Developmental and Comparative Immunology 49 (2015) 59–71

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

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

Article history:
Received 26 September 2014
Revised 6 November 2014
Accepted 6 November 2014
Available online 11 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).

http://dx.doi.org/10.1016/j.dci.2014.11.007
0145-305X/

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