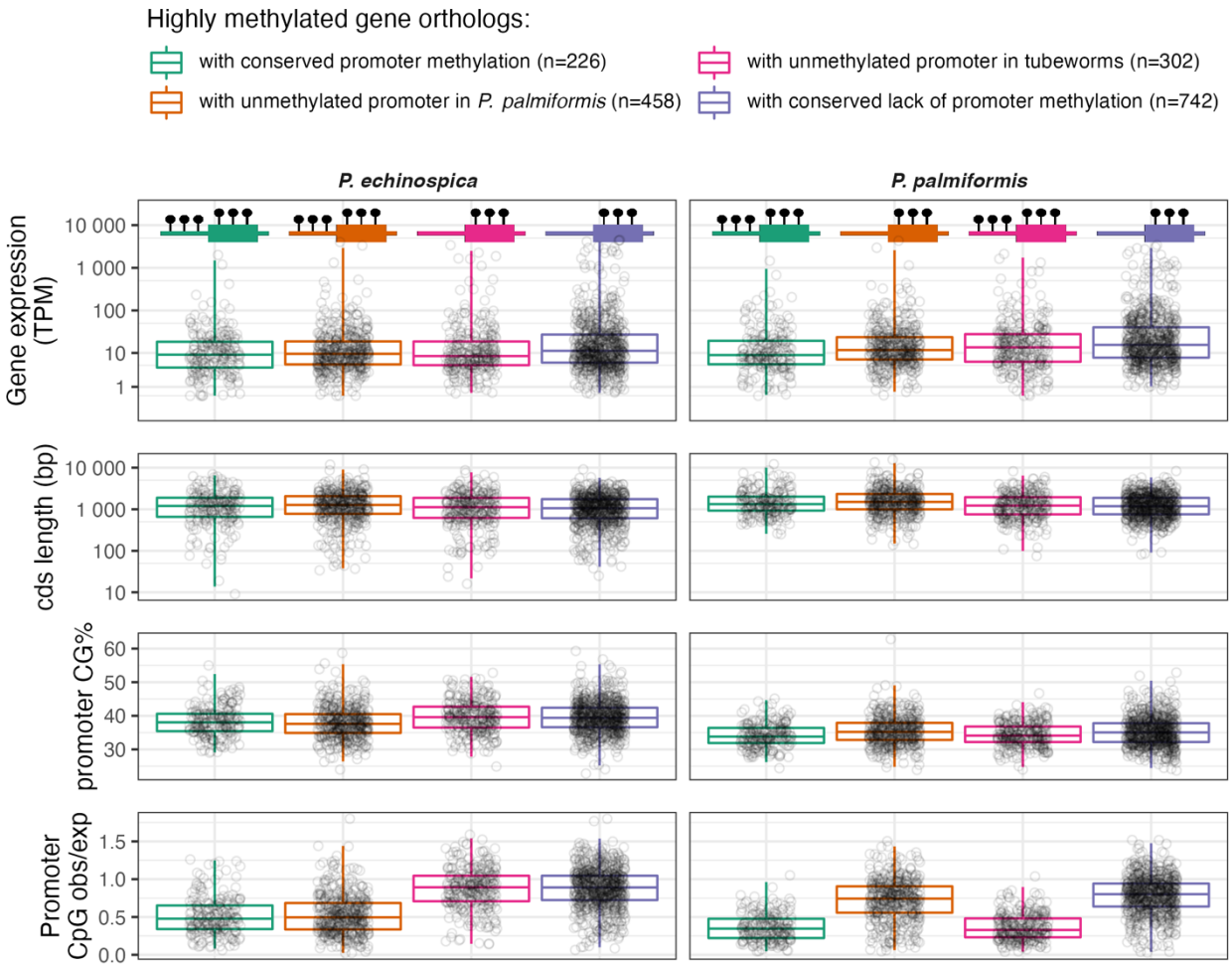

Figure S4.35 Promoter methylation variation across worm families. Distribution of gene expression, coding sequence length, gene CG content, and promoter CpG observed/expected for groups of orthologous genes classified according to their promoter (1Kbp upstream) methylation profile in the tubeworms and *P. palmiformis*. Promoter regions were considered unmethylated if they had low-intermediate methylation level (methylation depth ≤ 0.5 or methylation coverage ≤ 0.75). The whiskers extend to 95% of the data while the boxes represent the medians and interquartile ranges. Schematic representations of the promoter and gene methylation profiles of the orthologous gene groups are displayed above the gene expression distributions with methylation marks represented by black pins.

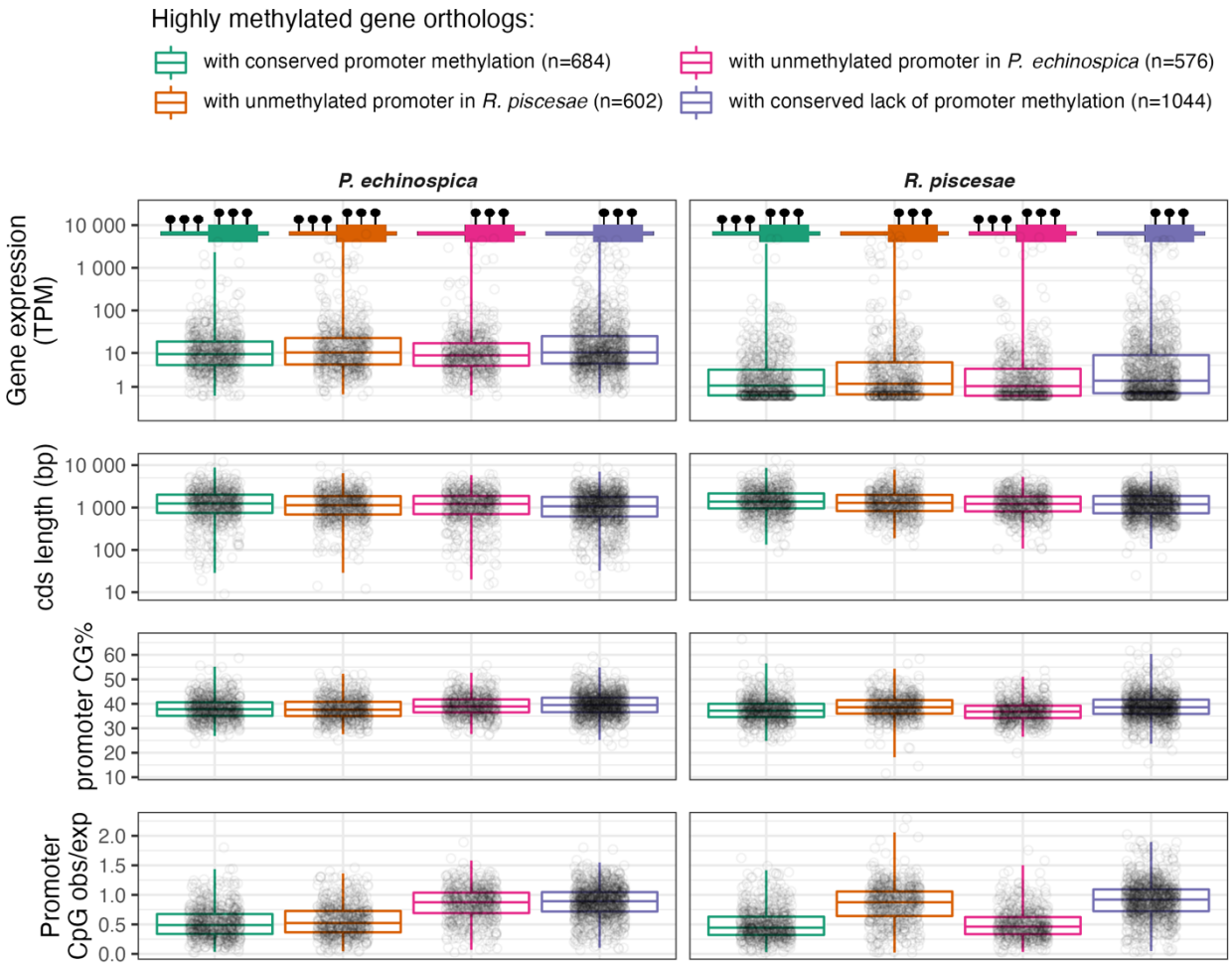

Figure S4.36 Promoter methylation variation across worm families. Distribution of gene expression, coding sequence length, gene CG content, and promoter CpG observed/expected for groups of orthologous genes classified according to their promoter (1Kbp upstream) methylation profile in *P. echinospica* and *R. piscesae*. Promoter regions were considered unmethylated if they had low-intermediate methylation level (methylation depth ≤ 0.5 or methylation coverage ≤ 0.75). The whiskers extend to 95% of the data while the boxes represent the medians and interquartile ranges. Schematic representations of the promoter and gene methylation profiles of the orthologous gene groups are displayed above the gene expression distributions with methylation marks represented by black pins.

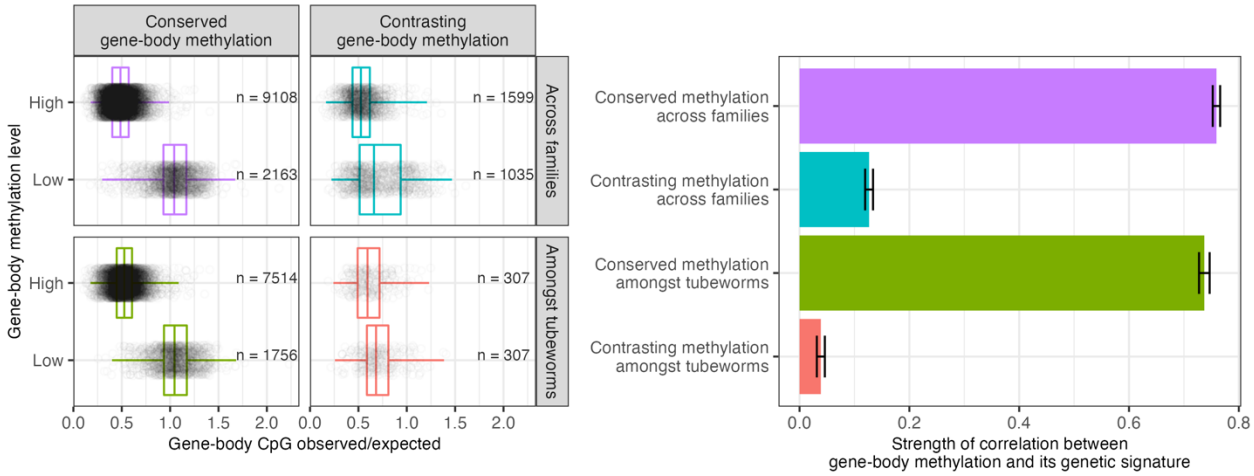

Figure S4.37 Stronger genetic signatures are associated with persistent gene-body methylation across taxa. Weaker association between promoter methylation and its genetic signature is found amongst tubeworms (*P. echinospica* vs *R. piscesae*) than across worm families (tubeworms vs *P. echinospica*) and amongst genes with non-conserved promoter methylation patterns. Left: Distribution of promoter CpG observed/expected for single copy gene orthologs grouped according to their promoter methylation profile across taxa. The whiskers extend to 95\\% of the data while the boxes represent the medians and interquartile ranges. Right: For each gene group, the correlation strength was estimated by the goodness of fit of the logistic regression of promoter CpG observed/expected on promoter methylation level. The mean and standard deviation of 1000 bootstrap samples are represented.

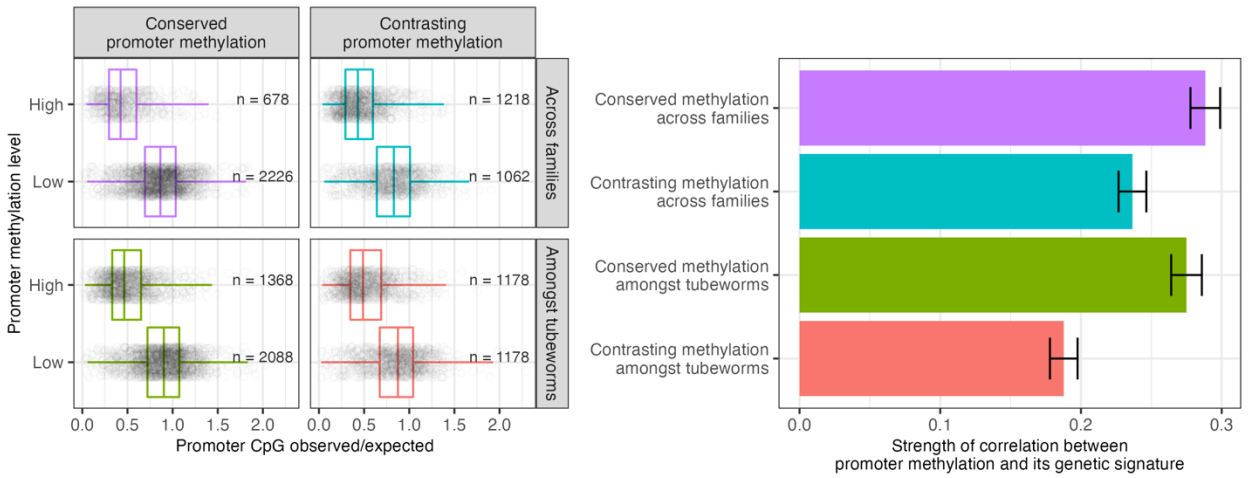

Figure S4.38 Stronger genetic signatures are associated with persistent promoter methylation across taxa. Weaker association between promoter methylation and its genetic signature is found amongst tubeworms (*P. echinospica* vs *R. piscesae*) than across worm families (tubeworms vs *P. echinospica*) and amongst genes with non-conserved promoter methylation patterns. Left: Distribution of promoter CpG observed/expected for single copy gene orthologs grouped according to their promoter methylation profile across taxa. The whiskers extend to 95% of the data while the boxes represent the medians and interquartile ranges. Right: For each gene group, the correlation strength was estimated by the goodness of fit of the logistic regression of promoter CpG observed/expected on promoter methylation level. The mean and standard deviation of 1000 bootstrap samples are represented.

Supplementary Tables

Tables S4.1 to S4.6 are presented within the Supplementary Methods.

Table S4.7 DNMT1 sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	DNMT1					
	DMAP1-binding	DNMT1-RDF	Zf-CXXC	BAH1	BAH2	Dcm
Pfam	**PF06464**	**PF12047**	**PF02008**	**PF01426**	**PF01426**	**PF00145**
P. echinospica						
PE_Scaf8378_0.5	5' partial	present	present			
PE_Scaf8378_1.9				present	present	present
pe4_TRINITY_DN50770_c0_g1	5' partial	present	present	present	present	present
R. piscesae						
NP493_146g03046	present	present	present	present	present	present
P. palmiformis						
LSH36_240g01020	present	present	present	present	present	present
TRINITY_DN10844_c0_g3_i1				present		
TRINITY_DN10844_c0_g2_i1						5' and 3' partial
P. dumerilii						
QVX32595.1	present	present	present	present	present	present

Table S4.8 DNMT3 sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	DNMT3		
	PWWP	ADD	Dcm
Pfam	**PF00855**	**PF17980**	**PF00145**
P. echinospica			
PE_Scaf12509_5.1			5' partial
reannotation of PE_Scaf12509_5.1	5' partial	present	present
Pec_combn_TRINITY_GG_25762_c1_g1_i1	5' partial		
Pec_combn_TRINITY_GG_6065_c2_g1_i1		present	3' partial
R. piscesae			
NP493_663g03006	present	present	present
P. palmiformis			
LSH36_780g00008	present		
LSH36_2519g00000	present	3' patial	
LSH36_2519g00013		5' partial	
LSH36_3177g00008			3' partial
LSH36_3177g00000			3' partial
reassembly of DNMT3 contigs	present	present	present
TRINITY_DN1240_c1_g1	present		
H. robusta			
ESN99160.1	present	present	present

Table S4.9 TET sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	TET
	Tet_JBP
Pfam	**PF12851**
P. echinospica	
PE_Scaf5346_7.1	5' partial
Pec2_TRINITY_DN1533_c0_g1_i41	present
R. piscesae	
NP493_370g03019	present
P. palmiformis	
LSH36_127g18064	present
TRINITY_DN23735_c0_g1_i1	5' and 3' partial
D. melanogaster	
M9NEY8	present

Table S4.10 MAT sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	MAT		
	S-AdoMet_synt_N	AdoMet_synt_M	AdoMet_synt_C
Pfam	**PF00438**	**PF02772**	**PF02773**
P. echinospica			
Pec_PE_Scaf1649_4.5	present	present	present
TRINITY_DN20854_c0_g1_i1	present	present	present
Pec_PE_Scaf1649_4.6	present	present	present
TRINITY_DN4765_c0_g2_i1	5' partial	present	present
R. piscesae			
NP493_347g03093	present	present	present
NP493_347g03092	present	present	present
TRINITY_DN27028_c0_g1_i1	present	present	present
P. palmiformis			
LSH36_68g16017	present	present	present
TRINITY_DN2991_c0_g4_i2	present	present	present
LSH36_132g00002	present	present	present
TRINITY_DN27028_c0_g1_i1	present	present	3' partial
P. caudatus			
Q6PTE5	present	present	present

Table S4.11 MTR sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	MTR				
	S-methyl_trans	Pterin_bind	B12-binding	B12-binding_2	Met_synt_B12
Pfam	**PF02574**	**PF00809**	**PF02310**	**PF02607**	**PF02965**
P. echinospica					
PE_Scaf12088_5.9	present	present	present	present	3' partial
PE_Scaf10753_11.6					5' partial
TRINITY_DN1482_c0_g4_i1	present	present	present	present	present
R. piscesae					
NP493_734g01015	present	present	present	present	present
TRINITY_DN80800_c0_g1_i1		3' partial			
P. palmiformis					
LSH36_405g01009	present	present	present	present	present
TRINITY_DN26095_c0_g1_i1	3' partial				
C. elegans					
Q09582	present	present	present	present	present

Table S4.12 BHMT sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	BHMT
	S-methyl_trans
Pfam	**PF02574**
P. echinospica	
PE_Scaf12331_0.15	present
pe4_TRINITY_DN818_c0_g1_i1	present
R. piscesae	
NP493_445g01015	present
H. sapiens	
Q9H2M3	present

Table S4.13 AHCYL sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	AHCYL
	AdoHcyase_NAD
Pfam	**PF00670**
P. echinospica	
PE_Scaf6625_1.3	5' partial
TRINITY_DN286_c0_g1_i10	present
TRINITY_DN286_c0_g1_i2	present
R. piscesae	
NP493_212g10017	present
TRINITY_DN78702_c0_g1_i1	5' partial
P. palmiformis	
LSH36_580g00068	
LSH36_580g00069	3' partial
TRINITY_DN1964_c0_g1_i4	present
H. sapiens	
O43865	present

Table S4.14 AHCY sequences recovered from the worms genomes and transcriptomes. Dark grey: gene model sequences; light grey: de novo transcriptome sequences.

	AHCY
	AdoHcyase_NAD
Pfam	**PF00670**
P. echinospica	
PE_Scaf7756_5.5	present
TRINITY_DN6635_c0_g1_i2	present
R. piscesae	
NP493_37g11015	present
TRINITY_DN1919_c0_g1_i1	present
P. palmiformis	
LSH36_915g00057	present
TRINITY_DN842_c0_g1_i1	present
H. sapiens	
P23526	present

Table S4.15 Results of the relaxed selection tests performed with RELAX.

clade	test branches	reference branches	selection parameter K	p-value	likelihood ratio
Vestimentifera	MATb	MATa	0.08	0.000	18.94
Terebellidae	MATb	MATa	0.81	0.398	0.72
Alvinellidae + Amphraetidae	MATb	MATa	0.92	0.428	0.63

Table S4.16 Spearman partial correlations between LTR methylation and insertion time controlling for epigenomic context (i.e. 4Kbp flanking regions methylation level). LTRs missing insertion time estimates (n=2) or flanking regions data (n=3) were excluded. *P. echinospica*: n = 1971; *R. piscesae*: n = 138; *P. palmiformis*: n = 84.

Controlled variables	Correlation estimate	P.value
P. echinospica		
none	0.2057149	0.00000
Epigenomic context	0.2253600	0.00000
R. piscesae		
none	-0.0453592	0.59731
Epigenomic context	-0.1518102	0.07658
P. palmiformis		
none	0.0843843	0.44536
Epigenomic context	0.0673971	0.54492

Table S4.17 Contrasting methylation level on LTRs compared to their flanking regions. Results of paired Wilcoxon rank sum tests between LTR methylation and genomic context (4 kbp upstream and downstream regions) methylation for LTRs with high and low methylation, respectively.

LTR count	LTR methylation	alternative hypothesis	p.value
P. echinospica			
221	high	methylation on LTR > flanking region	0.00000
1760	low-intermediate	methylation on LTR < flanking region	0.00000
R. piscesae			
26	high	methylation on LTR > flanking region	0.00000
114	low-intermediate	methylation on LTR < flanking region	0.00000
P. palmiformis			
6	high	methylation on LTR > flanking region	0.01563
78	low-intermediate	methylation on LTR < flanking region	0.00110

Table S4.18 Spearman partial correlations between gene methylation and expression. *P. echinospica*: n = 21252, transcriptome = same individual and tissue as epigenome; *R. piscesae*: n = 30089, transcriptome = same individual different tissues as epigenome; *P. palmiformis*: n = 24552, transcriptome = different individuals and tissues as epigenome.

Controlled variables	Correlation estimate	P.value
P. echinospica		
CDS length	0.3602243	0.00000
gene GC%	0.3252294	0.00000
gene GC% + CDS length	0.3011819	0.00000
R. piscesae		
CDS length	0.1948939	0.00000
gene GC%	0.1917303	0.00000
gene GC% + CDS length	0.1622550	0.00000
P. palmiformis		
CDS length	0.3918341	0.00000
gene GC%	0.3085895	0.00000
gene GC% + CDS length	0.3014488	0.00000

Table S4.19 The negative correlation between 1Kbp upstream methylation and gene expression breaks down with increasing distance to the TSS. Spearman partial correlations between mean methylation level of 1Kbp window upstream of transcription start site and gene expression for all highly methylated genes. Statistically significant negative correlations (pvalue < 0.1) are indicated in red. Corr. est.: Correlation estimate; n : gene count.

	P. echinospica			*R. piscesae*			*P. palmiformis*		
Controlled variables	**Cor. est.**	**P.value**	**n**	**Cor. est.**	**P.value**	**n**	**Cor. est.**	**P.value**	**n**
1 Kbp upstream									
CDS CG content	-0.099	0.00000	9949	-0.118	0.00000	18924	-0.177	0.00000	8861
CDS length	-0.093	0.00000	9949	-0.120	0.00000	18924	-0.201	0.00000	8861
Gene length	-0.047	0.00000	9949	-0.101	0.00000	18924	-0.169	0.00000	8861
CDS CG content + CDS length	-0.101	0.00000	9949	-0.124	0.00000	18924	-0.203	0.00000	8861
CDS CG content + Gene length	-0.055	0.00000	9949	-0.101	0.00000	18924	-0.169	0.00000	8861
2 Kbp upstream									
CDS CG content	-0.033	0.00106	9697	-0.056	0.00000	18693	-0.115	0.00000	8827
CDS length	-0.033	0.00120	9697	-0.058	0.00000	18693	-0.134	0.00000	8827
Gene length	-0.007	0.50777	9697	-0.042	0.00000	18693	-0.107	0.00000	8827
CDS CG content + CDS length	-0.035	0.00063	9697	-0.061	0.00000	18693	-0.134	0.00000	8827
CDS CG content + Gene length	-0.009	0.35703	9697	-0.042	0.00000	18693	-0.107	0.00000	8827
4 Kbp upstream									
CDS CG content	-0.010	0.32052	9459	-0.024	0.00101	18458	-0.026	0.01411	8752
CDS length	-0.011	0.27463	9459	-0.026	0.00041	18458	-0.043	0.00007	8752
Gene length	0.006	0.55256	9459	-0.014	0.05559	18458	-0.021	0.05470	8752
CDS CG content + CDS length	-0.011	0.26889	9459	-0.027	0.00022	18458	-0.041	0.00012	8752
CDS CG content + Gene length	0.005	0.62982	9459	-0.014	0.05604	18458	-0.020	0.06679	8752
6 Kbp upstream									
CDS CG content	0.010	0.34047	9270	-0.026	0.00041	18252	-0.008	0.47713	8705
CDS length	0.010	0.31590	9270	-0.028	0.00016	18252	-0.023	0.03375	8705
Gene length	0.025	0.01666	9270	-0.018	0.01594	18252	-0.003	0.77025	8705
CDS CG content + CDS length	0.009	0.38094	9270	-0.028	0.00012	18252	-0.021	0.05217	8705
CDS CG content + Gene length	0.023	0.02784	9270	-0.018	0.01591	18252	-0.002	0.85778	8705
8 Kbp upstream									
CDS CG content	0.017	0.10988	9093	-0.027	0.00031	18081	0.009	0.42306	8652
CDS length	0.018	0.09131	9093	-0.028	0.00016	18081	-0.003	0.79867	8652
Gene length	0.027	0.00977	9093	-0.020	0.00832	18081	0.011	0.30344	8652
CDS CG content + CDS length	0.016	0.12317	9093	-0.029	0.00010	18081	-0.000	0.97355	8652
CDS CG content + Gene length	0.025	0.01600	9093	-0.020	0.00863	18081	0.012	0.24954	8652
10 Kbp upstream									
CDS CG content	0.026	0.01436	8541	-0.021	0.00616	16860	0.008	0.45852	8100
CDS length	0.028	0.01055	8541	-0.022	0.00449	16860	-0.001	0.91712	8100
Gene length	0.037	0.00058	8541	-0.014	0.06478	16860	0.010	0.37836	8100
CDS CG content + CDS length	0.026	0.01643	8541	-0.023	0.00275	16860	0.001	0.92083	8100
CDS CG content + Gene length	0.035	0.00118	8541	-0.014	0.06607	16860	0.011	0.32011	8100

Table S4.20 The negative correlation between 1Kbp upstream methylation and BUSCO gene expression breaks down with increasing distance to the TSS. Spearman partial correlations between mean methylation level of 1Kbp window upstream of transcription start site and gene expression for complete BUSCO highly methylated genes. Statistically significant negative correlations (pvalue < 0.1) are indicated in red. Corr. est.: Correlation estimate; n : gene count.

	P. echinospica			*R. piscesae*			*P. palmiformis*		
Controlled variables	**Cor. est.**	**P.value**	**n**	**Cor. est.**	**P.value**	**n**	**Cor. est.**	**P.value**	**n**
1 Kbp upstream									
CDS CG content	-0.015	0.70006	690	-0.089	0.00075	1425	-0.061	0.09948	727
CDS length	-0.011	0.77334	690	-0.070	0.00810	1425	-0.079	0.03268	727
Gene length	-0.029	0.44628	690	-0.086	0.00115	1425	-0.115	0.00198	727
CDS CG content + CDS length	-0.012	0.75998	690	-0.079	0.00283	1425	-0.067	0.07309	727
CDS CG content + Gene length	-0.029	0.44139	690	-0.092	0.00054	1425	-0.106	0.00415	727
2 Kbp upstream									
CDS CG content	0.031	0.41541	689	-0.005	0.86185	1415	-0.050	0.18283	726
CDS length	0.041	0.28747	689	-0.012	0.64387	1415	-0.038	0.30304	726
Gene length	0.024	0.53730	689	-0.011	0.68741	1415	-0.069	0.06447	726
CDS CG content + CDS length	0.041	0.27829	689	-0.019	0.46941	1415	-0.025	0.50568	726
CDS CG content + Gene length	0.024	0.53816	689	-0.014	0.60759	1415	-0.060	0.10566	726
4 Kbp upstream									
CDS CG content	0.060	0.12326	673	0.012	0.65501	1401	-0.038	0.30544	719
CDS length	0.056	0.15065	673	-0.006	0.81981	1401	-0.054	0.14893	719
Gene length	0.057	0.13721	673	-0.002	0.93622	1401	-0.056	0.13623	719
CDS CG content + CDS length	0.059	0.12369	673	-0.006	0.83284	1401	-0.047	0.20597	719
CDS CG content + Gene length	0.058	0.13170	673	-0.002	0.94997	1401	-0.051	0.17407	719
6 Kbp upstream									
CDS CG content	0.077	0.04620	664	0.024	0.37877	1362	0.019	0.61840	718
CDS length	0.068	0.08034	664	0.016	0.55472	1362	0.006	0.87710	718
Gene length	0.069	0.07642	664	0.015	0.57938	1362	0.005	0.88540	718
CDS CG content + CDS length	0.069	0.07449	664	0.013	0.63825	1362	0.012	0.75859	718
CDS CG content + Gene length	0.069	0.07616	664	0.014	0.61694	1362	0.010	0.79845	718
8 Kbp upstream									
CDS CG content	0.049	0.20933	654	0.003	0.89932	1366	0.048	0.20085	717
CDS length	0.060	0.12566	654	0.002	0.94735	1366	0.039	0.29440	717
Gene length	0.054	0.16965	654	0.005	0.84300	1366	0.032	0.38691	717
CDS CG content + CDS length	0.062	0.11095	654	0.004	0.88272	1366	0.052	0.16326	717
CDS CG content + Gene length	0.054	0.17024	654	0.007	0.80946	1366	0.041	0.26814	717
10 Kbp upstream									
CDS CG content	0.068	0.09149	624	-0.017	0.55567	1278	0.049	0.20640	673
CDS length	0.068	0.09057	624	-0.021	0.45505	1278	0.042	0.27488	673
Gene length	0.067	0.09395	624	-0.016	0.55695	1278	0.027	0.48416	673
CDS CG content + CDS length	0.069	0.08447	624	-0.027	0.32757	1278	0.045	0.24084	673
CDS CG content + Gene length	0.067	0.09466	624	-0.019	0.49914	1278	0.029	0.45537	673