The genome of the Pacific oyster *Crassostrea gigas* brings new insights on the massive expansion of the C1q gene family in Bivalvia

Marco Gerdol [a], Paola Venier [b], Alberto Pallavicini [a,*]

[a] *Department of Life Sciences, University of Trieste, Via Licio Giorgieri 5, 34127 Trieste (TS), Italy*
[b] *Deparment of Biology, University of Padova, Via Ugo Bassi 58/B, 35121 Padova (PD), Italy*

ARTICLE INFO

Accepted 6 November 2014

Keywords:
C1q
Crassostrea gigas
Bivalvia
Innate immunity
Lectin-like

ABSTRACT

C1q domain-containing (C1qDC) proteins are regarded as important players in the innate immunity of bivalve mollusks and other invertebrates and their highly adaptive binding properties indicate them as efficient pathogen recognition molecules. Although experimental studies support this view, the molecular data available at the present time are not sufficient to fully explain the great molecular diversification of this family, present in bivalves with hundreds of C1q coding genes.

Taking advantage of the fully sequenced genome of the Pacific oyster *Crassostrea gigas* and more than 100 transcriptomic datasets, we: (i) re-annotated the oyster C1qDC loci, thus identifying the correct genomic organization of 337 C1qDC genes, (ii) explored the expression pattern of oyster C1qDC genes in diverse developmental stages and adult tissues of unchallenged and experimentally treated animals; (iii) investigated the expansion of the C1qDC gene family in all major bivalve subclasses.

Overall, we provide a broad description of the functionally relevant features of oyster C1qDC genes, their comparative expression levels and new evidence confirming that a gene family expansion event has occurred during the course of Bivalve evolution, leading to the diversification of hundreds of different C1qDC genes in both the Pteriomorphia and Heterodonta subclasses.

1. Introduction

The C1q domain was originally identified as the C-terminal domain of the three chains composing the complement C1q complex (Kishore and Reid, 1999). Structurally similar to the tumor necrosis factor domain (Shapiro and Scherer, 1998), C1q is a globular domain with remarkable ligand binding properties which has been involved in the activation of the classical complement pathway and in other functions such as apoptotic cell clearance, bacteria recognition, cell adhesion and cell growth modulation (Gaboriaud et al., 2003; Ghebrehiwet et al., 2012; Kishore et al., 2004). Several non-complement molecules, collectively named C1q domain containing (C1qDC) proteins have been discovered. They are usually characterized by a signal peptide, occasionally followed by a central collagen-like region involved in oligomerization, and a C-terminal C1q domain (Ghai et al., 2007). Changes in key amino acids, length of the collagen-like region and the association with other domains are responsible of the diversification of the C1qDC protein family which includes 31 members in humans (Tom Tang et al., 2005).

The lectin-like features of C1qDC proteins were recognized in mollusks only in 2004 when a sialic acid-binding lectin was identified in the snail *Cepaea hortensis* (Gerlach et al., 2004). Later, further evidence for a lectin-like role useful to pathogen recognition and clearance was reported in other bivalve species, including *Argopecten irradians, Azumapecten farreri, Crassostrea hongkongensis, Crassostrea ariakensis, Ruditapes philippinarum, Solen grandis, Mytilus coruscus* and *Mytilus galloprovincialis* (Allam et al., 2014; Gestal et al., 2010; He et al., 2011; Li et al., 2011; Liu et al., 2014a; Wang et al., 2012; Xu et al., 2012; Yang et al., 2012; Zhang et al., 2008). The range of pathogen recognition molecular patterns (PAMPs) possibly recognized by the globular C1q domain in bivalves seems to be very broad and includes Gram-positive and Gram-negative bacteria, Rickettsia-like organisms, fungi, protists and even metazoan parasites (Kong et al., 2010; McDowell et al., 2014; Morga et al., 2012; Perrigault et al., 2009; Prado-Alvarez et al., 2009; Taris et al., 2009). The abundance of C1qDC transcripts as well as C-type lectin and fibrinogen-related (FREPs) transcripts in *M. galloprovincialis* supports their role as pathogen recognition receptors (PRRs) (Venier et al., 2011). Nevertheless, factors other than PAMPs are somehow able to trigger the expression of C1qDC genes, such as the exposure to nanoparticles, heavy metals and benzo(a)pyrene (Gomes et al., 2013; Liu et al., 2014a, 2014b; Maria et al., 2013). C1qDC transcripts have been

Abbreviations: C1qDC, C1q domain-containing; FREPs, fibrinogen-related proteins; ORF, Open Reading Frame; NGS, next generation sequencing; PAMP, pathogen associated molecular pattern; PRR, pathogen recognition receptor; SRA, Sequence Read Archive.

* Corresponding author. Department of Life Sciences, University of Trieste, Via Licio Giorgieri 5, 34127 Trieste (TS), Italy. Tel.: +39 0405588736; fax: +39 0405582452.
E-mail address: pallavic@units.it (A. Pallavicini).

detected in bivalve hemocytes (Gestal et al., 2010; Liu et al., 2014a; Oliveri et al., 2014) and a C1qDC protein was reported as the most abundant component of the extrapallial fluid in *Mytilus edulis* (Hattan et al., 2001; Yin et al., 2005). Moreover, C1qDC sequences have been identified as highly expressed in the mantle (Liu et al., 2007), digestive gland (Kong et al., 2010; Wang et al., 2012) and multiple tissues (Li et al., 2011; Yang et al., 2012).

We have previously reported several hemocyte-specific C1qDC transcripts in *M. galloprovincialis*, with other members of this family highly expressed in gills, digestive gland and posterior adductor muscle of unchallenged mussels (Gerdol et al., 2011). Altogether, these data point out that a large number of C1qDC protein precursors are constitutively expressed in various tissues.

Following Sanger sequencing of the *M. galloprovincialis* transcriptome, we described at least 168 distinct C1qDC transcript sequences, but the resolution power of next generation sequencing (NGS) later suggested much larger amounts, since 524 and 232 C1qDC transcripts have been reported in *M. edulis* and in *M. galloprovincialis*, respectively (Gerdol et al., 2014; Philipp et al., 2012). The expansion of the C1qDC gene family is not restricted to *Mytilus* spp., as briefly outlined in the oyster genome paper and in the Crassostrea virginica transcriptome analysis (Zhang et al., 2012, 2014). Based on a comparative transcriptomics analysis, we have hypothesized that a massive expansion event occurred in the class Bivalvia independently from the establishment of a large complement of C1qDC genes in the Chordates lineage (Gerdol et al., 2011). The increasing accessibility of RNA-seq datasets from non-model organisms and the recent release of the *Crassostrea gigas, Pinctada fucata* and *M. galloprovincialis* draft genomes (Nguyen et al., 2014; Takeuchi et al., 2012; Zhang et al., 2012) are leading to the explosion of bivalve-omics, making deeper investigations finally possible (Suárez-Ulloa et al., 2013). Taking advantage of the valuable oyster genome data, we have investigated the expansion of the C1qDC gene family in bivalves, their functional and structural diversification and their expression levels during development and in adult tissues.

2. Materials and methods

2.1. Data sources

The fully sequenced and annotated genome of the Pacific oyster *Crassostrea gigas* (Zhang et al., 2012) was downloaded from EnsemblMetazoa. The assembly version used for the analyses was the latest released oyster_v9 (GCA_000297895.1). All RNA-seq datasets available at the NCBI Sequence Read Archive (SRA) database for *C. gigas* were also downloaded. The complete list of these SRA datasets is provided in Supplementary Appendix S1, Table S1. Sequencing reads were imported in the CLC Genomics Workbench v.7.0.4 (CLC Bio, Aarhus, Denmark) and processed as follows. Reads were trimmed according to quality scores (the quality threshold was set at 0.05) and terminal ambiguous nucleotides were removed. Following the trimming procedure, all the reads shorter than 40 base pairs were discarded.

The trimmed reads were used to produce a *de novo* assembly using two different softwares. First, we applied the *de novo* assembly tool of the CLC Genomics Workbench v.7.0.4, setting the graph parameters to "automatic word size" and "automatic bubble size". The minimum contig length was set at 200 base pairs and, due to the presence of paired-end reads, scaffolding was permitted. Second, we performed a *de novo* assembly using Trinity (release 20140413) with default parameters (Grabherr et al., 2011). The minimum allowed contig length was 200 base pairs. We chose to use two independent *de novo* assembly methods due to their peculiar characteristics: Trinity is largely reported as the most efficient *de novo* assembler for the detection of alternatively spliced isoforms and paralogous genes discrimination whereas the CLC Genomics Workbench assembler usually produces less redundant full-length contigs also in these situations.

2.2. Identification and characterization of C1qDC genes

The strategy applied to the identification and characterization of oyster C1qDC genes is summarized in Fig. 1. First, putative annotated C1qDC genes were identified from EnsemblMetazoa based on the presence of the Interpro IPR001073 signature. The transcriptomic contigs obtained with the two *de novo* assembly methods were separately subjected to TransDecoder (http://transdecoder.sourceforge.net) to predict the encoded proteins, whose minimum sequence length was set at 100 amino acids. Predicted proteins were scanned for the presence of the IPR001073 C1q domain with InterProScan v. 5.4–47.0 (Zdobnov and Apweiler, 2001).

Therefore, three sequence datasets were created: (a) putative C1qDC genes annotated in the oyster genome; (b) putative C1qDC transcripts identified in the CLC Genomics Workbench *de novo* assembly; (c) putative C1qDC transcripts identified in the Trinity assembly.

The assembled contigs were used as a query in BLASTn (Altschul et al., 1990) to identify their genomic location and annotation as oyster genes (a). Matches were identified using an e-value threshold of 1×10^{-50} and an identity threshold of 95%. Additional genomic locations showing significant similarity with the putative transcripts in the datasets (b) and (c) were also identified using the same e-value threshold settings but no identity threshold, therefore permitting both new gene predictions and homology-based identification of the correct intron/exon organization of genes lacking a perfect match.

Matching contigs were then aligned to the corresponding genomic locations with MUSCLE (Edgar, 2004) to allow the correct identification of exon boundaries, and also to verify and add oyster novel C1qDC genes annotations whenever needed. We only considered Open Reading Frames (ORFs), from the initial ATG to the STOP codon (5′ and 3′ UTR regions were disregarded due to the high sequence divergence of paralogous genes within these regions). Finally, contigs encoding full-length proteins devoid of any significant match in the oyster genome, likely encoded in genomic regions constituting gaps of the sequenced *C. gigas* genome, were added to a new list of "orphan transcripts".

We further processed only genes which were fully confirmed by transcriptomic data and marked all the remaining genomic sites as putative C1q loci which were later confirmed by an Hidden Markov Model scan (see section 2.5): namely, full genes whose complete organization could not be inferred by RNA-seq data, incomplete genes interrupted by "N-stretches" in the assembly, partial genes overlapping scaffold edges and pseudogenes. Full genes for C1qDC proteins were named with the same scheme previously used for *M. galloprovincialis* (Gerdol et al., 2011). Therefore, oyster genes were named "CgC1qX", where X is a progressive number.

All the gene sequences, and the corresponding annotations (included in a Generic Feature Format file) and genomic scaffold IDs are available as Supplementary material (Supplementary Appendix S1, Table S2 and Supplementary Appendices S2 and S3).

2.3. Characterization of predicted C1q proteins

Predicted oyster C1qDC proteins were characterized as follows. The presence of a signal peptide was detected with SignalP v. 4.1 (Nielsen et al., 1997) and discriminated from N-terminal transmembrane regions with Phobius (Käll et al., 2004). Sequences were scanned for the presence of additional transmembrane regions with TMHMM v. 2.0 (Krogh et al., 2001). Coiled-coil regions were identified and categorized as parallel/antiparallel dimers, trimers or tetramers with LOGICOIL (Vincent et al., 2013), using a MARCOIL

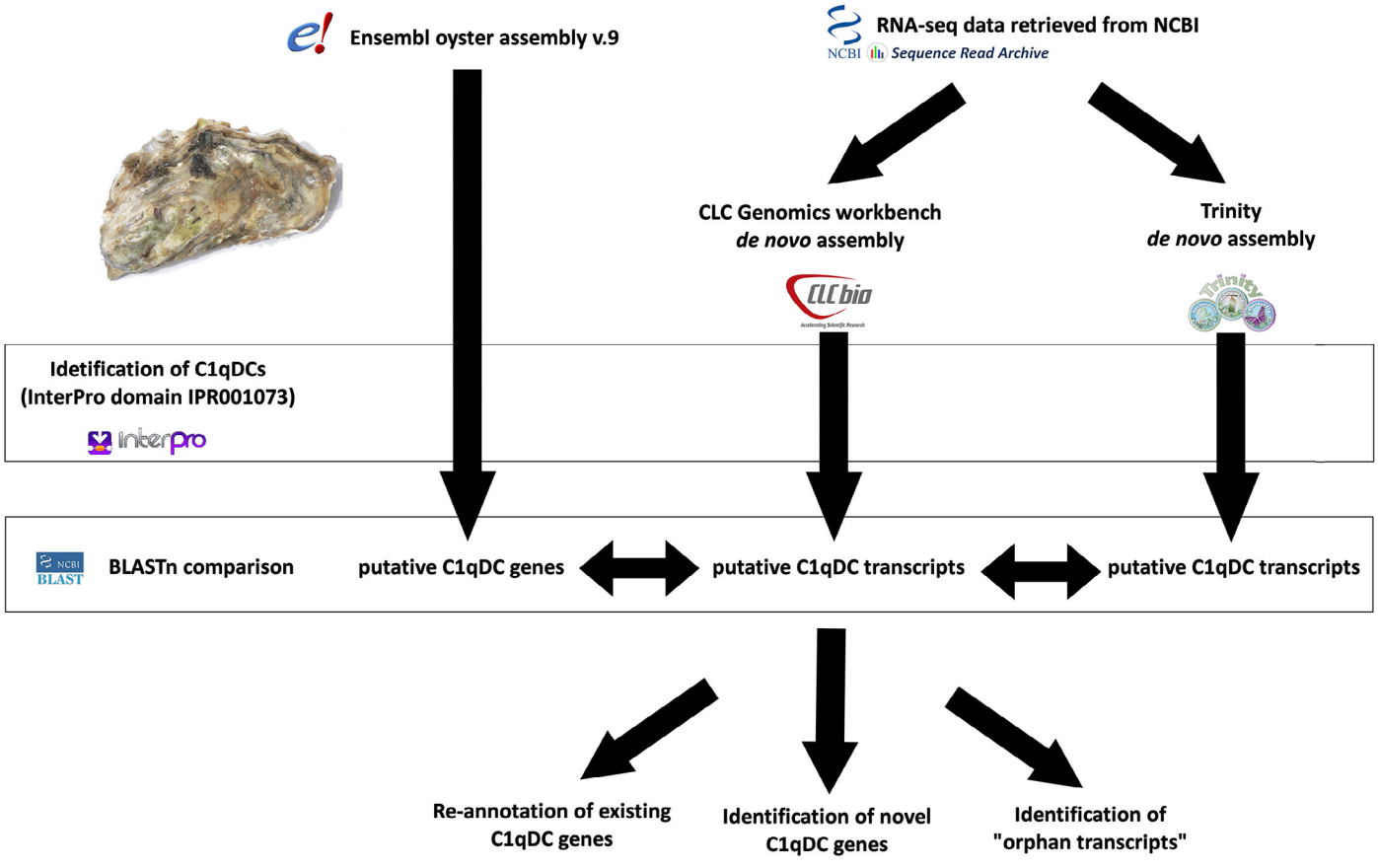

Fig. 1. Strategy used for the identification and re-annotation of oyster C1qDC genes. See section 2.2 for details.

threshold value of 50, and associated leucine-zippers were predicted with 2ZIP (Bornberg-Bauer et al., 1998). The presence and position of functional domains was assessed with InterProScan v. 5.4–47.0 (Zdobnov and Apweiler, 2001).

Tridimensional structures of selected oyster C1qDC proteins were modeled by homology with Phyre2, using a normal mode run (Kelley and Sternberg, 2009).

2.4. Digital gene expression analysis

After trimming, the RNA-seq datasets retrieved from SRA (Supplementary Appendix S1, Table S1) were individually mapped to the annotated C1qDC genes and orphan transcripts with the RNA-seq mapping tool of the CLC Genomics Workbench. Fifty-two highly expressed housekeeping genes (Supplementary Appendix S1, Table S3) were added to ensure a uniform mapping rate of all samples and to permit an accurate calculation of expression values. *Length and similarity fractions* were set to 0.75 and 0.95, respectively, whereas *mismatch/insertion/deletion penalties* were set to 3/3/3. The total number of reads matching exon regions only (*total exon reads*) were counted and used to calculate digital expression values as RPKM (Reads Per Kilobase per Million mapped reads) (Mortazavi et al., 2008).

Over 100 different RNA-seq datasets were analyzed, including digestive gland and gills subject to different experimental challenges. We could consider digestive gland from oysters challenged with heavy metals (Zn, Cd, Cu, Hg, Pb and Zn+Cd) and gills from oysters challenged with heavy metals (Zn, Cd, Cu, Hg, Pb and Zn+Cd), bacterial pathogens (*Vibrio anguillarum, V. tubiashii, V. aestuarianus*, two strains of *V. alginolyticus* and *Micrococcus lysodeikticus*), salinity (from 5‰ to 40‰), temperature (from 5° to 35°) and by exposure to air (up to 11 days). As the RNA-seq experiments we analyzed have been produced by other authors, we do not provide here experimental details, which can be retrieved from the NCBI SRA database (accession IDs of each RNA-seq experiment are provided in Supplementary Appendix S1, Table S1) and from the supplementary online material of the oyster genome paper (Zhang et al., 2012).

In order to evaluate the specific expression of C1qDC genes in adult oyster tissues we selected hemocytes, digestive gland, gills and mantle. In detail, we considered all genes reaching an average RPKM value of at least 30 in one of the mentioned tissues, the rates of normalized expression values were calculated for each of the possible pairwise comparisons and the genes showing an expression value higher than 4 times compared to the other three tissues were classified as tissue-specific. RPKM values were modified by square root transformation and subsequently used to create an expression heat map, using Euclidean distance as a similarity metric and average linkage as a clustering method.

Clusters of co-regulated genes were identified with the *K-means/K-medoids feature clustering tool* of the CLC Genomics Workbench. We applied both algorithms with the Euclidean distance used as a distance metric, setting the number of partitions to 30. Accordingly, a subset of strongly supported clusters was obtained.

2.5. Comparative genomics analysis

The RNA-seq transcriptome data available from bivalve mollusks were downloaded as raw data from SRA. The full list of the bivalve species analyzed is shown in Supplementary Appendix S1, Table S4. Following the trimming procedure by length and quality (see section 2.1 for assembly and trimming parameters), each dataset of Ilumina and 454 Life Sciences reads was independently *de novo*

assembled with the CLC Genomics Workbench. Assembled transcriptomes were translated into virtual proteins with TransDecoder (http://transdecoder.sourceforge.net), setting the minimum protein length to 100 amino acids. The proteins predicted from the genomes of the pearl oyster *P. fucata* and the gastropods *Lottia gigantea* and *Aplysia californica* were also used for the analysis in these species.

Based on the presence of the InterPro domain IPR001073, predicted C1qDC proteins were identified from each species with InterProScan v. 5.4–47.0 (Zdobnov and Apweiler, 2001). The relative abundance of C1qDCs was calculated as the rate between the proteins identified and the total number of TransDecoder predictions.

The presumptive total number of C1qDC loci in the genomes of *C. gigas, P. fucata* and *M. galloprovincialis* was also calculated as follows. Genomic scaffolds were first translated into the six possible reading frames with the EMBOSS Transeq tool (Rice et al., 2000) and the resulting translations were scanned with HMMER (Finn et al., 2011) for the presence of the C1q signature with an e-value threshold of 1×10^{-5}. As the globular C1q domain is most frequently encoded within a single exon and genes with multiple C1q domains are rare, the number of positive hits was considered as a rough indication of the number of genomic C1qDC loci.

3. Results and discussion

3.1. Re-annotation of C1qDC genes in the oyster genome

The annotated draft genome provided by Zhang and colleagues (Zhang et al., 2012) included a total of 322 C1qDC genes. Upon manual re-annotation, we confirmed only 54 accurate gene predictions, we modified 177 gene annotations and disregarded 91 genes that could not be reconstructed to their full length due to one of the following reasons: point mutations, insertion or deletions resulting in the premature termination of the ORF (48% of the cases); insufficient RNA-seq data (42% of the cases); ORF interruption by a N-stretch of by the scaffold edge (10% of the cases). While C1qDC loci falling in the two latter categories could still encode full-length functional proteins, those pertaining to the first one could be regarded as C1q pseudogenes (in the assumption that sequencing or assembly errors are not responsible of ORF terminations). We also identified 45 novel C1q complete loci missing from the Ensembl oyster genome annotation, bringing the total number of annotated and confirmed complete C1qDC genes to 276. In addition, in the *de novo* transcriptome assembly, we found 61 "orphan transcripts", lacking any significant match with genomic scaffolds. Overall, we therefore report the full length sequence of 337 C1qDC proteins which could be inferred from genomic sequence, from RNA-seq data, or from both.

Even though such number of orphan transcripts may seem quite high, the released oyster genome assembly is still rather fragmented, with over 7500 scaffolds and the longest 14% scaffolds covering more than 90% of the genomic sequences (Zhang et al., 2012). Therefore, the presence of almost 2000 scaffolds shorter than 1 Kb suggests that relevant assembly gaps may still exist.

The applied strategy of gene prediction based on a combination of homology-based, *ab initio* and RNA-seq methods (Zhang et al., 2012) could be scarcely effective in the specific case of C1qDC loci, thus explaining the low number of correctly annotated C1qDc genes. In detail: (i) oyster C1qDC genes result from a gene family expansion event which likely occurred selectively in bivalves (Gerdol et al., 2011); (ii) a coiled-coil domain N-terminal to the C1q domain is present in most oyster C1qDC genes (see section 3.3) and this feature, unique to bivalves, makes the prediction of the N-termini of oyster C1q by homology problematic.

Following rigorous criteria of genome annotation, stringent mapping thresholds were applied by Zhang and colleagues to obtain RNA-seq based predictions. However, stringent thresholds could at least partially explain the missing or incomplete annotation of many C1qDC genes such as several oyster C1qDCs showing poor expression levels in different tissues (see section 3.4).

3.2. Structure and genomic organization of C1qDC genes

We found the length of C1qDC genes to be very variable, from about 600 base pairs to over 10 Kb when considering the distance between the initial ATG and the final STOP codon. The longest re-annotated C1qDC gene is CgC1q184 (13,746 bp).

The number and position of introns is also extremely variable. While the presence of two or three exons within the CDS is the most frequent situation, being observed in 75.5% and 19.4% of the cases respectively, we detected up to 12 introns, particularly in genes encoding proteins with multiple C1q domains (see Fig. 2, panel A). A single case of an intronless gene was observed (CgC1q140). We also report the variable position of splice sites in the N-terminal region, as they can be present within the signal peptide, immediately after it or even at several hundred base of distance. Coiled-coil domains can be embedded within the first exon together with the signal peptide or they can be entirely encoded in the following ones, but in no case were they associated in the same exon with the C1q domain. As a matter of fact, C1q domains are encoded by a single exon whose splice acceptor site is found close to the N-terminal end of the C1q domain itself. Occasionally, this C1q exon is split in half by an additional intron (see Fig. 2, panel A).

Concerning their genomic organization, C1q genes are often associated in clusters. In most cases, these clusters are likely the result of gene duplication events because of the high conservation of their CDS, even though intron size can considerably vary among paralogous genes. As mentioned above, C1q pseudogenes are rather numerous and often appear within C1qDC clusters, as relicts which have lost their protein coding potential due to mutations or exon loss. An example of a 160 Kb long C1qDC cluster including 6 fully functional C1qDC genes and 3 pseudogenes is shown in Fig. 2, panel B. The relatively high fragmentation of oyster genomic scaffolds unfortunately prevents more detailed studies on the C1q gene clusters organization on a larger, chromosomic scale.

3.3. Structure of C1qDC proteins

We classified oyster C1qDC genes based on the predicted subcellular localization and on the domain organization of the predicted encoded proteins, according to the classification scheme proposed by Carland and Gerwick (2010). Namely, secreted C1qDCs are indicated by the prefix "*s*", those lacking a signal peptide and thus predicted to have an intracellular function by the prefix "*c*" and those which are predicted to be membrane-bound with the prefix "*tm*". Carland and Gerwick further classified C1qDC proteins into two main categories, globular head C1q (ghC1q, i.e. those containing only a C1q domain) and C1q-like (those containing a collagen region N-terminal to the C1q domain). However, we introduced additional C1qDC subclasses, based on the peculiar features of bivalve C1qDC proteins. Table 1 summarizes the number of C1qDC genes classified in each class.

In the Mediterranean mussel, the presence of a signal peptide addresses almost all C1qDCs to the secretory pathway (Gerdol et al., 2011). Our re-annotation of oyster C1q genes is consistent with this view, as only 13 predicted proteins (about 4% out of the total) did not display a signal peptide and, according to Phobius, most of them instead showed a N-terminal transmembrane region. A total of 110 oyster proteins display a simple domain organization, with a

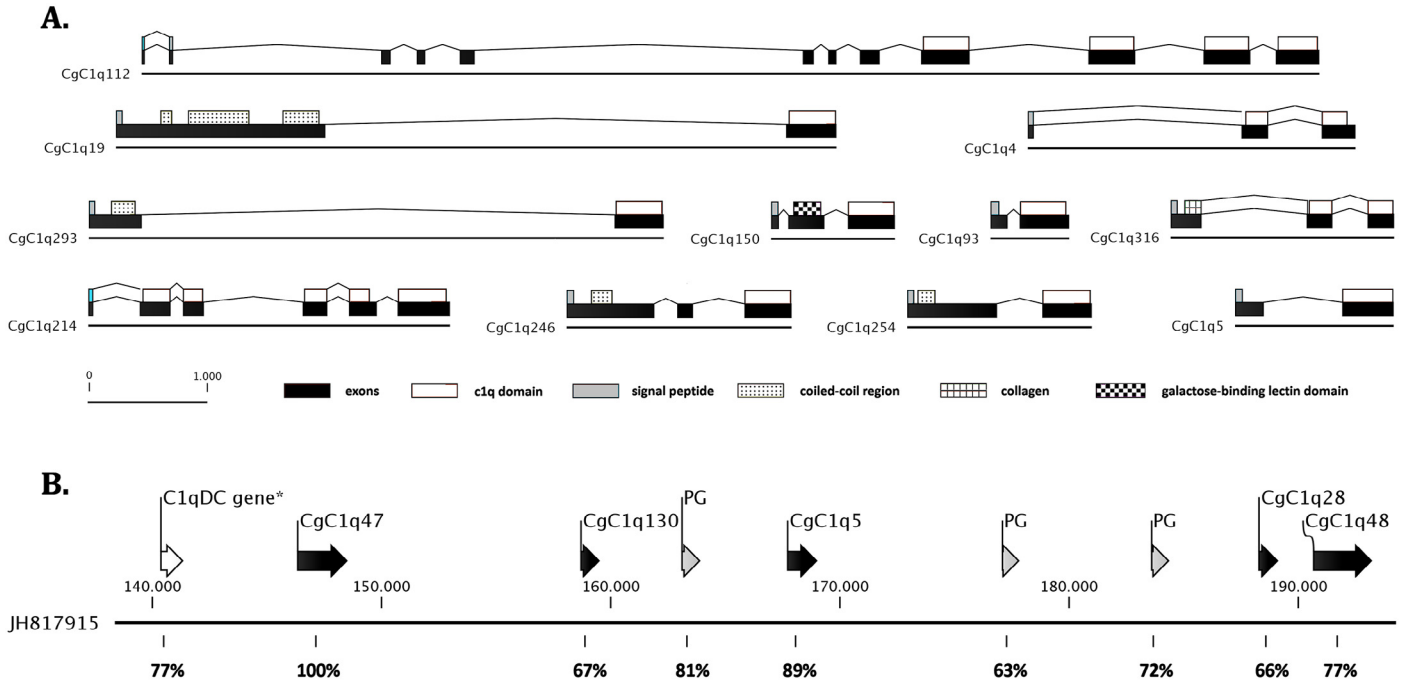

Fig. 2. Panel A: intron/exon organization of some oyster C1qDC genes. Black blocks represent exons, which are connected by lines. Blocks represent structural features and domains. Panel B: the C1qDC gene cluster contained in the genomic scaffold JH817915 (region between the nucleotides 138,000 and 194,000). A black arrow indicates complete C1qDC genes; a grey arrow indicates putative pseudogenes; an additional C1qDC gene, likely complete but whose coding sequence was interrupted by a N-stretch, is marked by an asterisk. Sequence identity percentages relative to CgC1q47 are also shown in the bottom part of the figure.

signal peptide immediately followed by a C1q domain (*s*ghC1q subclass).

A remarkably high fraction of the predicted oyster C1qDC proteins show coiled-coil regions N-terminal to the C1q domain. In detail, 210 oyster C1qDCs (62% of the total) share this feature. For comparison, we previously reported significantly lower amounts in *M. galloprovincialis* (32%), where *s*ghC1q is the largest subclass (Gerdol et al., 2011). Conversely, the abundance of coiled-coil associated leucine-zipper motifs in oyster is comparable to that of mussel (33 proteins, corresponding to 9% of the total). Collagen-like regions (characterizing the C1q-like subclass) are very common in vertebrate C1qDC proteins but they are almost completely absent in the mussel. Likewise, a single C1q-like protein was identified in oyster (CgC1q316). Since coiled-coil helices and leucine-zipper domains are well known to act as multimerization domains (Kammerer, 1997; Lupas et al., 1991; Tadokoro et al., 1999) also in vertebrate emilin/multimerin-like C1qDC proteins (Doliana et al., 1999; Hayward et al., 1995), the contemporary absence of collagen-like domains in bivalves suggests that these organisms may mainly rely on coiled-coils to organize and stabilize multimeric C1qDC complexes. For this reason, we classified all the C1qDC sequences bearing a coiled-coil domain within the novel C1q-like type 2 subfamily.

The classification of coiled-coil regions with LOGICOIL (Vincent et al., 2013) categorized most C1q-like type 2 proteins as trimers (44%), followed by parallel dimers (29%) and tetramers (23%). Antiparallel dimers were only predicted in 4% of the cases. Nevertheless, most of the times, the raw LOGICOIL scores were very similar for the three possible configurations. Therefore, in the absence of further experimental evidence, it is not currently possible to determine how C1qDC proteins interact with each other, and whether they cluster in homo- or hetero-oligomers.

With no exception, the globular C1q domain was located at the C-terminal end of the protein. Our re-annotation of C1q genes also revealed that the association between C1q and other InterPro domains in the same protein is extremely rare. In fact, only three proteins bearing a second functional domain, located N-terminal

Table 1
Classification of oyster C1qDC proteins, based on predicted subcellular localization and domain organization.

C1q subclass	Predicted subcellular localization	Domain organization	Number of genes identified
*c*C1q-like type 2	Cytoplasmic	CC + C1q	1
*s*multiC1q (2 domains)	Extracellular	SP + C1q + C1q	1
*s*C1q-like	Extracellular	SP + collagen + C1q	1
*c*ghC1q	Cytoplasmic	C1q	2
*s*multiC1q (4 domains)	Extracellular	SP + C1q + C1q + C1q + C1q	2
*s*SUEL/C1q	Extracellular	SP + SUEL + C1q	3
*tm*ghC1q	Transmembrane	TM + C1q	4
*tm*C1q-like type 2	Transmembrane	TM + CC + C1q	6
*s*multiC1q (3 domains)	Extracellular	SP + C1q + C1q + C1q	6
*s*ghC1q	Extracellular	SP + C1q	110
*s*C1q-like type 2	Extracellular	SP + CC + C1q	201

CC, coiled-coil; SP, signal peptide; TM, transmembrane domain; SUEL, D-galactoside/L-rhamnose binding SUEL lectin domain.

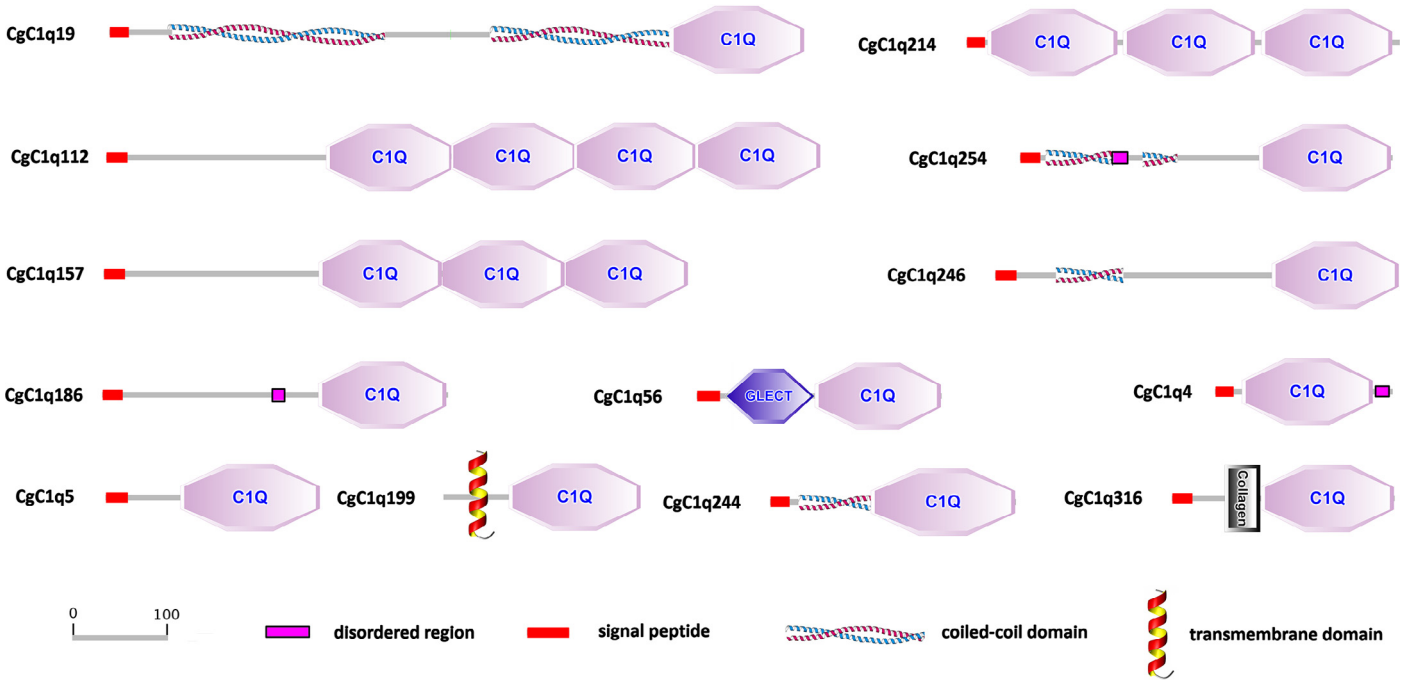

Fig. 3. Structural organization of virtually translated oyster C1qDC proteins (selected cases), namely CgC1q4, CgC1q5, CgC1q19, CgC1q56, CgC1q112, CgC1q157, CgC1q186, CgC1q199, CgC1q214, CgC1q244, CgC1q246, CgC1q254 and CgC1q316. The scale bar refers to 100 amino acid residues.

to C1q, were identified. Namely, these three oyster C1qDCs (C1qDC91, CgC1q150 and CgC1q194) are structurally similar and characterized by the presence of an immunity-related D-galactoside/L-rhamnose binding SUEL lectin domain (IPR000922, see Fig. 3). This domain is able to bind both D-galactoside and L-rhamnose and is active as a disulfide-linked homodimer (Ozeki et al., 1991), suggesting once again that bivalve C1qDCs are also organized in oligomeric complexes.

Likewise in mussel, multiC1q (proteins with multiple C1q domains) are present in oyster. More in detail, one multiC1q protein with two C1q domains, six with three C1q domains and two with four C1q domains were identified. These multiC1q proteins are expected to be secreted and only in two cases (CgC1q136 and CgC1q184) are they associated with a coiled-coil region.

3.4. *Tissue specificity of C1qDC genes*

A heat map summarizing the expression pattern of C1qDC genes in many developmental stages and adult tissues is shown in Fig. 4. Most oyster C1qDC genes display, to some extent, tissue-specific expression. According to our classification criteria (section 2.4) we identified 156 digestive gland-specific, 73 gills-specific, 5 hemocyte-specific and 5 mantle-specific C1qDCs. The expression profile of the remaining 91 genes did not evidence any preferential tissue of expression. Details on C1qDC genes found abundant in these four tissues are shown in Supplementary Appendix S4.

The comparative meta-analysis clearly indicated that a large variety of C1qDC transcripts (about 70% of the total) is predominantly expressed in the oyster digestive gland and gills, tissues where they could play multiple roles. The digestive gland is the tissue displaying both the largest number of expressed C1qDCs and the highest overall expression levels. On the contrary, just a very few C1qDC transcripts were categorized as mantle- or hemocyte-specific. The abundant expression of some C1qDC genes in the mantle of other bivalve species has been related to shell biomineralization (Hattan et al., 2001; Liu et al., 2007; Yin et al., 2005). Since a limited number of C1qDC transcripts are also highly expressed in oyster mantle, they may be involved in a similar function.

Despite the low number of hemocyte-specific C1qDC genes (see above), we could notice that CgC1q99, CgC1q163 and CgC1q315 accounted for over 80% of the total C1qDC expression in oyster circulating cells. In several cases, bivalve C1qDCs have been reported as mainly expressed in hemocytes, often inducible at very high levels in response to bacterial challenges (Gerdol et al., 2011; Gestal et al., 2010; Liu et al., 2014a; Oliveri et al., 2014). At the present time, only a single RNA-seq dataset is available from the hemocytes of unchallenged oysters and this could possibly explain the limited detection of oyster hemocyte-specific C1qDC transcripts.

According to our meta-transcriptome analysis, most C1qDC proteins are constitutively expressed in the digestive gland and in gills. The wide range of C1qDC transcripts found expressed in these tissues suggests that oysters constitutively produce an arsenal of C1qDc proteins in locations that are not classically considered primarily linked to immune functions, even if they are constantly at risk of pathogen invasion. We suppose that C1qDC proteins, constitutively expressed at significant levels, compose together with other lectins and humoral components the first line of defense in the local response to invading pathogens.

The high number of RNA-seq experiments included in our analysis permitted the identification of oyster C1qDC sequences whose expression levels are highly correlated and could be possibly regarded as clusters of co-regulated genes. Overall, we identified 23 independent clusters (see Supplementary Appendix S4). In detail, 13 consisted of digestive gland specific genes, 4 of gills-specific genes and one each of hemocytes- and mantle-specific genes. The remaining four clusters comprised genes which did not display any particular tissue preference. Each cluster comprised genes encoding widely different C1qDC proteins present on different genomic scaffolds, thus hinting that common regulatory elements might be present in neighboring genomic regions to assure an effective and coordinated expression.

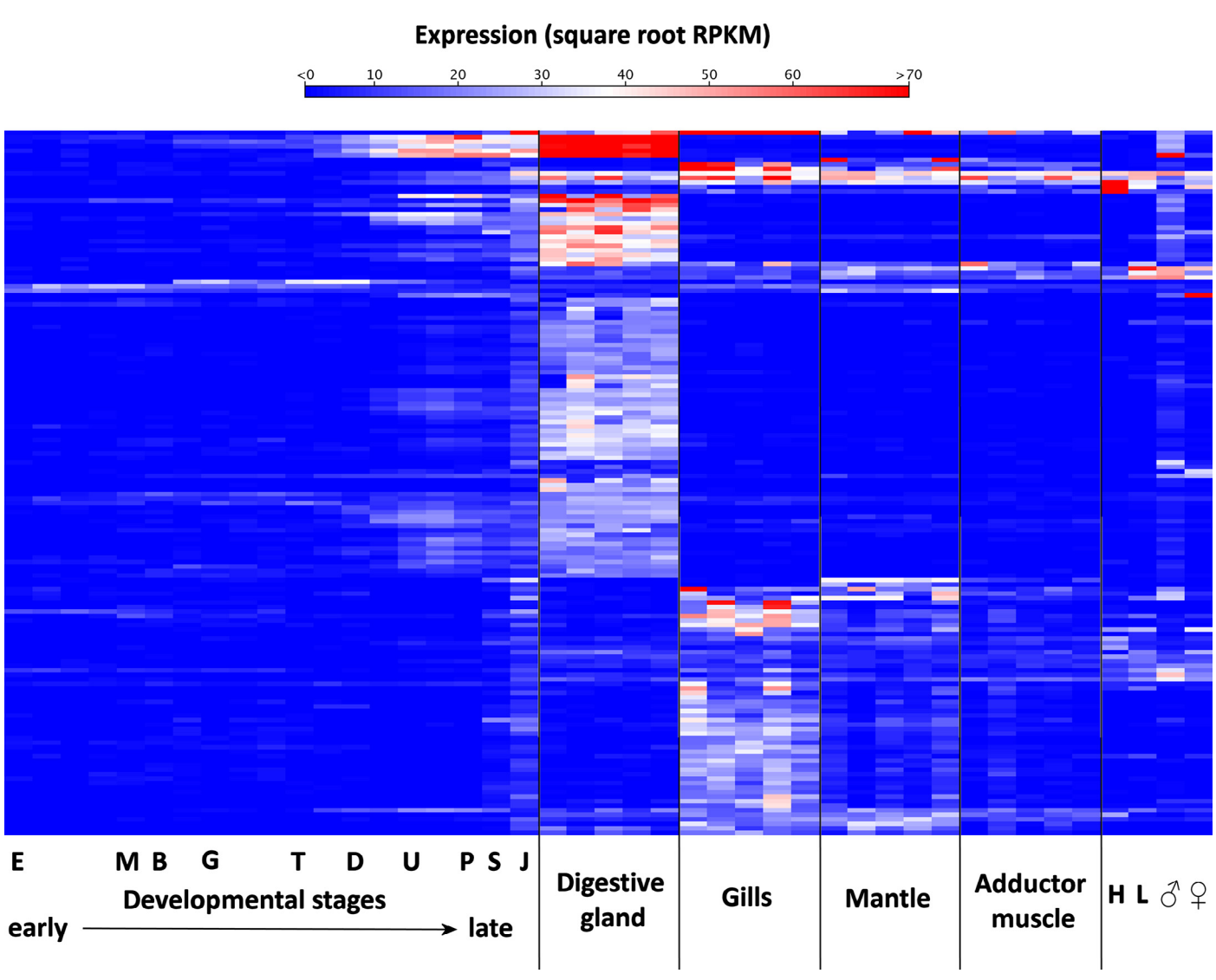

Fig. 4. Heat map summarizing the expression of individual C1qDC genes during development and in different adult tissues. E: egg; M: morula; B: blastula; G: gastrula; T: trochophore; D: D-shaped larva; U: umbo larva; P: pediveliger; S: spat; J: juvenile; H: hemocytes; L: labial palp; ♂: male gonad; ♀: female gonad. Only five representative samples are shown for adductor muscle, digestive gland, gills and mantle for the sake of simplicity. Complete gene expression data are available in Supplementary Appendix S1, Table S5.

3.5. C1qDC expression is strictly regulated during larval development

The expression of C1qDC genes is tightly regulated during early oyster development (Fig. 5). For the most part, C1qDC genes are not expressed at all until the larva settling. During the planktonic/benthic life transition, the cumulative expression of C1qDC genes gradually increases in the frame of a radical re-organization of body parts to suit a sedentary existence. At this stage, the main body tissues are rudiments of the adult shape, and the massive increase of C1qDC expression from spat to juvenile stages is consistent with the development of mature organs. More in detail, C1qDC expression is negligible up to the D-shaped larva stage, when it starts to progressively increase and reaches a peak at the umbo larva stage; after a slight decrease, it increases again, reaching a maximum at the juvenile stage, when a large variety of C1qDC transcripts, characterizing the main adult tissues, is clearly detectable (Figs. 4 and 5).

The constitutive expression of MgC1q at low levels in *M. galloprovincialis* was described throughout the first three months of larval development (Gestal et al., 2010). Oyster RNA-seq data are consistent with this information and further point out that most C1q genes show a very poor expression in larval stages compared to adult individuals.

In our analysis, only the expression of a single gene (CgC1q39, one of the few *tm*C1q-like type 2 proteins we have identified) is typical of developmental stages, as it was expressed at very high levels between the gastrula and umbo larva stages, and at much lower levels in adult tissues. Other genes important in adult oysters appear to be developmentally regulated (e.g. CgC1q60 and CgC1q82 in Fig. 5), even though their involvement in larval development remains to be understood.

3.6. Modulated expression of oyster C1qDC genes

As detailed in the materials and methods section, we used over 100 RNA-seq experiments to analyze the expression pattern of C1qDC genes in different tissues, developmental stages and in response to different stresses. Although experimental details of such samples were not always available, in particular when related to unpub-

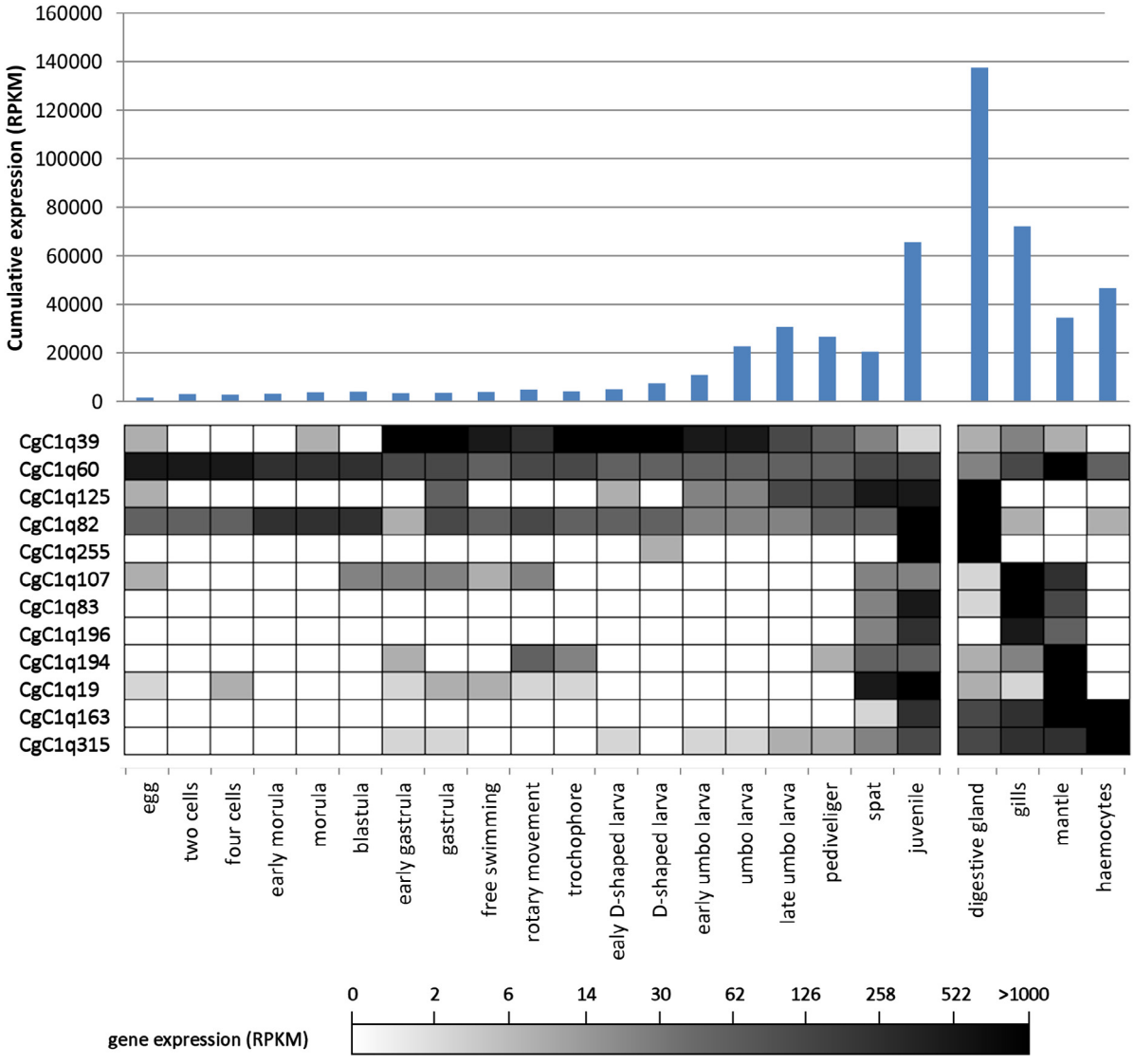

Fig. 5. Upper panel: Cumulative expression of C1qDC genes in different larval stages, compared to the cumulative expression in four adult tissues (digestive gland, gills, mantle and hemocytes). Bars represent the sum of the expression values of all C1qDC genes (measured as RPKMs). Bottom panel: expression levels of some C1qDC genes during development.

lished manuscripts, it was possible to assess the expression trends of the C1qDC family, both in terms of total cumulative expression and number of expressed genes. Charts summarizing the overall expression of the C1qDC gene repertoire in digestive gland and gills are shown in Supplementary Appendix S4. A detailed overview of the expression of oyster C1qDC genes across all the SRA samples is provided in Supplementary Appendix S1, Table S5.

Surprisingly, none of these challenges produced a significant up-regulation or down-regulation of C1qDC genes. Overall, the number of expressed genes in the digestive gland challenged with heavy metals remained very stable and the number of expressed C1qDC genes in the gills remained stable regardless of the type and intensity of challenge.

On the other hand, the relatively large differences observed among unchallenged samples (in both gills and digestive gland) suggest that a certain degree of variability exist in the basal expression of C1qDC genes even in unchallenged oysters. These differences may be attributable to factors which are often neglected or not reported at all in RNA-seq experiments, including sex, gonad development stage, age, size, season and geographic location of sampling.

However, RNA-seq data provided very useful information about the overall pattern of expression of C1qDC genes. Basically, this gene family, as a whole, is not definable as "stress-related", since no stressor was actually able to significantly affect the global expression patterns, at least not in the digestive gland or in gills where C1qDC transcripts were comparatively most abundant (Supplementary Appendix S4).

The absence of regulation following bacterial challenges (both Gram-positive and Gram-negative) in gills was also surprising, since many reports have linked an overexpression of bivalve C1qDC genes to the innate immune response. Increased expression of C1qDC transcripts has been reported in mussel, scallop and clam hemocytes upon stimulation with Gram-positive, Gram-negative bacteria and cell wall components (LPS, glucan and peptidoglycan) (Gerdol et al., 2011; Gestal et al., 2010; He et al., 2011; Li et al., 2011; Liu et al., 2014a; Wang et al., 2012; Yang et al., 2012; Zhang et al., 2008). The up-regulation of C1qDC genes in oyster hemocytes has been implicated in the response to Rickettsia-like organisms (*C. ariakensis*) (Xu et al., 2012) and to the protist *Bonamia ostreae* (*Ostrea edulis*) (Morga et al., 2012), but data concerning the response to bacteria

are still completely missing. RNA-seq data clearly point out that the expression of C1qDC genes is not inducible by bacterial challenges in gills, even though a consistent arsenal of C1qDC proteins is constitutively expressed in this tissue. Since numerous studies have previously evidenced a massive up-regulation of C1qDCs in the hemocytes of challenged bivalves, we hypothesize that a similar mechanism may occur also in oyster. Unfortunately, the absence of challenged hemocytes SRA samples prevented us to further investigate this aspect.

3.7. High variability of the C1q domain sequence

The C1q C-terminal domains of oyster C1DC proteins show a high degree of sequence variability (see Fig. 6). Following virtual translation and alignment of all oyster C1qDC CDS, only about 1/4th out of the 120–130 amino acids of this region resulted to be conserved in at least 50% of the sequences. Based on the tridimensional structure of the human complement C1q protein chain A (Gaboriaud et al., 2003), such residues appear to be mainly located in the beta strands regions. On the other hand, the regions showing the highest sequence variability are in particular the two large loops comprised between the first and the second, and between the 7th and the 8th beta strands. The alignment of oyster C1qDC proteins also evidenced regions where insertions and deletions of amino acids are permitted: in detail, alignment gaps were observed in the two large loops mentioned above and in the entire region encompassing the 6th strand and the following loop. Despite these differences, the modeled tridimensional structures revealed a remarkable structural conservation of the compact jelly-roll beta-sandwich fold typical of the C1q domain, such that all structures could be modeled based on a human template with a 100% Phyre2 confidence. The most visible structural variations can be observed in larger loop regions (Fig. 6).

Conserved residues may be important not only for the preservation of the tridimensional structure of the C1q domain, but also for guaranteeing an efficient interaction with other C1q domains in higher-order complexes. Indeed, the assembly of complement C1q heterotrimer in humans does not only rely on the interaction between the collagen-like tails, but also on hydrophobic and polar interactions between the interfaces of C1q domains themselves (Gaboriaud et al., 2003).

Taking into account the variability of C1qDCs, it is not surprising that all previous attempts at building phylogenetic trees using bivalve C1qDCs produced results with non-significant bootstrap values or inconsistencies in the species phylogeny (Gestal et al., 2010; Liu et al., 2014a; Xu et al., 2012; Yang et al., 2012). Indeed, we show that it is not possible to fully resolve bivalve and vertebrate C1qDC in two distinct branches of a phylogenetic tree (Supplementary Appendix S4). On the contrary human sequences are scattered across the tree in five different small groups. The molecular diversity between oyster and mussel C1qDC proteins is remarkably high, with the average pairwise identity between oyster and mussel putative orthologous sequences being just 46%.

3.8. Evolution and expansion of the C1qDC gene repertoire in bivalves

We previously reported that the large majority of protostomes do not show more than a dozen C1qDC genes in their genomes (Gerdol et al., 2011). The abundance and sequence diversity of C1qDC genes in bivalves represent a notable exception, even when compared to other molluscan classes.

As predicted from the *de novo* assembled transcriptomes, we estimated the number of C1qDC proteins and calculated their relative abundance in the virtual proteomes of each analyzed species. The accuracy of these calculations is likely influenced by many

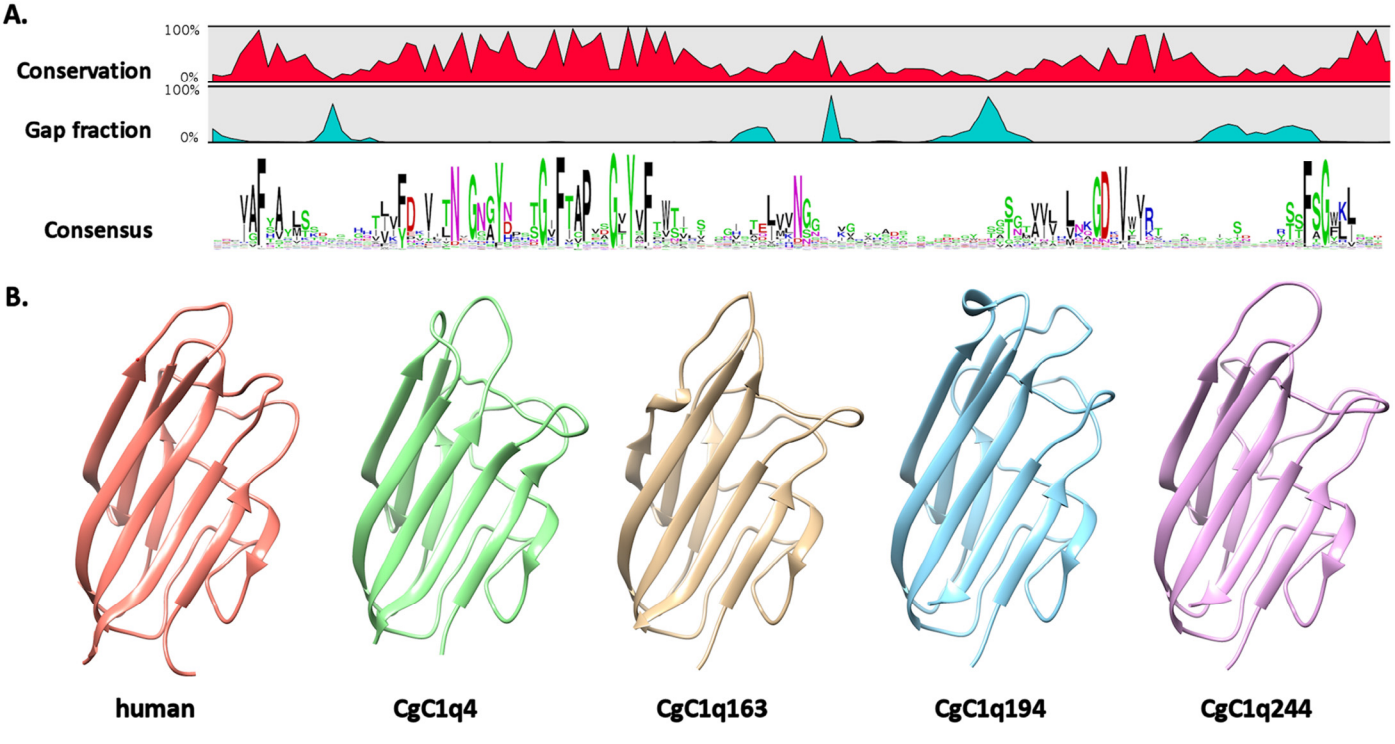

Fig. 6. Panel A: variability of the C1q domain in oyster C1qDC proteins. From top to bottom: (i) conservation plot based on the multiple alignment of the C1q domain only; (ii) gap fraction graph; (iii) consensus logo. Panel B: 3D structure of the globular head of human complement C1q (1PK6_A) and predicted 3D structures of oyster CgC1q4, CgC1q163, CgC1q194 and CgC1q244.

factors, such as the tissues used for RNA extraction, developmental stage, sex, environmental factors including the presence of stressors, etc. In addition, given the high number of C1qDC paralogous genes, similar transcripts are likely to be collapsed in a single contig during the *de novo* assembly procedure, thus leading to an under-estimation of the real abundance of C1qDC genes from transcriptomes. Hence, our calculations have to be considered as rough estimates which can nevertheless provide useful indications about the evolution of C1qDC genes and proteins in this large class of marine invertebrates.

The taxonomic and phylogenic classification of bivalves is a long debated issue, and multiple classification systems based on morphological, molecular data and a combination of both have been developed without reaching a widely accepted consensus within the scientific community (Bieler et al., 2014; Giribet, 2008). Taking this into account, we will refer to the updated classification by Bouchet and colleagues in the present paper (Bouchet et al., 2010).

The results of the comparative transcriptomic analysis are shown in Fig. 7 and detailed in Supplementary Appendix S1, Table S4. Despite

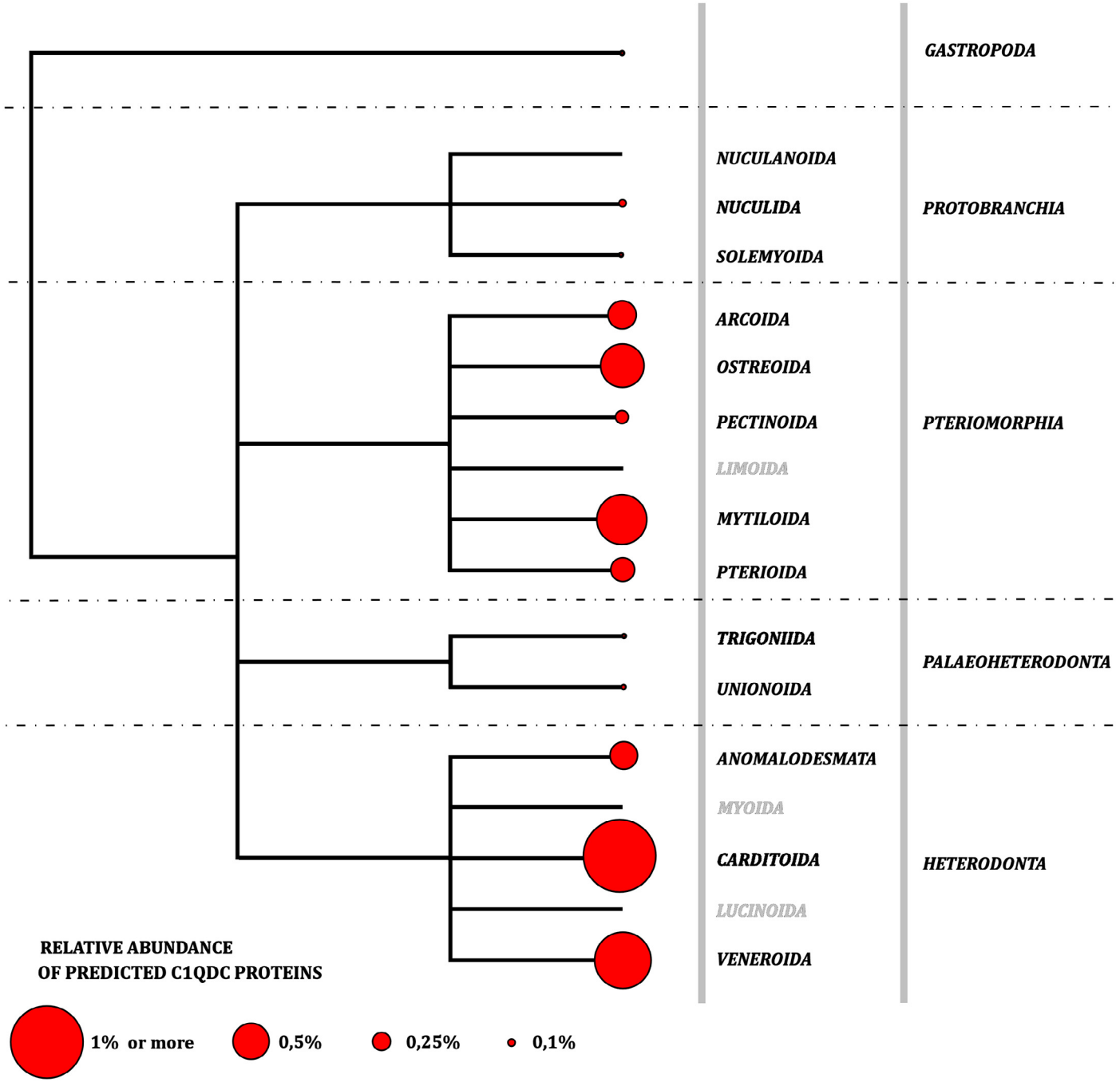

Fig. 7. Expansion of the C1qDC gene family across different taxonomic orders of bivalves, according to the classification of Bouchet et al. (2010). Relative abundances are calculated as the percentage of predicted C1qDC proteins according to *de novo* assembled bivalve transcriptomes. Circle size is directly proportional to the C1qDC protein abundance in each order, calculated as the average value among species when more than one species was available. Order names are shown in the left column, subclass names in the left one. Orders with no data available for the estimate are shown in light gray.

differences linked to the tissue of origin, the abundance of C1qDC transcripts appears clearly dependent on the bivalve taxonomic classification. In particular, the massive expansion of the C1qDC genes seems to have occurred only in two out of the four bivalve subclasses, Pteriomorphia and Heterodonta. C1qDC transcripts are indeed scarcely represented in Protobranchia (orders Nuculanoida, Nuculida and Solemyoida) and Palaeoheterodonta (orders Unionoida and Trigoniida), since their abundance is always lower than 0.1%. Based on transcriptomic data the total number of C1q genes in these organisms (1–30) seems comparable to other invertebrates and, in particular, to gastropod mollusks, with only six genes present in the genome of *L. gigantea* (Gerdol et al., 2011) and eight in the genome of *A. californica*. Despite major uncertainties concerning bivalve classification, today's protobranchs are regarded as organisms closely resembling the most primitive bivalve species (Cope, 2000). Therefore, a limited set of C1qDc proteins seems to be the plesiomorphic condition of bivalves which, besides modern protobranchs, is still observable also in palaeoheterodonts, including the freshwater mussels of the order Unionoida and the saltwater clams of the order Trigoniida.

On the contrary, C1qDC transcripts are very common in both Pteriomorphia and Heterodonta. Within the former subclass, Ostreoida and Mytiloida display a very similar abundance, in the range of 0.5 to 1% (see Fig. 7). Pterioida and Arcoida show a slightly lower relative abundance while the amount of C1qDC transcripts in Pectinoida is apparently smaller.

Concerning Heterodonta, which represent by far the largest and most diverse major group of bivalves (Taylor et al., 2007), the order Veneroida displays an average number of C1qDC proteins similar to that of Mytiloida and Ostreoida, or even higher in some species. Anomalodesmata, the most basal clade of the infraclass Euheterodonta (Bieler et al., 2014), also displays a minor expansion of the C1qDC gene family. The only species available for the infraclass Archiheterodonta (*Astarte sulcata*, order Carditoida) presents the highest relative abundance of C1qDC transcripts among all bivalves (1.93%).

Altogether, these computational data permit to estimate that the presumptive number of C1qDC genes in most pteriomorph and heterodont bivalves is well above one hundred. In the oyster genome, we confirmed 337 complete C1qDC genes, a number not including C1qDC genes which could not be fully confirmed due to insufficient RNA-seq data and pseudogenes (with the latter being quite relevant). A HMMER scan of oyster genomic scaffolds revealed 609 total putative C1qDC loci. The still incomplete status of the oyster genome is evidenced by the relatively high number of "orphan C1DC transcripts" (see section 3.1) and about 500 or slightly more functional C1qDC loci could be ascertained in the Pacific oyster *C. gigas*.

Concerning the pearl oyster *P. fucata*, 234 C1qDC genes were identified among the predicted models, but the HMMER scan evidenced as many as 720 C1qDC loci in the genomic scaffolds. The discrepancy between the numbers of annotated C1qDC genes and total presumptive loci is likely linked to the more fragmented nature of the *P. fucata* genome compared to *C. gigas* (Takeuchi et al., 2012). Assuming a total of 25,000–30,000 genes for these two species and considering the under-estimation given by paralogies, the relative abundance of C1qDC genes would be comprised between 1 and 2.5%, percentages quite in line with the values predicted from transcriptomes.

The transcriptomes of Mytiloida revealed an even higher number of non-redundant C1qDC transcripts (Supplementary Appendix S1, Table S4). Actually, we identified 1274 putative C1qDC loci in the *M. galloprovincialis* genomic scaffolds (Nguyen et al., 2014). The lack of annotation and the extremely fragmented and incomplete nature of the mussel genome do not allow an adequate analysis, though these molecular data suggest that the C1qDC gene family expansion event had a particular impact on *Mytilus* spp., leading to the diversification of at least 1500 members.

The most likely evolutionary scenery for the amazing expansion and diversification of the C1q gene family is that it might have dated back to the Cambrian/Ordovician bivalve radiation (Bieler et al., 2014; Cope, 2002) which brought to the flourishing of the over 9000 extant species today. This expansion clearly involved Pteriomorphia and Heterodonta without affecting Palaeoheterodonta. Given the uncertain phylogenetic placement of this third order with respect to the other two, it is not possible to know whether the actual abundance of C1qDC genes in Pteriomorphia and Heterodonta is the result of a single gene family expansion event in a common ancestor or of two separate events that occurred independently from each other.

3.9. Conclusion and perspectives

According to the available sequence data, we studied the C1qDC gene family in bivalves (gene organization and structure, sequence variability of the coding sequence, expression levels, regulative and evolutionary aspects) being aware that the sequences by themselves are not sufficient to solve structure/function relationships.

Massive expansions of innate immunity-related molecules, in particular PRRs, have been widely documented in invertebrates. We demonstrated that, in pteriomorph and heterodont bivalves, the number of C1qDC genes largely exceeds the hundred, possibly reaching more than a thousand members in certain mussels of the order Mytiloida. Such numbers are comparable to those of Sp185/333, scavenger receptor cysteine-rich (SRCRs) and NACHT domain and leucine-rich repeat (NRLs) proteins of sea urchin (Hibino et al., 2006) and largely exceed those of FREPs in mosquitoes (Wang et al., 2005) and lipopolysaccharide- and glucan-binding proteins (LGBPs) in *Daphnia* (Lee et al., 2000). Besides C1qDC genes, other lectin families have certainly undergone a remarkable expansion in bivalves. In particular, there is evidence of over a hundred C-type lectins and FREPs in oysters and mussels (Venier et al., 2011; Zhang et al., 2012). According to our sequence study, C1qDC transcripts seem to be the most abundant and largest PRR family in bivalves, possibly one of the most expanded and diversified family of lectin-like molecules in marine invertebrates.

Many aspects of the pathogen recognition mechanisms mediated by C1qDC proteins, however, remain unclear. First, the role of the hypervariability of this domain in the binding specificity; second, even though the presence of coiled-coil domains suggests that C1qDCs may act cooperatively in multimeric complexes, their precise mode of interaction is unknown.

On the other hand, changes in the expression of bivalve C1qDCs have also been associated to the response to heavy metals, nanoparticles and pollutants, and to shell mineralization, highlighting their involvement in non-immune functions. The high number of C1qDCs constitutively expressed in digestive gland and gills supports this view, and suggest that, following diversification of the globular C1q domain, a number of bivalve C1qDC proteins may have specialized in other, non-immune, functions likewise in human (Ghai et al., 2007). The investigation of the molecular functions of bivalve C1qDC proteins remains therefore a field to be further explored.

Overall, this study provides the first genome-scale and comparative description of C1qDC sequences from a bivalve mollusk in which repeated gene duplication events likely drove the diversification of more than 300 proteins expected to act as PRRs. Moreover, we provide to the scientific community interested in oyster immunity a manually curated genome annotation of this protein family. On the basis of the reported data, various studies could be undertaken to investigate unsolved aspects, for instance to the relationship between the structure of C1qDCs and their specific function in the interaction with PAMPs.

Acknowledgements

This work was supported by BIVALIFE (FP7-KBBE-2010-4).

References

Allam, B., Pales Espinosa, E., Tanguy, A., Jeffroy, F., Le Bris, C., Paillard, C., 2014. Transcriptional changes in Manila clam (Ruditapes philippinarum) in response to Brown Ring Disease. Fish Shellfish Immunol. 41, 2–11. doi:10.1016/j.fsi.2014.05.022; Special Issue: ISFSI 2013 review.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.

Bieler, R., Mikkelsen, P.M., Collins, T.M., Glover, E.A., González, V.L., Graf, D.L., et al., 2014. Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. Invertebr. Syst. 28, 32–115.

Bornberg-Bauer, E., Rivals, E., Vingron, M., 1998. Computational approaches to identify leucine zippers. Nucleic Acids Res. 26, 2740–2746.

Bouchet, P., Rocroi, J.-P., Bieler, R., Carter, J.G., Coan, E.V., 2010. Nomenclator of bivalve families with a classification of bivalve families. Malacologia 52, 1–184. doi:10.4002/040.052.0201.

Carland, T.M., Gerwick, L., 2010. The C1q domain containing proteins: where do they come from and what do they do? Dev. Comp. Immunol. 34, 785–790. doi:10.1016/j.dci.2010.02.014.

Cope, J., 2002. Diversification and biogeography of bivalves during the Ordovician Period. In: Crame, J.A., Owen, A.W. (Eds.), Palaeobiogeography and Biodiversity Change: The Ordovician and Mesozoic-Cenozoic Radiations. London, pp. 25–52.

Cope, J.C.W., 2000. A new look at early bivalve phylogeny. Geol. Soc. Lond. Spec. Publ. 177, 81–95. doi:10.1144/GSL.SP.2000.177.01.05.

Doliana, R., Mongiat, M., Bucciotti, F., Giacomello, E., Deutzmann, R., Volpin, D., et al., 1999. EMILIN, a component of the elastic fiber and a new member of the C1q/tumor necrosis factor superfamily of proteins. J. Biol. Chem. 274, 16773–16781.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37. doi:10.1093/nar/gkr367.

Gaboriaud, C., Juanhuix, J., Gruez, A., Lacroix, M., Darnault, C., Pignol, D., et al., 2003. The crystal structure of the globular head of complement protein C1q provides a basis for its versatile recognition properties. J. Biol. Chem. 278, 46974–46982. doi:10.1074/jbc.M307764200.

Gerdol, M., Manfrin, C., De Moro, G., Figueras, A., Novoa, B., Venier, P., et al., 2011. The C1q domain containing proteins of the Mediterranean mussel Mytilus galloprovincialis: a widespread and diverse family of immune-related molecules. Dev. Comp. Immunol. 35, 635–643. doi:10.1016/j.dci.2011.01.018.

Gerdol, M., De Moro, G., Manfrin, C., Milandri, A., Riccardi, E., Beran, A., et al., , 2014. RNA sequencing and de novo assembly of the digestive gland transcriptome in Mytilus galloprovincialis fed with toxinogenic and non-toxic strains of Alexandrium minutum. BMC Res. Notes 7, 722.

Gerlach, D., Schlott, B., Schmidt, K.-H., 2004. Cloning and expression of a sialic acid-binding lectin from the snail Cepaea hortensis. FEMS Immunol. Med. Microbiol. 40, 215–221. doi:10.1016/S0928-8244(03)00367-5.

Gestal, C., Pallavicini, A., Venier, P., Novoa, B., Figueras, A., 2010. MgC1q, a novel C1q-domain-containing protein involved in the immune response of Mytilus galloprovincialis. Dev. Comp. Immunol. 34, 926–934. doi:10.1016/j.dci.2010.02.012.

Ghai, R., Waters, P., Roumenina, L.T., Gadjeva, M., Kojouharova, M.S., Reid, K.B.M., et al., 2007. C1q and its growing family. Immunobiology 212, 253–266. doi:10.1016/j.imbio.2006.11.001.

Ghebrehiwet, B., Hosszu, K., Valentino, A., Peerschke, E.I.B., 2012. The C1q family of proteins: insights into the emerging non-traditional functions. Front. Immunol. 3, 52. doi:10.3389/fimmu.2012.00052.

Giribet, G., 2008. Bivalvia. In: Ponder, W., Lindberg, D.R. (Eds.), Phylogeny and Evolution of the Mollusca. Oakland, CA, pp. 105–141.

Gomes, T., Pereira, C.G., Cardoso, C., Bebianno, M.J., 2013. Differential protein expression in mussels Mytilus galloprovincialis exposed to nano and ionic Ag. Aquat. Toxicol. 136–137, 79–90. doi:10.1016/j.aquatox.2013.03.021.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. doi:10.1038/nbt.1883.

Hattan, S.J., Laue, T.M., Chasteen, N.D., 2001. Purification and characterization of a novel calcium-binding protein from the extrapallial fluid of the mollusc, Mytilus edulis. J. Biol. Chem. 276, 4461–4468. doi:10.1074/jbc.M006803200.

Hayward, C.P., Hassell, J.A., Denomme, G.A., Rachubinski, R.A., Brown, C., Kelton, J.G., 1995. The cDNA sequence of human endothelial cell multimerin. A unique protein with RGDS, coiled-coil, and epidermal growth factor-like domains and a carboxyl terminus similar to the globular domain of complement C1q and collagens type VIII and X. J. Biol. Chem. 270, 18246–18251.

He, X., Zhang, Y., Yu, F., Yu, Z., 2011. A novel sialic acid binding lectin with anti-bacterial activity from the Hong Kong oyster (Crassostrea hongkongensis). Fish Shellfish Immunol. 31, 1247–1250. doi:10.1016/j.fsi.2011.08.021.

Hibino, T., Loza-Coll, M., Messier, C., Majeske, A.J., Cohen, A.H., Terwilliger, D.P., et al., 2006. The immune gene repertoire encoded in the purple sea urchin genome. Dev. Biol. 300, 349–365. doi:10.1016/j.ydbio.2006.08.065.

Kammerer, R.A., 1997. Alpha-helical coiled-coil oligomerization domains in extracellular proteins. Matrix Biol. 15, 555–565, discussion 567–568.

Käll, L., Krogh, A., Sonnhammer, E.L.L., 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338, 1027–1036. doi:10.1016/j.jmb.2004.03.016.

Kelley, L.A., Sternberg, M.J., 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. 4, 363–371. doi:10.1038/nprot.2009.2.

Kishore, U., Reid, K.B., 1999. Modular organization of proteins containing C1q-like globular domain. Immunopharmacology 42, 15–21.

Kishore, U., Gaboriaud, C., Waters, P., Shrive, A.K., Greenhough, T.J., Reid, K.B.M., et al., 2004. C1q and tumor necrosis factor superfamily: modularity and versatility. Trends Immunol. 25, 551–561. doi:10.1016/j.it.2004.08.006.

Kong, P., Zhang, H., Wang, L., Zhou, Z., Yang, J., Zhang, Y., et al., 2010. AiC1qDC-1, a novel gC1q-domain-containing protein from bay scallop Argopecten irradians with fungi agglutinating activity. Dev. Comp. Immunol. 34, 837–846. doi:10.1016/j.dci.2010.03.006.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580. doi:10.1006/jmbi.2000.4315.

Lee, S.Y., Wang, R., Söderhäll, K., 2000. A lipopolysaccharide- and beta-1,3-glucan-binding protein from hemocytes of the freshwater crayfish Pacifastacus leniusculus. Purification, characterization, and cDNA cloning. J. Biol. Chem. 275, 1337–1343.

Li, C., Yu, S., Zhao, J., Su, X., Li, T., 2011. Cloning and characterization of a sialic acid binding lectins (SABL) from Manila clam Venerupis philippinarum. Fish Shellfish Immunol. 30, 1202–1206. doi:10.1016/j.fsi.2011.02.022.

Liu, H.-H., Xiang, L.-X., Shao, J.-Z., 2014a. A novel C1q-domain-containing (C1qDC) protein from Mytilus coruscus with the transcriptional analysis against marine pathogens and heavy metals. Dev. Comp. Immunol. 44, 70–75. doi:10.1016/j.dci.2013.11.009.

Liu, H.-L., Liu, S.-F., Ge, Y.-J., Liu, J., Wang, X.-Y., Xie, L.-P., et al., 2007. Identification and characterization of a biomineralization related gene PFMG1 highly expressed in the mantle of Pinctada fucata. Biochemistry (Mosc) 46, 844–851. doi:10.1021/bi061881a.

Liu, N., Pan, L., Gong, X., Tao, Y., Hu, Y., Miao, J., 2014b. Effects of benzo(a)pyrene on differentially expressed genes and haemocyte parameters of the clam Venerupis philippinarum. Ecotoxicology 23, 122–132. doi:10.1007/s10646-013-1157-7.

Lupas, A., Van Dyke, M., Stock, J., 1991. Predicting coiled coils from protein sequences. Science 252, 1162–1164. doi:10.1126/science.252.5009.1162.

Maria, V.L., Gomes, T., Barreira, L., Bebianno, M.J., 2013. Impact of benzo(a)pyrene, Cu and their mixture on the proteomic response of Mytilus galloprovincialis. Aquat. Toxicol. 144–145, 284–295. doi:10.1016/j.aquatox.2013.10.009.

McDowell, I.C., Nikapitiya, C., Aguiar, D., Lane, C.E., Istrail, S., Gomez-Chiarri, M., 2014. Transcriptome of American oysters, Crassostrea virginica, in response to bacterial challenge: insights into potential mechanisms of disease resistance. PLoS ONE 9, e105097. doi:10.1371/journal.pone.0105097.

Morga, B., Renault, T., Faury, N., Arzul, I., 2012. New insights in flat oyster Ostrea edulis resistance against the parasite Bonamia ostreae. Fish Shellfish Immunol. 32, 958–968. doi:10.1016/j.fsi.2012.01.026.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628. doi:10.1038/nmeth.1226.

Nguyen, T.T.T., Hayes, B.J., Ingram, B.A., 2014. Genetic parameters and response to selection in blue mussel (Mytilus galloprovincialis) using a SNP-based pedigree. Aquaculture 420–421, 295–301. doi:10.1016/j.aquaculture.2013.11.021.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6.

Oliveri, C., Peric, L., Sforzini, S., Banni, M., Viarengo, A., Cavaletto, M., et al., 2014. Biochemical and proteomic characterisation of haemolymph serum reveals the origin of the alkali-labile phosphate (ALP) in mussel (Mytilus galloprovincialis). Comp. Biochem. Physiol. Part D Genomics Proteomics 11, 29–36. doi:10.1016/j.cbd.2014.07.003.

Ozeki, Y., Matsui, T., Suzuki, M., Titani, K., 1991. Amino acid sequence and molecular characterization of a D-galactoside-specific lectin purified from sea urchin (Anthocidaris crassispina) eggs. Biochemistry (Mosc) 30, 2391–2394. doi:10.1021/bi00223a014.

Perrigault, M., Tanguy, A., Allam, B., 2009. Identification and expression of differentially expressed genes in the hard clam, Mercenaria mercenaria, in response to quahog parasite unknown (QPX). BMC Genomics 10, 377. doi:10.1186/1471-2164-10-377.

Philipp, E.E.R., Kraemer, L., Melzner, F., Poustka, A.J., Thieme, S., Findeisen, U., et al., 2012. Massively parallel RNA sequencing identifies a complex immune gene repertoire in the lophotrochozoan Mytilus edulis. PLoS ONE 7, e33091. doi:10.1371/journal.pone.0033091.

Prado-Alvarez, M., Gestal, C., Novoa, B., Figueras, A., 2009. Differentially expressed genes of the carpet shell clam Ruditapes decussatus against Perkinsus olseni. Fish Shellfish Immunol. 26, 72–83. doi:10.1016/j.fsi.2008.03.002.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.
Shapiro, L., Scherer, P.E., 1998. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. Curr. Biol. 8, 335–338.
Suárez-Ulloa, V., Fernández-Tajes, J., Manfrin, C., Gerdol, M., Venier, P., Eirín-López, J.M., 2013. Bivalve omics: state of the art and potential applications for the biomonitoring of harmful marine compounds. Mar. Drugs 11, 4370–4389. doi:10.3390/md11114370.
Tadokoro, S., Tachibana, T., Imanaka, T., Nishida, W., Sobue, K., 1999. Involvement of unique leucine-zipper motif of PSD-Zip45 (Homer 1c/vesl-1L) in group 1 metabotropic glutamate receptor clustering. Proc. Natl. Acad. Sci. U.S.A. 96, 13801–13806. doi:10.1073/pnas.96.24.13801.
Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., et al., 2012. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes 19, 117–130. doi:10.1093/dnares/dss005.
Taris, N., Lang, R.P., Reno, P.W., Camara, M.D., 2009. Transcriptome response of the Pacific oyster (Crassostrea gigas) to infection with Vibrio tubiashii using cDNA AFLP differential display. Anim. Genet. 40, 663–677. doi:10.1111/j.1365-2052.2009.01894.x.
Taylor, J.D., Williams, S.T., Glover, E.A., Dyal, P., 2007. A molecular phylogeny of heterodont bivalves (Mollusca: Bivalvia: Heterodonta): new analyses of 18S and 28S rRNA genes. Zool. Scr. 36, 587–606. doi:10.1111/j.1463-6409.2007.00299.x.
Tom Tang, Y., Hu, T., Arterburn, M., Boyle, B., Bright, J.M., Palencia, S., et al., 2005. The complete complement of C1q-domain-containing proteins in Homo sapiens. Genomics 86, 100–111. doi:10.1016/j.ygeno.2005.03.001.
Venier, P., Varotto, L., Rosani, U., Millino, C., Celegato, B., Bernante, F., et al., 2011. Insights into the innate immunity of the Mediterranean mussel Mytilus galloprovincialis. BMC Genomics 12, 69. doi:10.1186/1471-2164-12-69.
Vincent, T.L., Green, P.J., Woolfson, D.N., 2013. LOGICOIL – multi-state prediction of coiled-coil oligomeric state. Bioinformatics 29, 69–76. doi:10.1093/bioinformatics/bts648.
Wang, L., Wang, L., Kong, P., Yang, J., Zhang, H., Wang, M., et al., 2012. A novel C1qDC protein acting as pattern recognition receptor in scallop Argopecten irradians. Fish Shellfish Immunol. 33, 427–435. doi:10.1016/j.fsi.2012.05.032.
Wang, X., Zhao, Q., Christensen, B.M., 2005. Identification and characterization of the fibrinogen-like domain of fibrinogen-related proteins in the mosquito, Anopheles gambiae, and the fruitfly, Drosophila melanogaster, genomes. BMC Genomics 6, 114. doi:10.1186/1471-2164-6-114.
Xu, T., Xie, J., Li, J., Luo, M., Ye, S., Wu, X., 2012. Identification of expressed genes in cDNA library of hemocytes from the RLO-challenged oyster, Crassostrea ariakensis Gould with special functional implication of three complement-related fragments (CaC1q1, CaC1q2 and CaC3). Fish Shellfish Immunol. 32, 1106–1116. doi:10.1016/j.fsi.2012.03.012.
Yang, J., Wei, X., Liu, X., Xu, J., Yang, D., Yang, J., et al., 2012. Cloning and transcriptional analysis of two sialic acid-binding lectins (SABLs) from razor clam Solen grandis. Fish Shellfish Immunol. 32, 578–585. doi:10.1016/j.fsi.2012.01.012.
Yin, Y., Huang, J., Paine, M.L., Reinhold, V.N., Chasteen, N.D., 2005. Structural characterization of the major extrapallial fluid protein of the mollusc Mytilus edulis: implications for function. Biochemistry (Mosc) 44, 10720–10731. doi:10.1021/bi0505565.
Zdobnov, E.M., Apweiler, R., 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinforma. Oxf. Engl. 17, 847–848.
Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al., 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49–54. doi:10.1038/nature11413.
Zhang, H., Song, L., Li, C., Zhao, J., Wang, H., Qiu, L., et al., 2008. A novel C1q-domain-containing protein from Zhikong scallop Chlamys farreri with lipopolysaccharide binding activity. Fish Shellfish Immunol. 25, 281–289. doi:10.1016/j.fsi.2008.06.003.
Zhang, L., Li, L., Zhu, Y., Zhang, G., Guo, X., 2014. Transcriptome analysis reveals a rich gene set related to innate immunity in the Eastern oyster (Crassostrea virginica). Mar. Biotechnol. 16, 17–33. doi:10.1007/s10126-013-9526-z.