From head to rootlet: comparative transcriptomic analysis of a rhizocephalan barnacle Peltogaster reticulata (Crustacea: Rhizocephala)

Maksim A. Nesterenko (maxnest.research@gmail.com)
St Petersburg State University
Aleksei A. Miroliubov
Zoological Institute

Research Article

Keywords:

Posted Date: January 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1254205/v1

Abstract

Rhizocephalan barnacles stand out in the diverse world of metazoan parasites. The body of a rhizocephalan female is modified beyond revealing any recognizable morphological features, consisting of the interna, the system of rootlets, and the externa, a sac-like reproductive body. Moreover, rhizocephalans have an outstanding ability to control their hosts, literally turning them into “zombies”. Here we performed the first comparative transcriptomic analysis of rhizocephalan (*Peltogaster reticulata*) adult female body parts and established the spatial heterogeneity of gene expression in different regions. This resulted in an unexpected discovery of germ cells in the lumen of interna, which was additionally validated by histological methods. Our finding of germ cells in interna deprives externa from the exclusive role in propagation and renders it a simple brooding chamber. Furthermore, we confirmed that the neurons of the host are attracted to the rootlets of the interna, indicating a great intimacy of host-parasite relationship. Finally, based on the continual expression of development-associated genes, we suggest that rhizocephalans “got stuck in the metamorphosis”, even in their reproductive stage.

Introduction

Rhizocephalan barnacles (Crustacea: Rhizocephala) stand out among metazoan parasites. In the process of adaptation to parasitic lifestyle, they have changed beyond recognition, losing almost all the structures characteristic of other crustaceans. In particular, they have lost all normal systems of organs and even body axes [1, 2]. The body of an adult rhizocephalan female is represented by the interna, a system of hollow ramifying rootlets infiltrating the body cavity of the host (also crustaceans, usually decapods), and the externa, a sac-like body protruding outside the host. The interna is responsible for absorbing nutrients from the host hemolymph and their transportation to the externa [1] as well as for interactions with the host [3, 4]. The externa is a temporary structure thought to be an organ of sexual reproduction [5]. It usually contains two incorporated dwarf males and a special mantle chamber with developing embryos [2]. Some rhizocephalans can form several externas, sometimes as many as 2000 [6, 7], which is considered as a unique instance of modular/colonial organization among arthropods. Besides their unusual morphology rhizocephalans have evolved unique life cycle with a characteristic larval stage. The larva injects a few poorly differentiated cells into the host’s hemolymph, and what is left of the larva dies [8]. Noteworthy, the entire adult body originates from these cells and is thus a newly formed structure [9].

In addition to their morphological adaptations and weird life cycle, rhizocephalan barnacles show a remarkable ability of manipulating the host. These parasites can take control of the moulting cycle of the hosts, change their metabolism, behaviour and even body shape [2, 10–20]. Specialized sites responsible for host-parasite interactions have recently been described [3, 4], with a network of the host’s neurons enlacing the rootlets of parasite, but the molecular mechanisms of these interactions remained enigmatic. The authors suggested that the parasite should emit some signal molecules attracting the growth of the host’s neurons [3].

A rapid development of high-throughput sequencing technologies has enabled detailed molecular-biological studies of many living organisms (for example, [21–25]) but until recently molecular studies on Rhizocephala have been lacking. The research on body heterogony, host-parasite interactions and functional physiology was based only on the morphological and other classical methods.

In an attempt to fill this gap, we made a comparative transcriptomics analysis of different parts of the rhizocephalan female body. Our research object was *Peltogaster reticulata* (Rhizocephala: Peltogasteridae), whose females parasitize hermit crabs *Pagurus minutus* (Crustacea: Decapoda) and form one or, less often, several externas. On the one hand, *P. reticulata* is a typical rhizocephalan, but, on the other hand, it has a lesser degree of modularity than many other representatives of this group. Thus, it is a particularly convenient research model. In this study we present the transcriptome-based evidence of molecular and functional heterogeneity of the female rhizocephalan body. We also show that the ovary is diffused throughout the interna and host`s motor neuron axon are attracted to the interna rootlets. Phylostratigraphy and evolutionary transcriptomic analysis has been performed for Rhizocephala for the first time. Our results make it possible to trace evolutionary trends in the *P. reticulata* gene set.

Results

The de novo assembled transcriptome was characterized by high quality and completeness of assembly

The transcriptomes of the whole body (Fig. 1a), the thoracic part of the interna (Fig. 1a), the growing part of the interna (Fig. 1a) and the main trunk of the interna (Fig. 1a) as well as that of the externa (Fig. 1a) of adult *P. reticulata* were collected and sequenced in two biological replicates. More than 87% of read pairs remained in all the paired-end libraries after the removal of adapters, poor-quality regions, and short sequences [Tab.S1]. The potential contamination with human-derived sequences did not exceed 4.1% from the total number of trimmed read pairs in each library [Tab.S1].

All prepared libraries were merged, and the resulted libraries were used as input for Trinity software. In general, 353130 contigs with lengths greater or equal to 200 nucleotides were assembled de novo. After the clusterization of similar sequences, the *P. reticulata* transcriptome included 267188 contigs. TransRate assembly and optimal scores made up 0.3331 and 0.3835, respectively. More than 95% (256047 / 267188) of the contigs were well-assembled (“good”) according to the TransRate quality control results. The completeness analysis for “good” contigs using a database of the metazoan single-copy orthologues revealed that 96.3% (Single:57.8%, Duplicated:38.5%) of the orthologues were assembled completely.

P. reticulata reference transcriptome included transcripts of 12620 protein-encoding genes

Given the parasitic lifestyle of *P. reticulata*, the assembled transcriptome was checked for the presence of potential contamination. According to the RNAmmer analysis results, 8 and 9 sequences could be classified as 18S and 28S ribosomal RNAs, respectively. The comparison with NCBI nucleotide database revealed that the contigs successfully aligned with ribosomal sequences from Alveolata (HQ891115.2),

Fungi (AY382649.1; CP033152.1; CP030254.1; GQ336996.1; MF611880.1), and Metazoa (AY265359.1; EU082415.1; KY454201.1; EU370441.1; KU052603.1). Among the metazoan hits, only sequences from Crustacea were identified. The identity percentage with the database-derived host (KY454201.1) and parasite (EU082415.1) 18S sequences made up 88% (990 / 1123) and 99% (1200 / 1201), respectively. In this study, we used the MCSC hierarchical clustering algorithm for removing potential contamination from the assembled transcriptome. The decontaminated transcriptome included 80779 contigs and contained 81.6% (Single: 56.9%; Duplicated: 24.7%) of completely assembled single-copy metazoan orthologous. The number of paired-end reads successfully aligned to selected contigs varied from 10.74 (externa, first replicate) to 30.64 million (the thoracic part of the interna, second replicate) [Tab.S1].

The sequence expression level quantification was performed with Salmon by mapping selected read pairs to the decontaminated transcriptome. The mapping rate ranged from 85–91%. The transcript-to-gene map was used to obtain gene expression levels in TPM (Transcript-Per-Millions) values. After excluding genes with a low activity in all analysed samples (expression level < 1 TPM), the dataset contained TPM values for 20980 contigs. According to the TransDecoder results, 32990 contigs encoded proteins with lengths ≥ 100 amino acids.

Only the protein-coding genes with a noticeable expression were involved in the further analysis. After filtering by expression level (≥ 1 TPM in at least one sample) and encoded protein lengths (≥ 100 aa), we obtained the reference gene set for *P. reticulata* containing 12620 sequences. For each gene, the longest protein encoded by its splice variants was selected as a representative sequence. The comparison with a single-copy metazoan orthologues database revealed that in the reference protein set of *P. reticulata* 75.6% of the orthologues were assembled completely, whereas 21% of the sequences were absent.

Most of the sequences from the reference sets were annotated successfully

The prepared reference sets were compared with publicly available databases. According to the results, 3399, 8502, and 9690 genes had hits with NCBI nucleotide, SwissProt, and NCBI non-redundant databases, respectively. The overlap analysis revealed that 3240 genes were successfully annotated using each of the databases. Moreover, 6090 genes belonged to at least one Gene Ontology (GO) terms. The domain architecture of the encoded protein was identified for 8803 genes based on the comparison with the PfamA database. Annotation results are presented in Supplementary table S2.

The proteins set of P. reticulata **was similar to the reference proteome of another cirripede barnacle**, Amphibalanus amphitrite.

The identification of orthogroups in *P. reticulata* and other Crustacea involved in this study was carried out in 3 stages. First, the proteomes of *P. reticulata* and 8 reference crustacean species were analysed using OMA standalone. As a result, 24840 OMA Groups were discovered.

Secondly, a phylogenetic tree of the studied crustacean species was constructed based on the results [Fig.S1]. We selected 609 OMA Groups, containing at least 8 out of 9 species. The multiple protein

alignment results in each of the OMA Groups were concatenated into a supermatrix. After site selection, the final supermatrix contained 282907 sites. In the resulting phylogenetic tree, *P. reticulata* was united into the same taxon with another cirripede barnacle, *Amphibalanus amphitrite*, with full support [Fig.S1].

Thirdly, the resulting tree was used to refine the searching results for orthogroups. More than 20000 orthogroups were found: 24840 and 20874 OMA and Hierarchical Orthologous Groups were identified, respectively. Figure 1b shows the number of common OMA Groups for pairs of species. *Peltogaster reticulata* has the largest number of "common" OMA Groups (4354) with *A. amphitrite*. For comparison, *P. reticulata* had no more than 3490 “common” OMA Groups with other crustacean species, the smallest overlap being found with *Portunus trituberculatus* (1262 OMA Groups).

The externa was clustered separately from the interna based on the gene expression analysis results

We define the molecular signature of a body part as a set of genes with an expression ≥ 2 TPM in the body part considered. Each molecular signature included at least 8 thousand genes: the main trunk of interna (8070 genes), the growing part (8148), the thoracic part (8223), and externa (9233) [Tab. S3]. Approximately 54% (6829 / 12620) of the genes from the reference set were included into molecular signatures of all body parts (Fig. 1c). Figure 1c shows significant overlaps between the parts of the interna and an almost 10-fold difference in the number of “specific” genes between the interna parts and the externa. Based on the gene expression, the body parts were divided into two clusters (Fig. 1d). The first cluster contained all parts of the interna, while the second cluster contained only two replicates of the externa.

According to differential expression analysis results, the number of over-expressed genes varied from 204 (the main trunk of the interna) to 2224 (the externa) [Tab. S3]. The number of genes with an increased expression in the interna did not exceed 283. Figure 1e shows that (i) only 1 gene was over-expressed both in the externa and in the interna, (ii) the number of “shared” over-expressed genes between different interna parts ≤ 35, (iii) only 6 gene were over-expressed in the whole interna. The externa clustered separately from the interna, and all interna parts were remote from each other (Fig. 1f).

The results of identification of molecular signatures and differentially expressed genes are presented in Supplementary table S3.

Gene Set Enrichment Analysis (GSEA) results and histological studies indicate the presence of germ-like cells in the interna

Figure 2a,b shows Venn diagrams for lists of bioprocesses enriched with genes included in the molecular signatures (Fig. 2a) and having increased expression (Fig. 2b) in the female body parts considered. In contrast to the externa, where many active bioprocesses were associated with development, bioprocesses enriched in the interna were mainly connected with metabolism. Comparative analysis results revealed that 50 bioprocesses were active in at least 2 parts of the female body. There were among them “mitotic cell cycle process” (main trunk and the thoracic part of interna), “cell division” (growing and thoracic part

of interna), “homeostasis of number of cells” (same), “apoptotic signalling pathway” (same), “symbiotic process” (growing and main trunk of interna), “determination of adult lifespan” (same), “immune system process” (the growing part, the main trunk, and thoracic part of interna), and “autophagy” (same).

The bioprocess “germ cell development” was enriched in the main trunk and the growing part of the interna, whereas “female gamete generation” was found only among the lists of enriched bioprocesses in the thoracic interna part (Fig. 2c). At the same time, according to our results, meiosis probably occurred only in the externa (Fig. 2c). The results of histological studies also confirmed the presence of germ-like cells in the central lumen of the rootlets of the interna. Groups of floating small round cells with a high nuclear cytoplasmic ratio were found in the central lumen of the main trunk and peripheral rootlets (Fig. 2d-f). A Nuage body (Fig. 2d-f), which is a marker of germ cells, was present in each cell next to the nucleus.

No common bioprocess was found in different parts of the female body, enriched with over-expressed genes. The genes with over-expression in the externa were involved in bioprocesses associated with cuticle transformation and development of the nervous system. In the growing interna part, over-expressed genes were involved in “gland development,” “response to nutrient levels”, “organic acid transport”, as well as lipid and fatty acid metabolic processes. In addition to various metabolic processes, “determination of adult lifespan” and “intrinsic apoptotic signaling pathway” were also found among a set of enriched bioprocesses for the main trunk. Bioprocesses associated with response to various stimulus (bacterium / oxygen-containing compound / wounding), as well as “interspecies interaction between organisms”, “ion transmembrane transport”, “cellular homeostasis”, and “aging” were classified as “enriched” in the thoracic part of the interna.

All GSEA results are presented in Supplementary table S4.

Hundreds of genes encoding potential excretory / secretory proteins (ESP) were identified

The identification of potential ESP was performed *in silico*. We found 852 "classical" and 282 "non-classical" ESP, which, respectively, have or do not have classical N-terminal signal peptides. Figure 3a,b shows Venn diagrams for the sets of genes encoding “classical” (Fig. 3a) and “non-classical” (Fig. 3b) ESP, respectively, which were included in the molecular signatures of body parts. Approximately 35% (297 / 852) of the genes encoding “classical” ESP had a noticeable expression level in all body parts considered. At the same time, more than a half (143 / 282) of genes encoding “non-classical” ESP had this expression pattern. In both cases, the externa had the largest number of “specific” ESP: 325 and 56 “classical” and “non-classical” ESP, respectively. Dozens of ESP (101 “classical” and 35 “non-classical”) were common for the three considered interna parts.

Both “classical” and “non-classical” ESP were divided into families based on their sequence similarity. For “classical” ESPs, 35 families contained 2-4 proteins, while only 2 “non-classical” ESP were combined into one family. Most (579 / 852) of the “classical” ESP matched the MetazSecKB database, the hits being, e.g., mannose-binding proteins, serine proteinase, and cuticle proteins [Tab. S5]. Only 58 “non-classical”

ESP matched MetazSecKB, of which 35 were “uncharacterized proteins” [Tab. S5]. All ESP were also compared to the NeuroPep database, and only 14 “classical” ESP have hits with it [Tab. S5]. The latter included cerebellin-1, kininogen-1, insulin-like growth factors I and II, neuroparsin-A, nucleobindin-2, and 5 representatives of the Serpin family.

Figure 3c shows bioprocesses enriched with genes encoding “classical” ESP. Among the body-parts-specific bioprocesses were the “molting cycle” and “chitin-based cuticle development” in the externa, “positive regulation of cell communication” and “muscle organ development” in the growing interna part, “immune response” and “regulation of apoptotic signaling pathway” in the main trunk and the thoracic part of interna, respectively. The majority (21 / 34) of enriched bioprocesses were common for two or more body parts considered. For example, “regulated exocytosis” (externa and growing part), “mesoderm development” (the growing part and the main trunk), “cell recognition” (same), “proteolysis” (all body parts considered), “motor neuron axon guidance” (same), “cell adhesion” (same), and “neuron recognition” (externa and thoracic part of interna). The activity of the bioprocesses associated with the involvement of the nervous system is consistent with the fact that trophic rootlets of *P. reticulata* were enlaced by a network of host's neurons marked by a presence of a-tubulin and serotonin (Fig. 3d-f).

All ESP analysis results are presented in Supplementary table S5.

Significant differences between the transcriptome age indices (TAI) of body parts were revealed

Almost all (12618 / 12620) *P. reticulata* genes were distributed across 17 phylostrata, i.e. sets of genes that coalesce to founder genes having a common phylogenetic origin [26] [Tab. S6]. The three largest phylostrata were “Cellular organisms” (32.34%, 4082 genes), “Eukaryota” (22.25%, 2809 genes), and species-specific ones (13.7%, 1729 genes) (Fig. 4a). Less than a hundred genes were assigned to “Ecdysozoa” (50 genes), “Panarthropoda” (44), “Mandibulata” (63 genes), “Crustacea” (28), and “Hexanauplia” (6). Phylostratum “Cirripedia”, which included two barnacles, *A. amphitrite* and *P. reticulata*, consisted of 363 genes.

The phylostratigraphy results were also used for composition analysis of various gene sets (Fig. 4a). More than a half of the genes with expression ≥ 2 TPM in all body parts considered belonged to “Cellular organisms” and “Eukaryota” phylostrata. The proportion of species-specific genes in this set was approximately 7% (490 / 6829). Noticeable differences in the contributions of different phylostrata to the sets of over-expressed genes were found. For example, approximately 31% (88 / 283) of such genes in the thoracic interna part belonged to the species-specific phylostratum. In contrast, in other parts of the body the proportion of species-specific genes from the total number of over-expressed genes did not exceed 16%. A complex phylostratigraphic composition was also revealed for genes encoding both "classical" and "non-classical" ESP. “Cellular organisms” phylostratum made the greatest contribution to the “classical” ESP (32.16%), while the species-specific phylostratum contributed most to the “non-classical” ESP (38.3%).

We divided the phylostrata into two groups: before the divergence of Crustacea (from “Cellular organisms” to “Crustacea”) and after this event (from “Multicrustacea” to “*P. reticulata*”). The relative expression patterns of the phylostrata are shown in figure 4b,c. All phylostrata except the species-specific one had the highest relative expression in the externa, while the highest expression of the species-specific phylostratum was recorded in the thoracic part of the interna. At the same time, 14 out of 17 phylostrata had the least expression in the main trunk of the interna, the remaining three being “Bilateria”, “Pancrustacea”, and “Multicrustacea”.

One metric to quantify transcriptome conservation on a global scale is the Transcriptome Age Index (TAI) [27], which denotes the average transcriptome age throughout the biological process of interest [28]. TAI was measured for each part of the *P. reticulata* female body. The higher the value of TAI, the greater the contribution of the “young” phylostrata. Significant differences between the TAI of different body parts were revealed. The TAI of the thoracic part of the interna was the highest (4.33), whereas the TAI of the growing part of the interna was the lowest (4.21) (Fig. 4d). The partial contribution of the “*P. reticulata*” phylostratum to the TAI of the body part was about approximately 3 time greater than the contribution of any other phylostratum (Fig. 4e).

From the genes with the GO annotation, we extracted top-500 with the largest contribution to the TAI of body parts and performed GSEA for these gene sets (Fig. 5). Among the “specific” bioprocesses were “embryonic organ development” and “male sex differentiation” in the externa, “cell fate commitment involved in the formation of primary germ layer” and “positive regulation of chemotaxis” in the growing part of the interna, “formation of primary germ layer” and “gland morphogenesis” in the main trunk of the interna, and “response to external stimulus” and “cell population proliferation” in its thoracic part. Only 4 bioprocesses were common for all the female body parts considered: “stem cell population maintenance”, “regulation of anatomical structure morphogenesis”, “chitin-based cuticle development”, and “animal organ morphogenesis”. Developmental processes were registered in each of the female body parts studied (Fig. 5).

Phylostratigraphy and evolutionary transcriptomics results are presented in Supplementary tables S6 and S7, respectively.

Discussion

In this study we obtained the first transcriptomes of a rhizocephalan and made a comparative analysis of different body parts of an adult rhizocephalan female. Our results are a step towards understanding the functioning of these highly specialized parasites and the trends of their evolutionary history. Below we discuss 1) how the molecular signatures helped us to verify the functional role of each part of the parasite body; 2) potential excretory/secretory proteins and their putative role in host-parasite interactions; 3) the trends of rhizocephalan evolution derived from the phylostratigraphy and evolutionary transcriptomics results.

One of the main challenges in molecular biological studies of parasitic organisms is to identify and eliminate contamination. In our case this challenge was aggravated by the fact that both the parasite and its host were crustaceans. This means that approaches based on database comparison could be less effective than usual at separating the reads from different sources. For this reason, we chose an MCSC algorithm, which classifies sequences based on the analysis of their properties. Based on the results of a preliminary orthogroup reconstruction (data not shown) and the reconstruction after decontamination and considering that there was an approximately 14% reduction in the number of duplicates of single-copy metazoan orthologues, we can assume that at least part of the signal from the host was removed. In further analysis, we focused on 12,620 protein-coding genes with a noticeable expression level. The results indicate that the reference gene set obtained in our study corresponds to those of other crustacean species in quality and completeness. The results of the analysis of orthogroups revealed that the proteome of *P. reticulata* was more similar to that of *Amphibalanus amphitrite* than to any other crustacean species involved in our study. *Amphibalanus amphitrite* belongs to Thoracica, a sister group of Rhizocephala. Thoracica barnacles are mostly free-living but also very transformed crustaceans. A comparative analysis involving numerous representatives of these two sister taxa may uncover evolutionarily conservative mechanisms of transformation of adult rhizocephalans.

The body of a rhizocephalan female is divided into the externa and an extensive interna, which has different ultrastructural organization in different parts ([29] and Miroliubov unpublished data). Our aim was to make a detailed record of the genome's activity in different parts of the female body. Therefore, we used the soft threshold of 2 TPM as a condition for identifying genes whose expression contributed to the molecular signature of the sample under consideration. The results of the study of gene expression in different parts of the female body showed that: 1) slightly more than a half of the identified protein-coding genes worked in all the examined parts of the body, 2) the lists of genes with an increased expression differed greatly between body parts, 3) the externa always clustered separately from the interna, although the differences were also found between the sites of the interna. These results suggest that the morphological heterogeneity of the female body is reflected in the spatial differences of gene expression (molecular heterogeneity).

However, it should be kept in mind that the molecular signature of the body part is a derivative of the transcriptomes of all cell types included in the body region under consideration. It also depends on the organism's response to various stimuli and conditions (for example, the host's immune response, O_2 level and concentrations of different metabolites in the host's hemolymph) [13, 30]. As a consequence, the more similar the cellular composition of body parts or the set of factors affecting them is, the more similar their molecular signatures will be. In our study all the parts of the interna clustered together and were separate from the externa cluster. At the same time, the differences of the biological replicates of the externas could be explained by the fact that the externa contains embryos at different stages of development. Probably, the signal from the embryos is so strong that even the pooling of samples did not smooth out the differences.

Our results indicate that various processes are at work in different body parts of the female rhizocephalan. Active developmental processes were registered in the externa, which could be expected considering that it contains numerous embryos, while active metabolism processes were recorded in the interna (Fig. 2c). These results are consistent with the classical concepts of the functional role of individual parts of the rhizocephalan body [1, 2].

However, some of our findings are at odds with the classical views. Previous morphological studies have postulated that the ovary is located in the visceral mass of the externa [2]. However, the GSEA revealed that the female germ cells formed in the interna. Moreover, in histological sections we observed some cells, floating in the central lumen of the main trunk, that looked like primary germ cells. Such cells have been described before [31, 32] but the authors assumed them to be stem cells, responsible for the formation of new buds of externas. Based on histological data and the GSEA results, we suggest that these cells are more likely to be female germ cells. We suppose that female germ cells begin and/or continue to form in the interna and then migrate to the externa, where they mature and are fertilized. In our opinion, the ovary of rhizocephalans is diffused within the interna, while the externa serves merely as a brooding chamber.

Our finding of a “diffused” ovary prompts a reconsideration of the phenomenon of rhizocephalan “coloniality”. As noted above, some rhizocephalans form numerous externas. It has been suggested that each externa is a separate reproductive module, and the entire animal has been considered as a colony [2, 31]. However, if the ovary is in fact diffused and scattered across the interna and if numerous externas are merely brooding chambers, the term “coloniality” does not seem suitable for Rhizocephala. This issue calls for further research with the use of molecular and morphological methods.

Regardless of whether a rhizocephalan barnacle is a colony or an individual organism, it has to communicate with the host via special excretory molecules involved in particular bioprocesses [3]. We found that the composition of the excretome / secretome varied in different parts of the female rhizocephalan body and showed the character of the distribution of the secreted substances. For instance, in the externa there seems to be an active secretion of proteins involved in the storage of nutrients in the developing embryos as well as those involved in the moulting cycle. Proteins responsible for muscle development were some of the excretory proteins enriched in the growing part of the interna. These data confirm that the muscular system is formed in a growing tip of the main trunk [29].

Excretory proteins involved in neuron axon guidance were found in both the externa and the interna. While neuron axon guidance in the externa is probably associated with the offspring development, the evidence of this process in the interna is less expected and more interesting. A direct contact between the parasite and the host’s nervous system was shown in this study and in our previous research [3]. We can expect the parasite to emit attracting signals directing the host’s nervous system towards and along the interna. In addition, the excretory proteins involved in cell adhesion could also play an important role in the formation of a neural network around the rootlets. However, it cannot be ruled out that this transcriptome signal comes from the host tissues surrounding the rootlets of the parasite, since it is technically

impossible to completely separate interna from the host's tissues. Nevertheless, an intimate host-parasite interaction is an intriguing phenomenon requiring in-depth research.

The evolution of parasitic barnacles and their interactions with the hosts remains enigmatic [33]. Considering that the species traits are formed by a regulation of activity of a single genome, the understanding of how this genome was transformed in evolution has great value. Here we approached this issue with the help of phylostratigraphy, a method that makes it possible to divide the set of genes of the studied species into groups with a common phylogenetic origin, called "phylostratum" or "phylostrata" in plural. The results of the phylostratigraphy analysis indicate that almost all *P. reticulata* genes were successfully distributed into 17 phylostrata. The third largest phylostratum was species-specific one, which also probably includes genus- and family-specific orthologs. Given the transcriptome obtained in our study is the first reported rhizocephalan transcriptome, it is difficult to determine what percentage of proteins from this phylostratum belongs to each of these categories. Nevertheless, we are confident that a more detailed analysis of this particular phylostratum will allow us to identify the molecular basis of the Rhizocephala-specific biological traits.

It should be noted that we analysed the reference gene set reconstructed based on the transcriptomic data, and the transcripts of poorly expressed or completely silent genes may be absent in our data. However, since we obtained transcriptomes of different regions of the female body and used soft expression level threshold, we may be fairly sure that many genes active in the adult female were represented in the reference gene set.

Our results indicate that evolutionarily younger genes make a relatively large contribution to the signatures of the externa and the thoracic part of the interna. The transcriptional "youth" of the externa could be associated with the fact that the samples contained highly modified males. This idea is supported by the enrichment of "male sex differentiation" bioprocess by genes, with the greatest contribution to TAI by externa. On the contrary, the signature of the thoracic part probably does not have any additional components, and a high TAI of this region may be associated with the evolutionary transformation of the region itself. The transcriptional "oldness" of the growing part of interna is presumably due to the activity of conservative processes, including those associated with the cell cycle.

The GSEA results for genes with the greatest contribution to TAI, both in the interna and in the externa, revealed many developmental processes. One gets the impression that rhizocephalans have an incomplete and endless metamorphosis. Taking into account that the adult female body originates from a fraction of the larval body, we are inclined to agree with the hypothesis suggested by Glenner and Høeg [33]. It postulates that ancestors of rhizocephalans were filter-feeding epibiotic barnacles and the interna of an adult female originates from the part of the larval body homologous to the peduncle of a Goose barnacle, whose metamorphosis went the "wrong" way [33, 34]. Nevertheless, a contemporary species can serve only as a proxy for an ancestor model. More transcriptomic / genomic data on other rhizocephalans are necessary for a reliable reconstruction of the evolutionary history of this unique group of parasites.

To sum up, the first comparative transcriptomic results for rhizocephalans brought us closer to understanding molecular mechanisms underlying the biology of these amazing parasites. We established the molecular and functional heterogeneity of the female body in addition to the known morphological one. A similarity was found between the set of protein-coding genes of the *P. reticulata* and that of a free-living representative of the sister taxon. Both bioinformatic data analysis and histological results indicated the presence of germ cells in the lumen of the interna, which casts doubt on the rhizocephalan coloniality. The molecular basis of the interaction between the nervous system of the host and the parasite's interna was determined. We established the differences between body parts in terms of phylostrata expression and their contribution to molecular signatures. Our results could indicate that rhizocephalans “got stuck in the metamorphosis” even in their reproductive stage. Our study can serve as a basis for further research of the rhizocephalan evolution.

Methods

Sampling

Hermit crabs *Pagurus minutus* Hess, 1865 infected with *Peltogaster reticulata* Shiino, 1943 were collected in the Sea of Japan (Marine Biological Station “Vostok” of the Institute of Marine Biology of the Russian Academy of Sciences) (N: 42.893720, E: 132.732755). All the parasites were adults with fully developed externas.

The infected crabs were dissected in filtered sea water. The parasite was removed from the host’s body cavity and the remains of the host tissues were isolated from the interna. The body of each parasite was divided into 4 parts: 1) the externa, 2) distal part of the main trunk (the growing part), 3) the main part of the interna trunk (the main trunk), 4) the part of interna from the thorax of the host (the thoracic part) (Fig 1). For each body part the pooled sample was prepared, containing material from five parasitic individuals in two biological replicates. The samples were collected into the centrifuge tubes and frozen at -80 °C in IntactRNA (Evrogen, Moscow, Russia) reagent according to the manufacturer`s protocol.

Before RNA isolation, the IntactRNA-fixed samples were rinsed in 0.1M phosphate-buffered saline (PBS). The total RNA was isolated using Quick-RNA™ MiniPrep (R1054, Zymo Research, Irvine, California, USA). The libraries were synthesized using NEBNext Ultra Directional RNA Library Prep Kit for Illumina (E7760, New England BioLabs, Ipswich, Massachusetts, USA). Paired-end sequencing was carried out using Illumina HiSeq 2500 instrument (Illumina, San-Diego, California, USA).

Sampling was conducted in accordance with the European Community Council Directive of November 24, 1986 (86/609/EEC). All possible effort was made to minimize the number of animals used.

Preparation of reads libraries and *de novo* transcriptome assembly

The primary quality control of paired-end reads libraries was manually assessed using FastQC (v0.11.5). The potential sequencing error identification and correction was performed by Karect [35] (v1.0) with

params: –celltype=diploid –matchtype=hamming. Trimmomatic (v0.39) was used for removing the sequencing adaptors (ILLUMINACLIP:$ADAPTERS:2:30:10:2:TRUE), low-quality read regions (SLIDINGWINDOW:4:20 MAXINFO:50:0.8) as well as reads with length less than 25 nucleotides (MINLEN:25). Since libraries preparation and sequencing were performed in the laboratory where researchers also work with human medical samples, we checked the parasite data for the presence of the read pairs with a high identity to the *Homo sapiens* reference transcriptome (GENCODE v.31) using BBTools (v37.02) [URL: http://sourceforge.net/projects/bbmap/].

The prepared libraries were pooled and used for *de novo* reference transcriptome assembly using Trinity [36] (v2.5.1) with k-mer size and required minimal contig length equal to 25 and 200 nucleotides, respectively. The assembled contigs were renamed by adding the four-digit tag of the species, “Pret”, at the beginning of IDs. Isoforms were clustered on all assembled contigs using CDHIT-est [37] (v4.7) and sequence identity threshold equal to 95% (-c 0.95), accurate mode (-g 1), and both +/+ and +/- strands alignments (-r 1). TransRate [38] (v1.0.1) was used for the quality assessment of clustered sequences. Only the contigs classified as “good” by TransRate [38] were included in further analysis.

Removal of potential contamination

The 18S and 28S ribosomal RNA sequences were searched using RNAmmer [39] (v1.2) from the Trinotate pipeline (v3.1.1). The sequences obtained in this way were compared with the NCBI nucleotide database with BLASTn [40] (v2.6.0+) to identify their possible sources.

We used MCSC [41] for removing potential contamination, with “Arthropoda” as the target taxon and the clustering level equal to 5 (32 clusters). The parsing of BAM-files with reads alignment results created by Bowtie2 [42] was performed to extract only reads pairs that mapped to the decontaminated set of contigs.

Quantification of genes expression levels and identification of encoded amino acid sequences

Salmon [43] (v1.0.1) was used for expression level quantification (-l ISF –discardOrphans –seqBias –gcBias –validateMappings). The expression quantification results and the transcripts-to-genes map from Trinity output were provided to library “tximport” for R to obtain expression levels of genes. The tables with both unaveraged TPM values and TPM values averaged between biological replicates were prepared. Only the sequences with expression levels ≥ 1 TPM at least in 1 sample were included in further analysis.

TransDecoder (v5.5.0) [URL: http://transdecoder.github.io] was used for the identification of the amino acid sequences encoded by assembled contigs. Firstly, the long open reading frames with a length ≥ 100 amino acids and products of its translation were found. Secondly, identified proteins were compared with the NCBI non-redundant (DIAMOND BLASTp [44] (v0.9.22.123), e-value = 1e-3) and the PfamA [45] (HMMscan (v3.1b2) [URL: http://hmmer.org]) databases. Thirdly, the comparison results were provided to the TransDecoder to identify the likely coding regions and to obtain the probable set of proteins.

Reference gene set preparation

In our analysis we focused only on genes that successfully passed two filters: 1) noticeable expression level (i.e., the expression is ≥ 1 TPM at least in 1 sample) and 2) encoding of the proteins with a length greater than or equal to 100 amino acid. Only the longest protein and its coding transcript were selected as representatives for each gene and referred to as “reference sets”. The completeness of the protein reference set was evaluated by comparison with the database of single-copy orthologues of Metazoa (odb-9) using BUSCO [46,47] (v3.0.1) (e-values = 1e-3, mode = proteins).

Sequence annotation

For the annotation of the genes, their nucleotide and amino acid sequences were compared with publicly available databases: NCBI nucleotide collection (nt), NCBI non-redundant (nr), and SwissProt [48]. The similarity search was carried out with BLASTn megablast (nt) and the sensitive mode of DIAMOND BLASTp (amino acid databases), with an expected value (e-value) threshold equal to 1e-3 and a limit up to 10000 for the number of description and alignments [49]. The best BLAST hits (BBH) were selected with a custom script.

The potential domain architecture of the proteins was identified using the PfamA database (HMMScan) and custom script. The proteins were also analysed using the eggNOG-mapper web-resource [50] (v2) with default parameters.

Identification of orthogroups

Orthogroups were identified with the use of OMA standalone program (v2.5.0) [51] in three steps. Firstly, the reference proteomes of *Amphibalanus amphitrite* Darwin, 1854 (UP000440578), *Armadillidium nasatum* Budde-Lund, 1885 (UP000326759), *Armadillidium vulgare* Latreille, 1804 (UP000288706), *Daphnia magna* Straus, 1820 (Strain: Xinb3) (UP000076858), *Daphnia pulex* Leydig, 1860 (UP000000305), *Penaeus vannamei* Boone, 1931 (UP000283509), *Portunus trituberculatus* Miers, 1876 (UP000324222), and *Tigriopus californicus* Baker, 1912 (UP000318571) were downloaded from UniProt [48] database. Only the sequences with a length equal to or more than 100 amino acid were analysed. The OMA standalone was run with default parameters with the “bottom-up” algorithm for inference of HOGs, without a phylogenetic tree, but with the identification of 2 Daphnia species as an out-group. Secondly, we reconstructed the phylogenetic tree following the protocol of Dylus et al. [52]. Briefly, using the filter_groups.py provided, we selected OMA groups that included at least 8 of the 9 crustacean species involved in the analysis. Then, using MAFFT [53] (v7.487), multiple protein alignment in each orthogroup was performed (–maxiterate 1000 -localpair). The alignments were concatenated into a supermatrix using the concat_alignments.py. The selection of suitable sites in the supermatrix was carried out using tramAl [54] (-automated1). We used the ProtTest program [55,56] (v3.4.2) to determine the most appropriate sequence evolution model. The phylogenetic tree was reconstructed using the IQ-TREE [57,58] (v2.1.4-beta) with the following parameters: -m LG+I+G+F –seed 12345 -B 1000 –nmax 1000. The consensus tree was rooted by the out-group using the “ape” [59] library for R. Thirdly, the phylogenetic

tree was used when OMA standalone was re-run with default settings. The construction of a heatmap with the number of common OMA groups between the studied species was performed in RStudio using the “ggplot2” (v3.3.5), “pheatmap” (v1.0.12), and “RColorBrewer” (v1.1-2) libraries.

Reference gene set expression analysis

We define the “molecular signature” of a body part as a set of genes with an expression level ≥ 2 Transcripts-Per-Million (TPM) in the body part. The expression threshold value was chosen in accordance with the results of studies by Wagner, Kin, and Lynch, according to which “genes with more than two transcripts per million transcripts (TPM) are highly likely from actively transcribed genes” [60]. If a gene had an expression ≥ 2 TPM in all the body parts, we classified it as “commonly expressed”.

Significant variation of gene expression between samples was detected using “RNentropy” library [61] (v1.2.2) for R. The analysis was carried out using a table with TPM values unaveraged between replicates. We used the corrected global sample specificity test $P < 0.01$ according to the Benjamini-Hochberg method, and local sample specificity test $P < 0.01$.

The overlaps between the molecular signatures and the sets of over-expressed genes were visualized with the InteractiVenn [62].

Multidimensional Scaling

We performed Multidimensional Scaling (MDS) for the molecular signatures and the sets of over-expressed genes. The presence / absence matrices were used as input. It was indicated in the matrix for each gene (row) whether the gene was included in the molecular signature of the body part or had an increased expression in it (“1”) or not (“0”). The optimal number of clusters was determined using the “silhouette” method implemented in the “factoextra” library (v1.0.7) for R. We used the metaMDS function from the “vegan” library (v2.5-7) with the following parameters: distance = "manhattan", try = 100, trymax = 100000, autotransform = FALSE, binary = TRUE, k = the optimal number of clusters. The seed was set to 1234 both when the optimal number of clusters was determined and in MDS. To visualize the results, we used ggscatter function from “ggpubr” (v0.4.0) library for R.

Potential ESP identification and analysis

The *in silico* identification of potential ESP was performed according to the pipelines described by Garg and Ranganathan [63]. Firstly, all proteins from the reference set were analysed with SignalP [64] (v5.0b). Based on the analysis results, the proteins were divided into potential “classical” (SP ≥ 0.5) and “non-classical” (SP < 0.5) ESP. Secondly, the “non-classical” ESP were analysed using SecretomeP [65] (v1.0). Only proteins that had NN-scores ≥ 0.9 and at the same time were predicted not to contain a signal peptide were selected. Thirdly, all potential ESP were scanned for the presence of the mitochondrial transit peptide with TargetP [66] (v2.0). The proteins with this signal were excluded. Fourthly, TMHMM [67] (v2.0c) was used to detect transmembrane domains, and only proteins without them were considered

as potential ESP. Out of these potential ESP, we selected only those that were included in at least one molecular signature. The overlap analysis between ESP sets were performed using InteractiVenn [62].

We run all against all BLASTp searching using DIAMOND [44] (–evalue 1e-3) for both classical and non-classical ESP. Similarity search results were used in Silix [68] (v1.2.11) (-r 0.9), which assigns proteins to putative gene families. For annotation, all ESP were compared against the NeuroPep [69] and MetazSecKB [70] databases using DIAMOND BLASTp with the following params: –sensitive –max-target-seqs 10000 –evalue 1e-3. The best BLAST hits were selected using a custom script.

GSEA

The GSEA using “topGO” library (v2.40.0) for R was performed for 1) whole molecular signatures, 2) sets of over-expressed genes, 3) sets of potential “classical” and “non-classical” ESP. Only the GO terms describing biological processes were considered. We used Fisher`s exact test and extracted from the results (GO terms with P-value <0.01) only the terms including at least 10 significant genes. Redundancy was reduced with “rrvgo” library (v1.0.2) for R. We used the minus log10-transformed p-values as scores, org.Dm.eg.db as database, Relevance as similarity measures methods, and 0.7 as the threshold for reduceSimMatrix. We used the “wordcloud” (v2.6) library for R to build word clouds based on the redundancy reduction results. The more often the parental bioprocess was found in the list of enriched bioprocesses, the larger the word size. Each bioprocess was assigned a colour from the “viridis” (v0.6.1) library palette.

Phylostratigraphy and TAI measuring

The phylostratigraphy analysis of the *P. reticulata* reference protein set was performed using “phylostratr” library [71] (v0.2.1) for R. We used reference proteomes of Crustacea prepared beforehand as well as the prebuilt dataset of prokaryotes, human, and yeast. Similarity search between proteins was carried out with BLASTp [40] (v2.6.0+). The tables with BLAST results in “6” output format were used as input for “phylostratr” for protein distribution between phylostrata: 1) “Cellular organisms”, 2) “Eukaryota”, 3) “Opisthokonta”, 4) “Metazoa”, 5) “Eumetazoa”, 6) “Bilateria”, 7) “Protostomia”, 8) “Ecdysozoa”, 9) “Panarthropoda”, 10) “Arthropoda”, 11) “Mandibulata”, 12) “Pancrustacea”, 13) “Crustacea”, 14) “Multicrustacea”, 15) “Hexanauplia”, 16) “Cirripedia”, and 17) “*Peltogaster reticulata*”.

The phylostratigraphic composition was analysed for *P. reticulata* reference gene set, the set of genes with noticeable (≥ 2 TPM) expression in all the female body parts considered, the sets of overexpressed genes as well as the sets of genes encoding potential “classical” and “non-classical” ESP. The results were visualized using “ggplot2”, “viridis” (v0.6.1), and “reshape” (v0.8.8) libraries for R.

TAI definition was performed for *P. reticulata* body parts using phylostratigraphy results and tables with averaged TPM-values. The analysis was carried out using library “myTAI” [28] (v0.9.3) for R. Genes with an expression level < 2 TPM at all the body parts were excluded. The analysis was carried out on log2(TPM + 1) transformed values. We used the FlatLineTest function to quantify the statistical

significance of the global TAI pattern. For analysis with the use PlotRE and PlotBarRE functions, the phylostrata were divided into groups “before” (phylostrata 1-13), and “after” (phylostrata 14-17) the division of Crustacea.

Using the pMatrix function from “myTAI”, the contributions of genes to TAI of body parts were determined. For each of the body parts, 500 genes with the largest contribution were selected out of the genes with the GO annotation. Further, GSEA for the selected gene sets was performed similarly to GSEA for the molecular signature.

Histology

For histological and light microscopic examination, the dissected internae were fixed with Bouin solution. Paraffin sections 5 μm thick were made using standard histological methods with the help of a Leica RM-2265 microtome and stained with hematoxylin-eosin. The sections were examined under a Leica DM2500 microscope. The photos were taken with a Nikon DS-Fi1 camera and processed with ImageJ software (FiJi [72]).

CLSM

Samples of interna for immune labelling were fixed with 4% paraformaldehyde (PFA; Sigma-Aldrich) in phosphate-buffered saline (PBS; Fluka) at 4 °C for 4 hours and then rinsed three times with PBS. Before the immunocytochemical staining, the fixed material was incubated with PBST (PBS + 0.1 % Triton-X100; Sigma-Aldrich) during the 24 hours at 4 °C. Then the samples were incubated in primary antibodies and anti-acetylated α-tubulin (Sigma Aldrich, Germany, T6793, produced in mice) and anti-serotonin (Sigma Aldrich, Germany, S5545, produced in rabbit) for three days. After incubation the specimens were rinsed in PBS three times and incubated in secondary antibodies anti-mouse IgG CFTM 633 (Sigma Aldrich, Germany, SAB4600138) and anti-rabbit IgG CFTM 488A (Sigma Aldrich, Ger- many, SAB4600030),

The specimens were rinsed with PBS three times and stained with the DAPI nuclei stain (1 μg/ml; Sigma) for 30 min, rinsed in PBS and mounted in DABCO-glycerol. The samples were examined using the confocal laser scanning microscope Leica TCS SP5 in the Resource Center “Microscopy and Microanalysis” of the Research Park of Saint Petersburg State University. The images were processed by ImageJ software (FiJi [72]).

SEM

Specimens for SEM were fixed at 4 °C in 2.5% glutaraldehyde, dehydrated in ethanol series and acetone, critical point-dried in Hitachi critical point dryer HCP- 2, mounted on stubs, coated with platinum with the use of Giko IB-5 Ion coater, and viewed under FEI Quanta 250 scanning electron microscope in “Taxon” Research Resource Center (http://www.ckp-rf.ru/ckp/3038/) of the Zoological Institute of the Russian Academy of Sciences.

Declaration

Acknowledgements

We thank the staff of the Resource Centers “Microscopy and Microanalysis”, “Molecular and Cell Technologies”, and “The Bio-Bank” of the Research Park of St Petersburg State University and “Taxon” Research Resource Center (http://www.ckp-rf.ru/ckp/3038/) of the Zoological Institute of the Russian Academy of Sciences for technical assistance. Data analysis was performed at the Bioinformatics Shared Access Center of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. We are grateful to Dr. Olga Korn and the diving team of the Marine Biological Station “Vostok” of National Scientific Center of Marine Biology, Far East Branch of Russian Academy of Science, for the help with the collection of specimens. We are grateful to Dr. Igor Adameyko and Natalia Lentsman for help with the manuscript preparation. We also express our gratitude to Ilya Borisenko for the help with the management of the research. This study was supported by the Grant No. 21-74-00018 of the Russian Science Foundation, by a grant of the Ministry of Science and Higher Education of the Russian Federation (№ 075-15-2021-1069) and the state laboratory theme of the Zoological Institute 1021051402849-1 (Biodiversity of parasites, life cycles, biology and evolution).

Author contributions

A.A.M. collected and fixed samples. A.A.M. and M.A.N. extract total RNA samples. M.A.N. conducted the bioinformatic analysis. Histology was performed by A.A.M. M.A.N. and A.A.M. wrote the manuscript and prepared the figures. All authors read and approved the final manuscript.

Data Availability

Data supporting the conclusions of this article are included within the article and its additional file. BioProject has been deposited at NCBI under accession PRJNA798055.

Competing interests

The authors declare that they have no competing interests.

References

1. Bresciani, J. & Høeg, J. T. Comparative ultrastructure of the root system in rhizocephalan barnacles (Crustacea: Cirripedia: Rhizocephala). *J. Morphol.* **249**, 9–42 (2001).
2. Høeg, J. T. The biology and life cycle of the Rhizocephala (Cirripedia). *J. Mar. Biol. Assoc. United Kingdom* **75**, 517–550 (1995).
3. Miroliubov, A. *et al.* Specialized structures on the border between rhizocephalan parasites and their host’s nervous system reveal potential sites for host-parasite interactions. *Sci. Rep.* **10**, 1–11 (2020).

4. Lianguzova, A. D., Ilyutkin, S. A., Korn, O. M. & Miroliubov, A. A. Specialised rootlets of *Sacculina pilosella* (Rhizocephala: Sacculinidae) used for interactions with its host's nervous system. *Arthropod Struct. Dev.* **60**, 101009 (2021).
5. Høeg, J. T. & Lützen, J. Life cycle and reproduction in the Cirripedia, Rhizocephala. *Oceanogr. Mar. Biol. an Annu. Rev.* **33**, 427–485 (1995).
6. Lützen, J. & Du, P. T. Three colonial rhizocephalans from mantis shrimps and a crab in Vietnam, including *Pottsia serenei*, new species (Cirripedia: Rhizocephala: Thompsoniidae). *J. Crustac. Biol.* **19**, 902–907 (1999).
7. HØEG, J. T. & LÜTZEN, J. Comparative morphology and phylogeny of the family Thompsoniidae (Cirripedia, Rhizocephala, Akentrogonida), with descriptions of three new genera and seven new species. *Zool. Scr.* **22**, 363–386 (1993).
8. Glenner, H. Cypris metamorphosis, injection and earliest internal development of the Rhizocephalan *Loxothylacus panopaei* (Gissler). Crustacea: Cirripedia: Rhizocephala: Sacculinidae. *J. Morphol.* **249**, 43–75 (2001).
9. Miroliubov, A. A. *et al.* Muscular system in the interna of *Polyascus polygenea* and *Sacculina pilosella* (Cirripedia: Rhizocephala: Sacculinidae). *Invertebr. Zool.* **16**, 48–56 (2019).
10. Alvarez, F., Hines, A. H. & Reaka-Kudla, M. L. The effects of parasitism by the barnacle *Loxothylacus panopaei* (Gissler) (Cirripedia: Rhizocephala) on growth and survival of the host crab *Rhithropanopeus harrisii* (Gould) (Brachyura: Xanthidae). *J. Exp. Mar. Bio. Ecol.* **192**, 221–232 (1995).
11. Vásquez-López, H. Affectation of Swimming Capacity in *Callinectes rathbunae* (Crustacea: Brachyura) Caused by *Loxothylacus Texanus* (Crustacea: Rhizocephala). *Res. J. Fish. Hydrobiol.* **5**, 76–80 (2010).
12. Toscano, B. J., Newsome, B. & Griffen, B. D. Parasite modification of predator functional response. *Oecologia* **175**, 345–352 (2014).
13. Alvarez, F., Alcaraz, G. & Robles, R. Osmoregulatory disturbances induced by the parasitic barnacle *Loxothylacus texanus* (Rhizocephala) in the crab *Callinectes rathbunae* (Portunidae). *J. Exp. Mar. Bio. Ecol.* **278**, 135–140 (2002).
14. Zacher, L. S., Horstmann, L. & Hardy, S. M. A field-based study of metabolites in sacculinized king crabs *Paralithodes camtschaticus* (Tilesius, 1815) and *Lithodes aequispinus* Benedict, 1895 (Decapoda: Anomura: Lithodidae). *J. Crustac. Biol.* **38**, 794–803 (2018).
15. Takahashi, T., Iwashige, A. & Matsuura, S. Behavioral manipulation of the shore crab, *Hemigrapsus sanguineus* by the rhizocephalan barnacle, *Sacculina polygenea*. *Crustacean Research* **26**, 153–161 (1997).
16. Belgrad, B. A. & Griffen, B. D. Rhizocephalan infection modifies host food consumption by reducing host activity levels. *J. Exp. Mar. Bio. Ecol.* **466**, 70–75 (2015).
17. BISHOP, R. K. & CANNON, L. R. G. Morbid behaviour of the commercial sand crab, *Portunus pelagicus* (L.), parasitized by *Sacculina granifera* Boschma, 1973 (Cirripedia: Rhizocephala). *J. Fish Dis.* **2**,

131–144 (1979).

18. Innocenti, G., Pinter, N. & Galil, B. S. Observations on the agonistic behavior of the swimming crab *Charybdis longicollis* Leene infected by the rhizocephalan barnacle *Heterosaccus dollfusi* Boschma. *Can. J. Zool.* **81**, 173–176 (2003).
19. Vázquez-lópez, H., Alvarez, F., Franco, J., Moran, A. & Cházaro, S. Observations on the Behavior of the Dark Crab *Callinectes rathbunae* Contreras Parasitized with the Rhizocephalan *Loxothylacus texanus* Boschma. *Int. J. Zool. Res.* **2**, 344–353 (2006).
20. Larsen, M. H., Høeg, J. T. & Mouritsen, K. N. Influence of infection by *Sacculina carcini* (Cirripedia, Rhizocephala) on consumption rate and prey size selection in the shore crab *Carcinus maenas*. *J. Exp. Mar. Bio. Ecol.* **446**, 209–215 (2013).
21. Fuchs, B. *et al.* Regulation of Polyp-to-Jellyfish Transition in Aurelia aurita. *Curr. Biol.* **24**, 263–273 (2014).
22. Nesterenko, M. *et al.* Molecular signatures of the rediae, cercariae and adult worm stages in the complex life cycles of parasitic flatworms (Psilostomatidae, Trematoda). *Parasit. Vectors* **13**, 1–21 (2020).
23. Torruella, G. *et al.* Global transcriptome analysis of the aphelid *Paraphelidium tribonemae* supports the phagotrophic origin of fungi. *Commun. Biol.* **1**, 1–11 (2018).
24. Almudi, I. *et al.* Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings. *Nat. Commun.* **11**, 1–11 (2020).
25. Wang, J. *et al.* Evolutionary transcriptomics of metazoan biphasic life cycle supports a single intercalation origin of metazoan larvae. *Nat. Ecol. Evol.* **4**, 725–736 (2020).
26. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 531–533 (2007).
27. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–819 (2010).
28. Drost, H.-G., Gabel, A., Liu, J., Quint, M. & Grosse, I. MyTAI: Evolutionary transcriptomics with R. *Bioinformatics* **34**, 1589–1590 (2018).
29. Miroliubov, A. A. Muscular system in interna of *Peltogaster paguri* (Rhizocephala: Peltogastridae). *Arthropod Struct. Dev.* **46**, 230–235 (2017).
30. Shirley, S. M., Shirley, T. C. & Meyers, T. R. Hemolymph responses of Alaskan king crabs to rhizocephalan parasitism. *Can. J. Zool.* **64**, 1774–1781 (1986).
31. Shukalyuk, A., Isaeva, V., Kizilova, E. & Baiborodin, S. Stem cells in the reproductive strategy of colonial rhizocephalan crustaceans (Crustacea: Cirripedia: Rhizocephala). *Invertebr. Reprod. Dev.* **48**, 41–53 (2005).
32. Isaeva, V. V., Shukalyuk, A. I., Trofimova, A. V., Korn, O. M. & Rybakov, A. V. The structure of colonial interna in *Sacculina polygenea* (Crustacea: Cirripedia: Rhizocephala). *Crustacean Research* **30**, 133–146 (2001).

33. Glenner, H. & Høeg, J. T. A Scenario for the Evolution of the Rhizocephala. in *Modern Approaches to the Study of Crustacea* (eds. Escobar-Briones, E. & Alvarez, F.) 301–310 (Springer, Boston, MA, 2002). doi:10.1007/978-1-4615-0761-1
34. Glenner, H. & Hebsgaard, M. B. Phylogeny and evolution of life history strategies of the Parasitic Barnacles (Crustacea, Cirripedia, Rhizocephala). *Mol. Phylogenet. Evol.* **41**, 528–538 (2006).
35. Allam, A., Kalnis, P. & Solovyev, V. Karect : accurate correction of substitution , insertion and deletion errors for next-generation sequencing data. *Bioinformatics* **31**, 3421–3428 (2015).
36. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
37. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
38. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate : reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
39. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
40. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
41. Lafond-Lapalme, J., Duceppe, M. O., Wang, S., Moffett, P. & Mimee, B. A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm. *Bioinformatics* **33**, 1293–1300 (2017).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
43. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
44. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
45. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
46. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V & Zdobnov, E. M. BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
47. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2017).
48. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
49. Shah, N., Nute, M. G., Warnow, T. & Pop, M. Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics* **35**, 1613–1614 (2019).

50. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation , Orthology Assignments , and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **msab293**, (2021).
51. Altenhoff, A. M. *et al.* OMA standalone: Orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).
52. Dylus, D. *et al.* How to build phylogenetic species trees with OMA. *F1000Research* **9**, 511 (2020).
53. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
54. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
55. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
56. Guindon, S. & Gascuel, O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52**, 696–704 (2003).
57. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
58. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
59. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
60. Wagner, G. P., Kin, K. & Lynch, V. J. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci.* **132**, 159–164 (2013).
61. Zambelli, F. *et al.* RNentropy: An entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Res.* **46**, e46 (2018).
62. Heberle, H., Meirelles, V. G., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16**, 1–7 (2015).
63. Garg, G. & Ranganathan, S. *In silico* secretome analysis approach for next generation sequencing transcriptomic data. *BMC Genomics* **12**, 1–10 (2011).
64. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
65. Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G. & Brunak, S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* **17**, 349–356 (2004).
66. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
67. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

68. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 1–9 (2011).
69. Wang, Y. *et al.* NeuroPep: A comprehensive resource of neuropeptides. *Database* **2015**, 1–9 (2015).
70. Meinken, J., Walker, G., Cooper, C. R. & Min, X. J. MetazSecKB: The human and animal secretome and subcellular proteome knowledgebase. *Database* **2015**, 1–14 (2015).
71. Arendsee, Z. *et al.* Phylostratr: A framework for phylostratigraphy. *Bioinformatics* **35**, 3617–3627 (2019).
72. Schindelin, J. *et al.* Fiji : an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

Figures

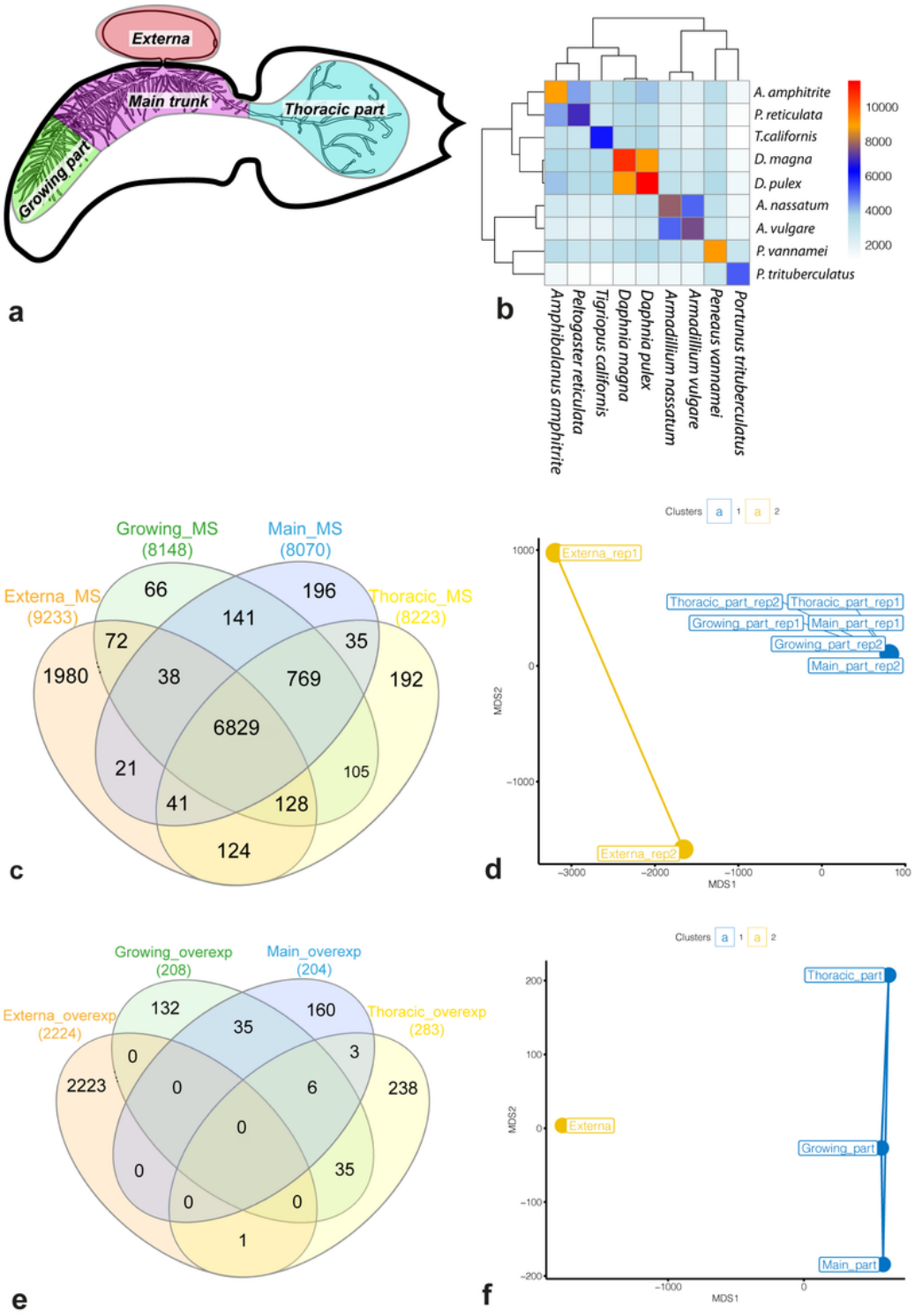

Figure 1

Molecular signatures of *Peltogaster reticulata* female. (a) Generalized scheme of *P. reticulata* female in the host. Colour sectors indicate the body parts examined in our study: externa (red), growing part of interna (green), main trunk of interna (purple), and thoracic part (blue). (b) The number of common OMA groups. The colour key on heatmap shows the number of shared OMA groups between species. (c, e) Venn diagram for a set of genes either included in molecular signatures of the body parts (c) or over-

expressed in the body parts (e). (d, f) Multidimensional Scaling (MDS) plots for molecular signatures (d) and sets of over-expressed genes (f). Different clusters in MDS plots are marked with colours. Abbreviations: MS/overexp – the molecular signature or set of over-expressed genes for the body part, respectively; rep1/2 – biological replication identifier.

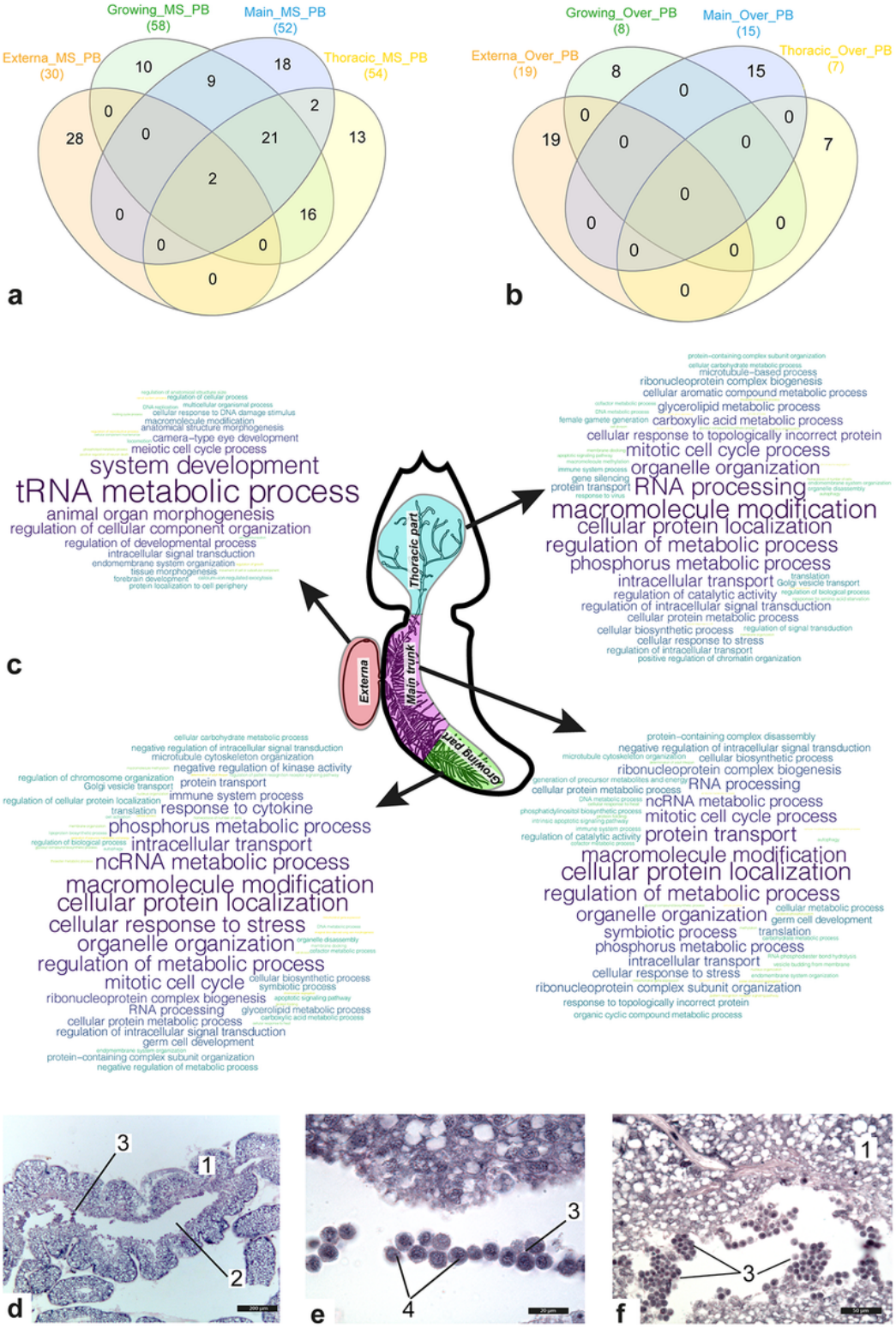

Figure 2

Gene Set Enrichment Analysis (GSEA) results for the gene sets. (a, b) Venn diagram for sets of parental bioprocesses enriched with either genes included in the molecular signature of the body part (a) or genes over-expressed in body part considered (b). (c) Clouds of enriched parental bioprocesses for the body parts. The more often the bioprocess was found in the list of enriched bioprocesses, the larger the word size. Abbreviations: MS/Over – bioprocesses enriched with genes included either in molecular signature or in sets of over-expressed genes; PB – parental bioprocesses. (d-f) Histological sections of the main trunk of *P. reticulata* 1 - the wall of the main trunk; 2 - central lumen; 3 - groups of the floating cells; 4 - Nuage body. Scale bars: d - 200μm, e - 20μm, f - 50μm.

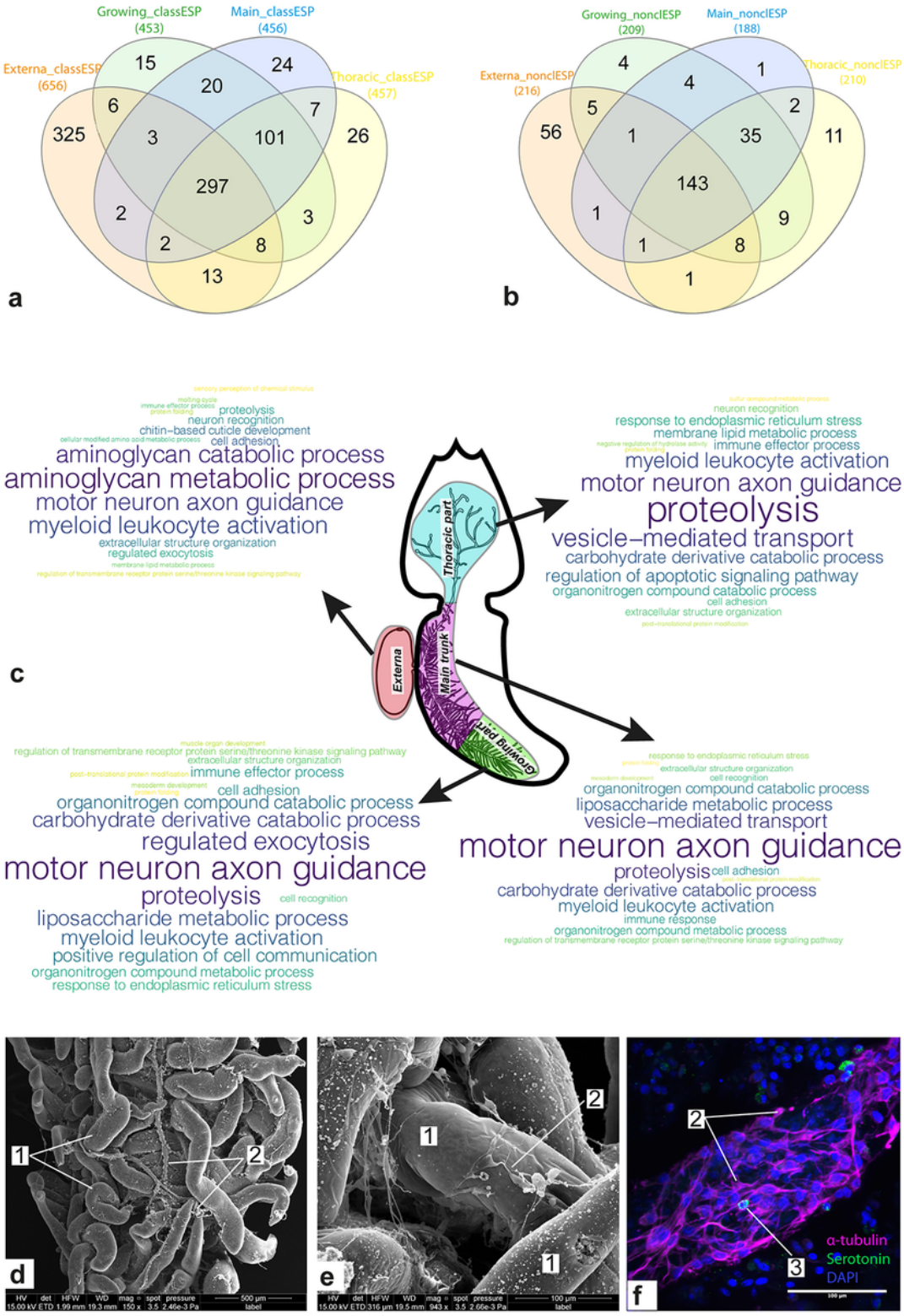

Figure 3

Gene Set Enrichment Analysis (GSEA) results for the identified sets of excretory/secretory (ES) proteins. (a, b) Venn diagram for sets of potential “classical” (a) and “non-classical” (b) ES proteins encoded by genes from molecular signatures of the body parts. (c) Clouds of parental bioprocesses enriched by genes encoding “classical” ES. The more often the bioprocess was found in the list of enriched bioprocesses, the larger the word size. (d-e) SEM photos of the interna of *P. reticulata*, 1 - rootlets; 2 – host

tissues enlacing rootlets. (f) CLSM photo of the interna of *P. reticulata*, scale bar 100μm, 2 - host tissues stained with antibodies against α-tubulin; 4 - cells stained with antibodies against serotonin. Abbreviations: class/nonclassES – “classical” and “non-classical” ES proteins, respectively.

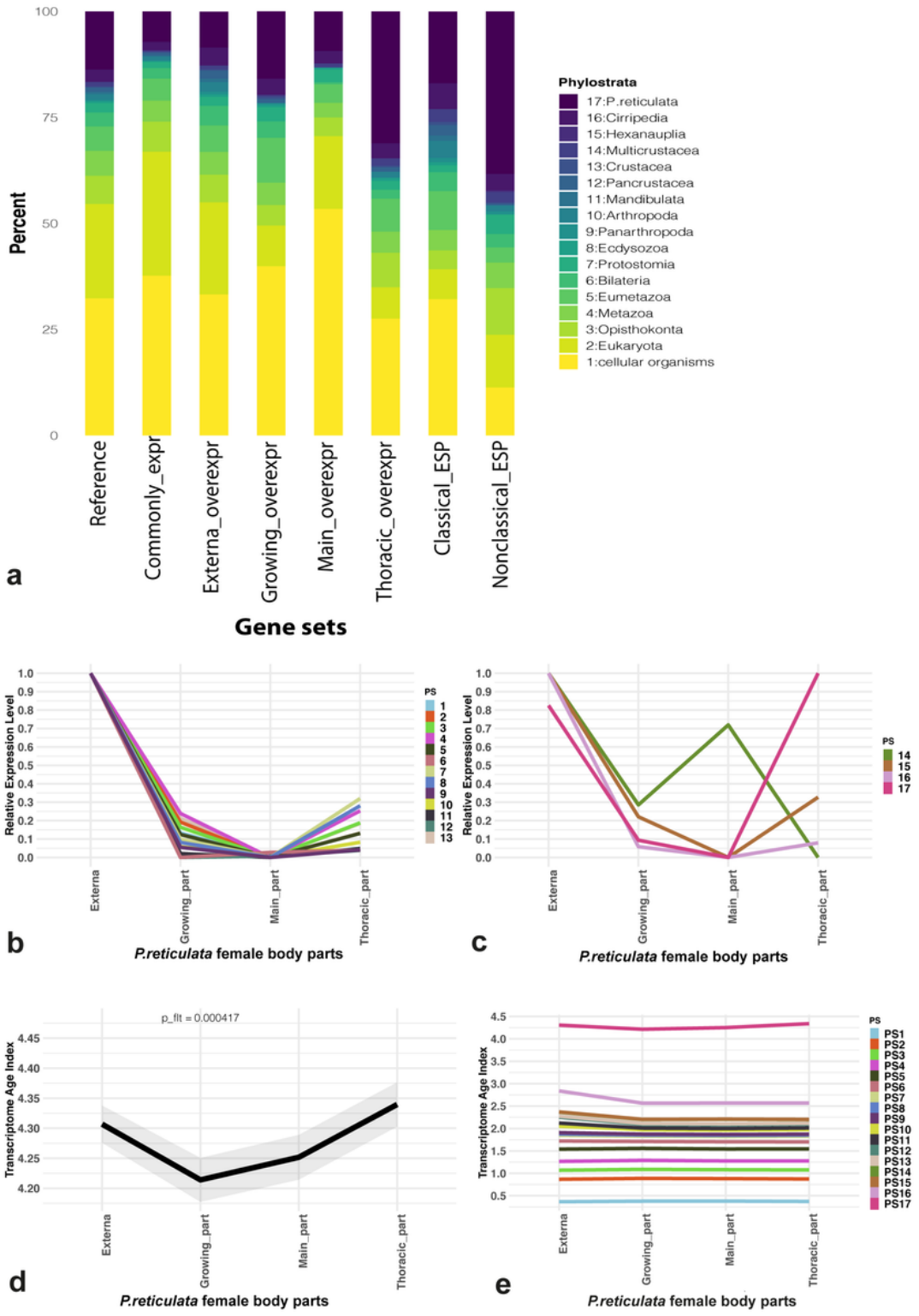

Figure 4

Phylostratigraphy and evolutionary transcriptomics results. (a) Phylostratigraphic composition analysis results for *P. reticulata* reference gene set ("reference"), set of genes with noticeable (≥ 2 Transcripts-Per-Million) expression in all female body parts considered ("common expr"), sets of genes over-expressed (externa/growing/main trunk/thoracic_overexpr) and sets of genes encoding potential "classical" ("classical_ESP") and "non-classical" ("nonclassical_ESP") excretory/secretory proteins (ESP). (b, c) Relative mean expression levels of phylostrata which occurred "before" (b) or "after" (c) division of Crustacea. (d) Transcriptome Age Indices (TAI) variation for female body parts considered. A lower TAI value describes an "older" transcriptome, whereas a higher TAI denotes a "younger" one. (e) The cumulative phylostrata contribution to the final (global) TAI profile.

Figure 5

Gene Set Enrichment Analysis (GSEA) results for top-500 annotated genes with the largest contribution to Transcriptome Age Index (TAI). Clouds of parental bioprocesses enriched with top-500 genes with both Gene Ontology annotation and large contribution to TAI of the body part under consideration. The more often the bioprocess was found in the list of enriched bioprocesses, the larger the word size.

Supplementary Files

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

- SupplementarytableS1.xlsx
- SupplementarytableS2.xlsx
- SupplementarytableS3.xlsx
- SupplementarytableS4.xlsx
- SupplementarytableS5.xlsx
- SupplementarytableS6.xlsx
- SupplementarytableS7.xlsx
- FigS1.pdf