MgC1q, a novel C1q-domain-containing protein involved in the immune response of *Mytilus galloprovincialis*

Gestal, C.[1], Pallavicini, A[2], Venier, P[3], Novoa, B. [1]and Figueras, A.[1*]

[1]*Instituto de Investigaciones Marinas (CSIC). Vigo, Spain.*

[2]*Department of Life Sciences, University of Trieste, Trieste, Italy*

[3] *Department of Biology, CRIBI biotechnology Center, University of Padova, Padova, Italy*

* Corresponding author
A. Figueras
Instituto de Investigaciones Marinas (CSIC)
Eduardo Cabello 6, 36208, Vigo, Spain
e-mail: antoniofigueras@iim.cisc.es

Keywords: *M. galloprovincialis*, MGC1q, gene characterization, immune gene, bacterial infection, expression level.

Abbreviations: SSH, suppression subtractive hybridization; C1q-DC, C1q-domain-containing

Abstract

In this study, we present the characterization of a newly identified gene, MgC1q, revealed in suppression subtractive hybridization and cDNA libraries from immunostimulated mussels. Based on sequence homology, molecular architecture and domain similarity, this new C1q-domain-containing gene may be classified as a member of the C1q family and, therefore, part of the C1q-TNF superfamily. The expression of MgC1q was detected all along the mussel ontogeny, being detectable within 2 h post-fertilization, with a notable increase after 1 month and continuing to increase until 3 months. Measurable transcript levels were also evident in all analyzed tissues of naïve adult mussels, and the hemocytes showed the highest expression levels. Experimental infection of adult mussels with Gram positive or Gram negative bacteria significantly modulated the MgC1q expression, and confirmed it as an immune-related gene. Intra- and inter-individual sequence analyses revealed extraordinary diversity of MgC1q at both the DNA and cDNA levels. While further research is needed to define its function, our data indicate that MgC1q is a pattern recognition molecule able to recognize pathogens during innate immune responses in *Myitilus galloprovincialis*. The high sequence variability suggests that somatic diversification of these nonself recognition molecules could have occurred.

1. Introduction

Recognition of surface patterns that are common to large and diverse groups of potential pathogens, but are absent in the host, represents an effective strategy for immune response [1]. The complement system is a highly sophisticated and powerful defense mechanism composed of more than 30 soluble serum and cell surface proteins that play essential roles in both innate and adaptive host defense [2, 3]. Classically, the complement system can be activated through three pathways: the antibody-dependent classical pathway, the antibody-independent alternative pathway and the lectin pathway. C1q, as a subcomponent of the complement C1 complex, is the target recognition protein of the classical complement pathway, playing a crucial role in the recognition of pathogen surface structures and antibody-antigen complexes to initiate this pathway and mediate adaptive immunity [4]. In addition, C1q is considered to be a versatile pattern recognition protein (PRP), binding directly to a broad range of pathogen–associated molecular patterns (PAMPs) of bacteria, viruses, and parasites, as well as enhancing pathogen phagocytosis [5,6]. Acting as a lectin, C1q activates the complement system, as reported in lower vertebrates or jawless fish such as the lamprey [7]. For this reason, C1q is considered to be a major connecting link between classical pathway-driven innate immunity and acquired immunity [8].

The C1q-domain-containing (C1q-DC) proteins are a family of proteins characterized by a globular domain of around 140 residues in the C-terminus with eight highly conserved residues, followed in most of them by a collagen-like region and a short amino-terminal region [9]. Crystallographic and molecular architecture studies on C1qDC proteins of C1q family member revealed an asymmetrical trimer of beta sandwich subunits, each of which has a 10-strand jelly-roll folding topology similar to that observed within the conserved C-terminal TNF homology domain of the tumor necrosis factor (TNF) family proteins. In addition, C1q and TNF family proteins also have similar gene structures. Their globular domains are each encoded within one exon, whereas introns in both families are restricted to respective N-terminal collagen or stalk regions. These features suggest an evolutionary link between these two families, arising by divergence from a common precursor of the C1q/TNF superfamily [10, 11, 12, 13, 4].

C1q-DC proteins are widely studied in vertebrates, and new members of the C1q domain family have been recently described. This family includes not only the complement component C1q as an immunological mediator, but also non-complement proteins such as adiponectin, precerebellin, hibernation protein, multimerin, ovary specific protein, and others with different functional roles [8]. However, both C1q and TNF family proteins appear to play major roles in immunity as well as homeostasis. C1q/TNF superfamily members have been found to be involved in host defense, inflammation, apoptosis, autoimmunity, and cell differentiation [4]. Interestingly, C1q and TNF-α can be produced in response to infection as inducers of proinflammatory activators [14] and cell-cycle arrest [15], and they can promote cell survival through the NF-κB pathway [16, 17].

The complement system has been studied extensively in mammals, but considerably less is known in lower vertebrates and invertebrates. The central C3 component has been identified not only in deuterostomes but also in cnidarians, such as the anthozoan coral, and the protostome horseshoe crab, indicating the very ancient origin of C3 [18]. Recently, it has been also identified in the lophotrochozoan Hawaiian bobtail squid *Euprymna scolopes* [19]. However, few reports have focused on the C1q-like complement component proteins in lower vertebrates and invertebrates. The C1q-DC proteins have been described in different teleost fish, containing a collagen domain followed by a C1q domain. An N-acetylglucosamine (GlcNAc)-binding lectin, LC1q, has been identified in the lamprey as an ortholog of mammalian C1q, bearing structural similarity to mammalian C1q [7], whereas C1q-DC proteins with and without collagen domains have been recently described in mandarin fish *Siniperca chuatsi* and color crucian carp *Carassius auratus* and in the scallop *Clamys farrery,* [20, 21, 22]. Notably, a total of 52 independent sequences for C1q-DC proteins have been discovered in zebrafish *Danio rerio* [23], and a number of models encoding C1q-like proteins have been identified in the purple sea urchin genome [24]. However, based solely on primary structure and the C1q domain, it is difficult to recognize the involvement of these molecules in the classical complement pathway.

To date, there is no evidence of specific adaptive immunity in invertebrates; however, these organisms are equipped with innate immune systems consisting of circulating cells and a large variety of molecular effectors able to recognize conserved surface epitopes, known as PAMPs, exemplified by LPS, lipoteichoic acids, peptidoglycans, and β- glucans [25]. Until few years ago, innate and adaptive immunity were considered to be two independent mechanisms. At present, they appear to be more intricately linked [26, 27]. New findings such as gene rearrangement in lampreys and somatic hypermutation/alternative splicing in mollusks suggest the need for additional studies on this invertebrate group [28, 29, 30].

Humoral and cellular responses involving soluble proteins and active phagocytic hemocytes, respectively, are well documented in bivalves [31, 32, 33, 34, 25]. However, relatively little is known about the molecular mechanisms of recognition, activation, and production of effector molecules in response to pathogens [35, 36, 37, 38, 39]. The only reference related to the presence of the complement system in bivalves has been published recently [40], describing a new C3 protein, RD-C3, and a new factor B-like protein in the carpet shell clam *Ruditapes decussatus,* demonstrating the presence of complement components in bivalve mollusks. Transcripts putatively identifying the C3 complement component in the mussel *Mytilus galloprovincialis*, MGC07073 and MGC05748, have been also discovered by large-scale EST sequencing [41]. No other complement components have been identified in bivalves to date.

In this study, we provide sequence analysis of a new C1q-domain containing protein from *M. galloprovincialis,* a mollusk species of high economic importance in aquaculture, and describe the high variability found at the genomic and transcriptional levels. We also present data on the expression of this protein during the ontogeny and in different organs of adult naïve mussels as well as in response to Gram negative and Gram positive bacterial infections.

2. Material and methods

2.1 EST selection, identification and characterization of a new C1q-domain containing cluster

Suppressive subtractive hybridization (SSH) and primary cDNA libraries previously constructed in our laboratories yielded a highly expressed transcript in *M. galloprovincialis* challenged with a mixture of dead bacteria and poly I:C [38, 41]. This transcript was identified as a gene bearing a high degree of similarity to oyster mantle gene 4, containing a C1q domain. In addition, another group of 8 ESTs with high similarity to a sialic acid binding protein, containing a C1q domain and similar sequence characteristics to mantle gene 4 proteins, were identified in mussel libraries. All raw EST data were manually analyzed and clustered. Raw chromatograms were analyzed with Chromas 231 software (Technelysium). A search for similarities to known genes was performed using BLAST (http://www.ncbi.nlm.nih.gov/blast/), specifically BlastX, and the best annotated hit from the similarity search was retained. Translation and protein analysis were carried out using the ExPaSy tools (http://us.expasy.org/tools). Multiple sequence alignments were generated by Clustal W [42]. Only ESTs with a full coding sequence were selected for further analysis.

2.2 Sequence analysis of a new C1q cluster (MgC1q)

After analysis and clustering of sequences, the complete open reading frame (ORF) of the most common and representative transcript of the new C1q cluster was characterized.

The presence of a signal peptide and location of cleavage sites was evaluated with SignalIP 3.0 software at http://www.cbs.dtu.dk/services/SignalP/ [43]. Protein domains were predicted with the Simple Modular Architecture Research Tool (SMART) version 4.0 (http:// www.smart.emblheidelberg.de/). We used the SOPMA [44], PredictProtein [45], and PORTER [46] software programs to predict alpha-helix, beta-sheet, and coil motifs in the secondary protein structure.

2.3 Phylogenetic analysis

Protein sequences from complete C1q domains obtained from the GenBank database included TNF superfamily member *N. vectensis* TNF receptor binding (XP001638169), *Chlamys farreri* C1q domain-containing protein (ABS50435), Sialic acid binding lectin *Cepaea hortensis* (CAD83837), Sialic acid binding lectin *Helix pomatia* (ABF00124), C1q B chain *Homo sapiens* (NP_000482), Adiponectin precursor of *Homo sapiens* (NP_004788), Cerebelin *Homo sapiens* (AAQ89429), C1q-containing protein Gliacolin-like *H. sapiens* (NP872334), HM MYTED Ep protein precursor *Mytilus edulis* (AAQ63463), C1q-like protein *Lamprea japonica* (BAD22833), C1q-Gliacolin *C. gigas* (translated from CX069305), C1q *Argopecten irradians* (translated from CB416625), mantle gene 4 *Pinctada fucata* (AAZ76258), and C1q *Ruditapes decussatus* (translated from EY255091). These sequences were aligned, together with the sequences obtained in this study, using ClustalW [42] and built-in MEGA4 [47]. The alignment was manually edited using the seed alignment of the Pfam PF00386 C1q domain as a template [48], and a phylogenetic tree was built using the Neighbor-Joining (NJ), the Minimum evolution, and the Maximum parsimony algorithms of MEGA4 [47]. Statistical confidence on the inferred phylogenetic relationships was assessed by performing 10,000 bootstrap replicates. We show the NJ tree after a manual verification of topology consistency. Branches corresponding to partitions reproduced in less than 50% of bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (10,000 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The distances matrix was computed using the Poisson correction method. All positions containing alignment gaps and missing data were eliminated only in pair wise sequence comparisons.

2.4 Genomic analysis

To study the genomic intra- and inter-individual variability, total DNA from the muscle tissue of 7 different mussels was extracted using the Phenol-Chloroform method [49]. PCR was performed using the designed pair of conserved primers (A10C1q complete 2F: 5'-CATCACAAGGGACTTTGGTAGA-3' and A10C1q complete R: 5'GATCTGGCGAACTTTGTTCC 3') to amplify the C1q gene from genomic DNA, thus undertaking a preliminary assessment of genetic polymorphism. Following cloning of the PCR products, 3-6 clones per mussel were randomly selected and subjected to complete sequencing in both the forward and reverse directions.

To study of the intra- and inter-individual variability of the transcribed C1q sequence, RNA was extracted from 6 different mussels using Trizol (Invitrogen), converted to cDNA, and amplified using the same pair of conserved primers. Similarly, 4-6 clones per mussel from 6 different mussels were sequenced in both directions following cloning of individual PCR products.

In addition, a comparison between genomic DNA (from a piece of muscle after extraction using Phenol-Chloroform) and cDNA (from a piece of muscle after extraction of RNA using Trizol and conversion to cDNA) from one individual mussel was performed.

To identify possible introns in the DNA sequence, all genomic DNA sequences obtained were analyzed using the Wise2 software program [50]. Confirmation of the presence of one intron in the genomic DNA sequence and its variability was investigated by amplification and sequencing of a total of 3-6 clones each from 8 different mussels using a pair of specific primers: C1q mussel INT 2F: 5'-TCACTTGGACCATCTGCGTA-3' and C1q A10 complete R: 5'-GATCTGGCGAACTTTGTTCC-3'.

The PCR profile consisted of an initial denaturation of 5 min at 94ºC; 35 cycles of 1 min denaturation at 94ºC, 1 min of annealing at 60ºC, and 1 min of elongation at 72ºC; and a final extension for 10 min at 72ºC. Excess primers and nucleotides were removed by enzymatic digestion using 10 U and 1 U of Exonuclease I (ExoI) and Shrimp Alkaline Phosphatase (SAP), respectively (Amersham Biosciences), at 37ºC for 1 h, followed by inactivation of the enzymes at 80ºC for 15 min. DNA sequencing was performed using a BigDye terminator Cycle Sequencing Ready Reaction Kit and an automated DNA Sequencer (ABI 3730).

2.5 Molecular variant analysis

We mined Mytibase [41] for the cluster corresponding to the mantle gene 4 protein. We identified a cluster (MGC00284) composed of 109 ESTs. These sequences were aligned using ClustalW and edited with MEGA. The appropriate substitution model was selected by the hierarchical likelihood ratio test implemented in Modeltest [51]. Using the selected model, we reconstructed an unrooted NJ tree. The bootstrap consensus tree inferred from 5,000 replicates was taken to represent the cladistic tree of the taxa analyzed. The inferred tree was used for selection analysis.

Polymorphism data, nucleotide variability, and neutrality tests were conducted with the program DnaSP [52]. To determine whether positive selection is operating in any of the codon sites of the 109 variants, we compared data obtained from the CodeML program to data from the PAML software package [53] and SLR methods [54].

2.6 Expression analysis by Q-PCR

1. Expression profile of MGC1q in early larval developmental stages and in different tissues

Larvae were produced at the laboratory after fertilization of adult mussels. Spawning of individual mussels was induced by increasing the seawater temperature from room temperature (18ºC) to 22-23ºC and maintaining the animals out of the water for 30 min. Mussels were placed in independent tanks of 1 L capacity. The number of gametes from individual mussels was quantified in a Neubauer chamber, and a proportion of 10:1 (spermatozoid:oocyte) was used to induce fertilization. Aliquots from the same samples used in fertilization were centrifuged at 2000xg for 10 min at 4ºC. Pellets with gametes were resuspended in 1 ml Trizol and maintained at -80 ºC until RNA was isolated. Larvae were kept growing in a flow-through system of aerated sea water at 18ºC and fed daily with *Isochrysys galbana* (12 x 10^8 cells/animal), *Tetraselmis suecica* (10^7 cells/animal) and

Skeletonema costatum (3 x 10^8 cells/animal). To evaluate and quantify the relative expression of C1q during the ontogeny of mussels through the larval developmental stages, larval pools were sampled at 2, 20 hours, 3,7,9,14, 32, 38 days and 1.5, 2 and 3 months. After centrifugation at 2000xg for 10 min at 4ºC, pellets were also resuspended in Trizol, and RNA was extracted. The cDNAs obtained were used to amplify the new MgC1q gene.

In addition, several tissues of naïve adult mussels (hemocytes, gills, mantle, muscle, gonad and digestive gland) were pooled per tissue (three pools of 4 individuals each), RNA was isolated following the Trizol reagent's instructions, and used to generate cDNA. Amplification of the new MgC1q gene was carried out to investigate the possibility of constitutive gene expression.

2. Experimental infections with Gram positive and Gram negative bacteria

Adult *M. galloprovincialis* mussels were obtained from a commercial shellfish farm from the Ría de Vigo (NW of Spain). Animals were acclimated for 1 week in the laboratory in open circuit filtered seawater tanks at 15ºC with aeration prior to experiments.

Two hundred and forty mussels were notched at the shell adjacent to the posterior adductor muscle. Eighty mussels were injected with 100 µl (containing 10^7 cells/ml) of Gram positive bacteria *Micrococcus lysodeikticus* into the adductor muscle and additional eighty mussels were injected with the same concentration of Gram negative bacteria *Vibrio anguillarum*, kindly donated by Philippe Roch from Université de Montpellier 2 (France). The remaining 80 mussels were injected with 100 µl filter sea water (FSW) and were used as controls. After the challenge, mussels were returned to the tanks and maintained at 15ºC. Samples were collected at 1, 3, 6, 24 and 72 hours.

From each sampling point, hemolymph (1-2 ml) was withdrawn from the adductor muscle of each animal with a disposable syringe and a piece of adductor muscle and digestive gland were also collected and pooled in four pools of four individuals from each of the three mussel groups. The hemolymph was centrifuged at 2500xg for 15 minutes at 4ºC. The pellet with hemocytes, as well as adductor muscle and digestive gland were resuspended in Trizol reagent (Invitrogen) and RNA was extracted according to the manufacturer's protocol.

3. Q-PCR analysis

To determine the expression of MgC1q in different larval developmental stages in different tissues of adult mussel and to evaluate and quantify its relative expression at different times after bacterial injection, a real time SYBR Green PCR assay was carried out using cDNAs obtained from 5 µg of the original RNA, after reverse transcription was performed using SuperScript II RNAase H-Reverse Transcriptase (Invitrogen).

Quantitative PCR assays and data analysis were performed using a 7300 Real Time PCR System (Applied Biosystems). The 25 µl PCR mixture included 12.5 µl of SYBR Green PCR master mix (Applied Biosystems) with 0.5 µl of 10 µM primers pairs (0.2 µM final concentration) designed for the selected sequences (A10C1q2F-Q-PCR: 5'-TCACTTGGACCATCTGCGTAGA-3' and A10C1q2R Q-PCR: 5'-GGAGCATGCGTCGTCGTACT-3'), and 1 µl of a 1:10 cDNA dilution. Amplification was carried out at the standard cycling conditions of 95ºC for 10 min, followed by 40 cycles of 95ºC 15 s and 60ºC for 1 min. The comparative CT method (2-ΔΔCT method) was used to determine the expression level of analyzed genes [55]. The expression of the candidate genes was normalized using 18S as a housekeeping gene detected by the specific primers 18S-mussel Q-PCR F: 5'-CACTGAAGGAATCAGCGTGTCT-3' and 18S-mussel Q-PCR R: 5'-CGTAATCAACGCGAGCTTATGA-3'. Fold units were calculated by dividing the normalized expression values of infected tissues by the normalized expression values of controls. Results are given as the mean and standard deviation of three replicates and pools.

3. Results

3.1 EST cluster analysis

During the analysis of 1400 reliable ESTs obtained by random clone sequencing of three primary cDNA libraries and two SSH libraries from hemocytes of immunostimulated mussels previously constructed in our laboratory, 402 ESTs grouped in 29 consensus sequences (370 and 24 ESTs from the cDNA and SSH libraries, respectively) showed similarity to an oyster mantle gene 4 or C1q-TNF protein containing a complement-related C1q domain. The most abundant cluster is actually stored in Mytibase [41] with identification code MGC00284 and, after large scale sequencing of a normalized cDNA library, is now composed of 109 ESTs. This cluster was selected for variability analysis, and its consensus sequence, coincident with the consensus sequence of the ESTs obtained by SSH libraries, was used as a reference for characterization and expression analysis.

In addition, from the initial ESTs sequencing approach described above, 8 ESTs putatively identified as a sialic acid binding lectin containing a C1q-domain (cluster MGC00322 in Mytibase) were also recovered. The cDNA and partial genomic sequence of MgC1q were deposited in GenBank under accession numbers FN 563147 and FN 563148.

3.2 Sequence analysis of a new C1q protein (MgC1q)

The complete cDNA consensus sequence of the mussel MgC1q protein presented a 507 bp ORF encoding 169 amino acid residues. The 5' untranslated region (UTR) is 50 bp long, and the 3’ UTR is 69 bp long, with a regular poly (A) tail. The deduced protein structure reveals the predominant organization of C1q-domain proteins, including a leading signal peptide and a C-terminal C1q domain, but without a collagen-like region. The signal peptide was localized in the N-terminal region and is composed of 22 amino acid residues. A complete C1q domain was identified in the ORF sequence, from the 25th to the 168th amino acid residues (Fig.1). The well defined cleavage site indicates that the mature protein can be released as soluble protein.

A multiple sequence alignment of the deduced C1q-domain amino acid sequence with those described in other marine invertebrate species, including mollusks as well as humans, revealed a relative similarity (Fig. 2), with 32-27% identity compared to the C1q-domain of the C1qRD protein of *R. decussatus* and the C1q-domain of Sialic acid binding lectin of *Helix pomatia* and 8 % identity compared to the C1q-domain of the C1q-like protein of *L. japonica* and the Cerebellin protein of *H. sapiens*.

There are five highly conserved residues throughout the C1q family. In addition, the eight invariant residues observed in human c1qDC proteins were conserved in MgC1q, but not in sialic acid binding proteins. The deduced mature protein has 145 amino acid residues with a predicted molecular mass of 16.2KDa and estimated isoelectric point of 7.95.

Neighbor-Joining phylogenetic trees constructed using the amino acid sequences of the two C1q-contaning consensuses of *M. galloprovincialis* together with C1qDC sequences of vertebrate and invertebrate species revealed that mussel C1q is similar to clam RdC1q as well as C1q-gliacolin of *C. gigas*, but separated from other bivalve C1q sequences (Fig. 3). The complexity of the evolutionary origin of C1q domains is too complex to be addressed by a simple analysis. The sequence variability between C1q family members in the same species is surely higher than the interspecific homologues C1q-DC genes. Here, we simply highlight the fact that MgC1q has two homologues in oysters and clams as demonstrated by collapsing the branches under the 50% bootstrap reproducibility threshold.

3.3 Genomic analysis

After amplification of genomic DNA from different mussels with the pair of primers (A10C1q complete 2F/R) and cloning of PCR products, the resulting clones were randomly selected for complete sequencing in the forward and reverse directions to analyze gene structure (Supplementary Figures). In addition to the features described above, the genomic organization of MgC1q showed the presence of one intron of 426 bp (Fig. 1). The intron variability was also analyzed after sequencing in forward and reverse direction with specific the primers (C1q mussel INT 2F/ C1q A10 complete R).

A preliminary assessment of polymorphism levels was performed on 25 partial genomic MgC1q sequences obtained from 7 different mussels. The alignment of a 577 nucleotids from the exonic region showed that all the sequences were different at least in one nucleotide position (Supplementary Figure 1). We have to remark that this level of polymorphism is probably overestimated due to the error of the DNA polymerase used for the fragment amplification. Nevertheless, 29 parsimony-informative sites (out of 577 positions) show nucleotide variation in at least two independent sequences. We also detected one indel of 11 bp fallen at the 105 sequence position.

The MgC1q intron showed also a peculiar variability (Supplementary Figure 2). From the analysis of 28 sequences we could detect 22 parsimony-informative sites (out of 441 positions), a ratio comparable to that of the exon sequence. Four different indels contribute to the intron variability.

Although possibly biased by errors introduced during the cloning phase, these data indicate substantial sequence variability and lead us to analyze the ESTs codifying MgC1q. through specific algorhythms. Similar variability was also found at intra-individual level after analysis of DNA and cDNA sequences obtained from one single mussel (data not shown).

3.4 Inference of adaptive molecular features

We multialigned the 109 ESTs forming the cluster MGC00284, likely codified by a single gene, to perform selection analysis (Fig 4). The tree was inferred using K80 + I as the best-fit model of evolution for the MGC00284 dataset identified by a hierarchical likelihood ratio test (hLRT) comparison. The number of segregating sites (S), average number of nucleotide differences per site (π), as well as nucleotide diversity based on the proportion of segregating sites ($\theta_{W)}$ for these transcripts are listed in Table 1. From the same table we can infer that the hypothesis of neutrality within the loci is rejected based the Tajima' D, the Fu and Li's D and the Zeng & Fu & Shi & Wu's E tests. On the contrary, the H test could not reject the null hypothesis of neutral evolution. Moreover there is no evidence of recombination among the MGC00284 cluster (Maxchi test, p 0.01) [56]. These preliminary data allows us to perform a phylogeny-based test for positive selection. Three alternative programs were used to infer molecular diversity events. CodeML, from the software package PAML [53], uses a numerical optimization algorithm to maximize the log-likelihood values under a specific model of evolution. The M1a model (nearly neutral) estimates a single parameter, $p0$, the frequency of conserved sites with $\omega 0 = 0$ and the remaining sites with frequency $p1$ ($p1 = 1 - p0$) assuming $\omega 1 = 1$. The M2a model (positive selection) adds a class of positively selected sites with frequency $p2$ (where $p2 = 1 - p1 - p0$), with ratio $\omega 2$ estimated from the data. Thus, whereas M1a estimates a single parameter ($p0$), M2a estimates three parameters ($p0$, $p1$, and $\omega 2$). Following the instructions of the PAML manual [53], this pair of models seems more able to classify positively selected sites in LRTs. Sites 1, 6, 12, 19, 40, 43, 44, 46, 54, 71, 88, 93, 88, 120, and 145 are likely to be under positive selection, with a naïve posterior probability (empirical Bayes, NEB) > 0.95.

The SLR (sitewise likelihood-ratio method [54] uses a site by site approach to test for neutrality using the entire alignment to determine evolutionary distances common to all sites. At the end of the test, a necessary correction for multiple testing is completed. This analysis seems to be more stringent, as only 3 sites (40, 71, and 120) are under positive selection, with an ω value of 4.514, 6.113, and 4.450, respectively. Two of these sites (40 and 120) are in a region predicted to be completely exposed in the 3D structure, signaling possible host-pathogen interaction regions (Fig. 4).

3.5 Expression analysis by Q-PCR

Analysis of MgC1q expression during the ontogeny of mussels through the larval developmental stages showed that MgC1q was expressed in all developmental larval stages, from 2h post fertilization with a notable increase after 1 month and continuing to increase until 3 months (Fig. 5A). However, the relative expression was lower than in oocytes and sperm.

In addition, MgC1q was expressed at different levels in all analyzed tissues of naïve adult mussels (Fig. 5B). The highest expression of MgC1q was detected in hemocytes, followed by gonads in both female and male individuals, digestive gland, gills, muscle and mantle tissues.

Experimental infections of adult mussels with Gram positive or Gram negative bacteria induced significant changes in the expression pattern of MgC1q. MgC1q expression was highly increased (22 and 18 fold increase, respectively) early in the infection (1 h post-infection) in the digestive gland of both Gram positive and Gram negative infections, decreasing at 3, 6, 24 and 72h post-infection. However, in muscle tissues, the highest expression level was observed at 3 h post-infection, showing 7 and 8 fold increases after infection with Gram positive and Gram negative bacteria, respectively (Fig. 6A). In hemocytes, the highest expression level was also observed at 1 h post-infection with both Gram positive and Gram negative bacteria, showing a 2 and 16 fold increases, respectively, compared to non-infected controls (Fig. 6B).

4. Discussion

In the present work, we have characterized a new gene, MgC1q, identified from SSH and cDNA libraries previously prepared from immunostimulated mussels. Based on sequence homology, molecular architecture characteristics, and similarity in domain structure, MgC1q can be classified as a new member of the C1q family and therefore integrated into the C1q-TNF superfamily. Although the sequence homology between the C1q and the TNF families is known to be relatively modest, the crystal structure of the conserved hydrophobic core and residues responsible for trimer assembly through conserved hydrophobic surfaces allows the inclusion of these proteins in the same superfamily. Indeed, the C1q family seems to be a growing branch of the C1q-TNF superfamily tree, begging the question whether C1q family proteins can perform immune functions similar to those described for TNF [4]. In human and zebrafish, as vertebrate models, the C1q-DC family is comprised by about 30 to 50 genes [57, 58]. In amphioxus, 50 genes can be enumerated [59], but in the sea urchin genome, only 7 genes are described [24] and just 2 were found in *Ciona intestinalis* [60]. From a preliminary analysis of Mytibase, the mussel EST database [41], we can identify at least 100 different C1q-DC transcripts, which is a conservative estimate keeping in mind the multi-individual origin of our data. Further analysis of these transcripts is being performed to confirm their single genome origin.

MgC1q appears to be an immune-related gene. The importance of this molecule in the immune system of mussels is suggested by: (1) the sequence homology and characteristic architecture of the C1q domain; (2) the fact that its expression was up-regulated in SSH and cDNA libraries derived from infected mussels; (3) the expression changes observed by Q-PCR after experimental infections of adult mussels with Gram positive and Gram negative bacteria. In addition, elevated levels of MgC1q expression were observed soon after bacterial injection, which is in agreement with the profile of pro-inflammatory proteins such as TNF. MgC1q expression was detected at different levels in all analyzed tissues. The highest expression was detected in hemocytes, suggesting that this C1q-DC molecule could be involved in host defense, as has been reported previously in other organisms. It is well known that bivalves have an open circulatory system. Circulating hemocytes are fundamental elements of the defense response, since they are responsible for cellular and humoral immune innate defense reactions, including phagocytosis after migration to different organs and infiltration of injured tissues. The presence of MgC1q in non-immune tissues could be due to the migration of hemocytes to those organs. However, more focused research should be performed to validate this hypothesis.

C1q-DC transcripts have been also recently identified in other bivalve mollusks. The transcript identified in the zhikong scallop *C. farreri* has certain similarities with MgC1q, containing a C1q domain without a collagen region. CfC1q-DC was also found to be constitutively expressed in a wide range of host tissues, but with lower presence in hemocytes. In the hemocytes, peak expression in *C. farreri* after *in vivo* and *in vitro* infection with Gram negative bacteria was reported at 6 h post infection [22], in contrast to the MgC1q peak observed at 1 h post infection in this study. A general decrease in expression after the first hour post-infection has been already reported for other immune-related genes, particularly in mollusks infected by parasites of the genus *Perkinsus* [61]. Moreover, RdC1q, a C1q-DC protein described recently with a C1q domain without a collagen region involved in host defense, followed a similar expression pattern in carpet clams *R. decussatus* infected by *P. olseni* [62]. This evidence supports the idea that host expression of C1q-DC is not restricted to bacterial infection responses, but also occurs in response to other pathogens, such as parasites. Together with the modulatory effect of bacterial infection, expression of MgC1q was detected early in larval development, suggesting a pro-survival role for this gene.

The MgC1q sequences analyzed in mussels at intra and inter-individual level are extraordinarily diverse not only at the genomic level, but also at the cDNA level. Such elevated variability recalls the diversification of nonself recognition molecules already observed in the gastropod mollusk *Biomphalaria glabrata* [63]. One unique sequence out of 6935 ESTs from a cDNA library from *C. farreri* was previously reported as a C1q-DC protein [22]. In contrast, we found a relatively high prevalence and sequence variability of this mussel transcript, that has not been identified before in any bivalve mollusk. Variability analysis in the primary sequences of MgCq1 reveals that several single nucleotide polymorphisms are nonsynonymous substitutions. A more specific study using molecular adaptive analysis methods suggests that substitutions at three particular sites may have undergone positive selection based on a tertiary model prediction. Classically, C1q is known to be the target recognition protein of the classical complement pathway, playing a crucial role in adaptive immunity. However, C1q is also considered to be a versatile Pattern Recognition Protein (PRP), able to bind directly to pathogens by engaging a broad range of Pathogen–Associated Molecular Patterns (PAMPs) via its C1q domain, triggering rapid phagocytosis enhancement [5, 6]. Due to its ability to recognize and bind to pathogen surface molecules acting as a lectin, C1q is also able to activate the lectin pathway, as has been observed in lower vertebrates or jawless fish such as lamprey [7]. This property has led to C1q being considered to be a major connecting link between classical pathway-driven innate immunity and acquired immunity [8]. Originally, innate and adaptive immunity were considered to be two independent mechanisms. At present, they appear to be more intricately linked with innate immunity playing a key role in stimulating the subsequent clonal adaptive immune response [64, 65]. New discoveries of mechanisms used in lampreys and mollusks suggest that a reconsideration of our fundamental views is now essential [26, 27]. While further studies are needed to define the function of MgC1q, our results confirm that MgC1q may act as a Pattern Recognition Molecule (PRM) in the innate immune system of mussels. Due to its similarity to lamprey C1q, it may also be able to activate the lectin pathway. This could be the first indication that a more highly evolved immune system could be present in bivalves than has been previously recognized. Pathogen associated lectin patterns can vary based on the environment, and the variability observed in MgC1q in mussels can be an efficient strategy to counteract possible pathogens.

The other C1q-DC protein identified in this study bears a sequence with high similarity to a Sialic acid binding lectin, containing a C1q domain. Sialic acid binding immunoglobulin-like lectins are a family of acidic sugars expressed in animals of the deuterostome lineage and are the largest family of endogenous vertebrate receptors that recognize glycoconjugates containing sialic acids. These proteins promote cell-cell interactions and regulate the functions of cells in the innate and adaptive immune systems through glycan recognition, potentially playing a role in triggering endocytosis following pathogen recognition [66].

The fact that MgC1q clusters with the sialic acid binding immunoglobulin-like lectin in the phylogenetic tree might be in line with the hypothesis that the classical pathway emerged after the lectin pathway and that the activation mechanism of the latter was partially conserved. In addition, the high variability of the MgC1q cluster could indicate that a somatic diversification of these nonself recognition molecules occurred, probably differently than the mechanism for diversification of vertebrate immunoglobulins, as has been proposed for other invertebrate immune genes, such as FREPs or vCRL1 [63, 67]. It is known that complement components are not variable between individuals. However, as has been reported [67], early in the history of chordate evolution, components of the complement system showed a high degree of variation. Thus, MgC1q could emerge as a lectin, functioning as an initial recognition molecule in the complement system showing a high degree of diversification and acting as a primitive immunoglobulin. Further studies will help us to confirm this hypothesis.

Acknowledgements

This work has been funded by the EU Integrated Project FOOD-CT-2005-007103 and AGL2008-05111/ACU from the Spanish Ministerio de Ciencia e Innovación. Camino Gestal wishes to acknowledge additional funding from the Spanish Ministerio de Educación y Ciencia through the "Ramón y Cajal" Contract.

References

[1] Medzhitov R, Janeway CA Jr. Innate immunity: the virtues of a nonclonal system of recognition. Cell 1997; 91(3):295-298.

[2] Nokana M, Smith SL. Complement system of bony and cartilaginous fish. Fish Shellfish Immunol 2000; 10(3): 215-228.

[3] Boshra H, Li J, Sunyer JO. recent advances on the complement system of teleost fish. Fish Shellfish Immunol 2006; 20: 239-262.

[4] Kishore U, Gaboriaud C, Waters P, Dhrive AK, Greenhough TJ, Reid KB, Sim RB, Arlaud GJ. C1q and tumor necrosis factor superfamily: modularity and versatility. Trends Immunol 2004; 25:551-561.

[5] Medizhitov R, Janeway CA. Decoding the patterns of self and nonself by the innate immune system. Science 2002; 12: 296(5566): 298-300.

[6] Bohlson SS, Fraser DA, Tenner AJ. Complement proteins C1q and MBL are pattern recognition molecules that signal immediate and long-term protective immune functions. Mol Immunol 2007; 44(1-3): 33-43.

[7] Matsushita M, Matsushita A, Endo Y, Nakata M, Kojima N, Mizuochi T, Fujita T. Origin of the classical complement pathway: lamprey orthologue of mammalian C1q acts as a lectin. Proc Natl, Acad Sci USA 2004; 101(27): 10127-10131.

[8] Kishore U, Reid KBM. C1q: structure, function and receptors. Immunopharmacol 2000; 42, 15-21.

[9] Tom Tang Y, Hu T, Arterburn M, Boyle B, Bright JM, Palencia S, Emtage PC, Funk W D. The complete complement of C1q-domain-containing proteins in *Homo sapiens*. Genomics 2005; 86(1): 100-111.

[10] Jones EY, Stuart DL, Walker NP. Structure of tumor necrosis factor. Nature 1989; 338 (6212): 225-228.

[11] Eck MJ, Sprang SR. The structure of tumor necrosis factor-ά at 2.6 Å resolution. Implications for receptor binding. J Biol Chem 1989; 246: 17595-17605.

[12] Shapiro L, Scherer PE. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. Curr Biol 1998; 8: 335-338.

[13] Bodmer JL, Schneider P, Tschopp J. The molecular architecture of the TNF superfamily. Trends Biochem Scie 2002; 27, 19-26.

[14] Van den Berg RH, Faber-Krol MC, Sim RB, Daha M R. The first subcomponent of complement, C1q, triggers the production of IL-8, IL-6 and monocyte chemoattractant peptide-1 by human umbilical vein endothelial cells. J Immunol 1998; 161: 6924-6930.

[15] Bording S, Tan X. C1q arrest the cell cycle progression of fibroblasts in G(1) phase: role of the cAMP/PKA-1 pathway. Cell Signal 2001; 13:119-123.

[16] Amanullah A, Azam N, Balliet A, Hollander C, Hofman B, Fornace A, Liebermann D. Cell signalling: cell survival and a Gadd45-factor deficiency. Nature 2003; 424: 741 discussion 742.

[17] Yamada M, Oritani k, Kaisho T, Ishikawa J, Yoshida H, Takahashi I, Kawamoto S, Ishida N, Ujiie H, Masaie H, Botto M, Tomiyama Y, Matsuzawa Y. Complement C1q regulates PLS-induced cytokine production in bone marrow-derived dendritic cells. Eur J Immunol 2004; 34: 221-230.

[18] Kimura A, Sakaguchi E, Nonaka M. Multi-component complement system of Cnidaria: C3, Bf, and MASP genes expressed in the endodermal tissues of a sea anemone, *Nematostella vectensis*. Immunobiol 2009; 214 (3): 165-178.

[19] Castillo M, Goodson M, McFall-Ngai M. Identification and molecular characterization of a complement C3 molecule in a lophotrozoan, the Hawaiian bobtail squid *Euprymna scolopes*. Dev Comp Immunol 2009; 33: 69-76.

[20] Lao, H., Sun, Y. Yin, Z., Wang, J., Chen, Ch., Weng, S., He, W., Guo, Ch., Huang, X., Yu, X., He, J. Molecular cloning of two C1q-like cDNAs in mandarin fish *Siniperca chuatsi*. Veterinary Immunol Immunopathol 2008; 125: 37-46.

[21] Chen B, Gui J. Identification of a novel C1q family member in color crucian carp (*Carassius auratus*) ovary. Comp Biochem Physiol 2004; 138: 285-293.

[22] Zhang H, Song L, Li Ch, Zhao J, Wang H, Qiu L, Ni D, Zhang Y. A novel C1q-domain-containing protein from Zhikong scallop *Chlamys farreri* with lipopolysacharide binding activity. Fish Shellfish Immunol 2008; 25: 281-289.

[23] Mei J, Gui J. Bioinformatic identification of genes encoding C1q-domain-containing proteins in zebrafish. J Genet Genomics 2008; 35(1):17-24.

[24] Hibino T, Loza-Coll M, Messier C, Majeske AJ, Cohen AH, Terwilliger DP, Buckley KM, Brockton V, Nair SV, Berney K, Fugmann SD, Anderson MK, Pancer Z, Cameron RA, Smith LC, Rast JP. The immune gene repertoire encoded in the purple sea urchin genome. Dev Biol 2006; 300: 349-365.

[25] Gestal C, Roch P, Renault T, Pallavicini A, Paillard C, Novoa B, Oubella R, Venier P, Figueras A. Study of diseases and the immune system of bivalves using molecular biology and genomics. Rev Fish Sci 2008; 16: 131-154.

[26] Flajnik MF, Du Pasquier L. Evolution of innate and adaptive immunity: can we draw a line?. Trends Immunol 2004; 25(12): 640-644.

[27] Vivier E, Malissen B. Innate and adaptative immunity: specificities and signalling hierarchies revisited Nat Immunol 2005; 6(1): 17-21.

[28] Espiritu DJ, Watkins M, Dia-Monje V, Cartier GE, Cruz LJ, Olivera BM.Venomous cone snails: molecular phylogeny and the generation of toxin diversity. Toxicon 2001; 39(12):1899-916.

[29] Moy GW, Vacquier VD. Bindin genes of the Pacific oyster *Crassostrea gigas*. Gene 2008; 423(2):215-20.

[30] Moy GW, Springer SA, Adams SL, Swanson WJ, Vacquier VD. Extraordinary intraspecific diversity in oyster sperm bindin. Proc Natl Acad Sci USA 2008; 105(6):1993-8.

[31] Canesi L, Gallo G, Gavioli M, Pruzzo C. Bacteria-hemocyte interactions nad phagocytosis in bivalves. Microsc Res Tech 2002; 57: 469-476.

[32] Olafsen JA. Role of lectins (C-reactive protein) in denfense of marine bivalves against bactiera. Adv Exp Med Biol 1995; 371A: 343-348.

[33] Ordás MC, Novoa B, Figueras A. Modulation of the chemiluminiscence response of Mediterranean mussel (*Mytilus gallorpovincialis*) haemocytes. Fish Shellfish Immunol 2000; 10: 611-622.

[34] Tafalla C, Novoa B, Figueras A. Production of nitric oxide by mussel (*Mytilus gallorpovincialis*) hemocytes and effect of exogenous nitric oxide on phagocytic functions. Comp Biochem Physiol B 2002; 132: 423-431.

[35] Gueguen Y, Cadoret JP, Flament D, Barreau-Roumiguière C, Girardot AL, Garnier J, Hoareau A, Bachère E, Escoubas A. Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas*. Gene 2003; 303:139-145.

[36] Kang YS, Kim YM, Park KLl, Cho SK, Choi KS, Cho M. Analysis of EST and lectin expression in hemocytes of Manila clams (*Ruditapes phylippinarum*) (Bivalvia: Mollusca) infected with *Perkinsus olseni*. Dev Comp Immunol 2006; 30 (12): 1119-31.

[37] Gestal C, Costa M, Figueras A, Novoa B. Analysis of differentially expressed genes in response to bacterial stimulation in hemocytes of the carpet-shell clam *Ruditapes decussatus*: identification of new antimicrobial peptides. Gene 2007; 406: 134-143.

[38] Pallavicini A, Costa MM, Gestal C**,** Novoa B, Dreos R, Venier P, Figueras A. Sequence variability of myticins identified in haemocytes from mussels stimulated with Vibrio and Poly I:C suggests ancient host-pathogen interactions. Dev Comp Immunol 2008; 32: 213-226.

[39] Costa MM, Dios S, Alonso-Gutierrez J, Romero A, Novoa B, Figueras A. Evidence of high individual diversity on miticin C in mussel (*Mytilus gallorpovincialis*). Dev Comp Immunol 2009; 33: 162-170.

[40] Prado-Alvarez M, Rotllant J, Gestal C, Novoa B, Figueras A. Characterization of a C3 and factor B-like in the carpet-shell clam, *Ruditapes decussatus*. Fish Shellfish Immunol 2009a; 26(2): 305-315.

[41] Venier P, De Pittà C, Bernante F, Varotto L, De Nardi B, Bovo G, Roch P, Novoa B, Figueras A, Pallavicini A. Lanfranchi G. Mytibase: a knowledgebase of mussel (*M. galloprovincialis*) transcribed sequences. BMC Genomics 2009; 10:72.

[42] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 1997; 25: 4876-4882.

[43] Dyrløv Bendtsen J, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004; 340:783-795.

[44] Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 1995; 11(6):681-4.

[45] Rost B, Yachdav G, Liu J. The PredictProtein Server. Nucleic Acids Research 2004; 32(Web Server issue):W321-W326.

[46] Pollastri G, McLysaght A. Porter: a new accurate server for protein secondary structure prediction. Bioinformatics 2005; 21(8), 1719-20.

[47] Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 2007; 24: 1596-1599.

[48] Finn RD, Tate J, Mistry J, Coggill PC, Sammut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res 2008; Database Issue 36:D281-D288.

[49] Strauss WM. Preparation of genomic DNA from mammalian tissue. Coligan JE, Bierer B, Margulies DH, Shevach EM, Strober W, Coico R, edithors. Protoc Immunol 2001;Chapter 10:Unit 10.2.

[50] Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res 2004; 14: 988-995.

[51] Posada D, Crandall KA. Modeltest: testing the model of DNA substitution. Bioinformatics 1998; 14(9): 817-818.

[52] Rozas J, Sánchez-Del Barrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorohism analyses by the coalescent and other methods. Bioinformatics 2003; 19: 2496-2497.

[53] Yang Z. PAML 4: a program package for phylogenetic analysis by maximum likehood. Mol Biol Evol 2007. 24: 1586-1591.

[54] Massingham T, Goldman N. Detecting amino acid sites under positive selection and purifying selection. Genetics 2005; 169: 1853-1762.

[55] Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 2001; 25: 402–408.

[56] Smith JM. Analyzing the mosaic structure of genes. J Mol Evol 1992; 34: 126-129.

[57] Mei J, Gui J. Bioinformatic identification of genes encoding C1q-domain-containing proteins in zebrafish. J Genetics and Genomics 2008; 35 (1): 17-24.

[58] Tang YT, Hu T, Arterburn M, Boyle B, Bright JM, Palencia S, Emtage PC, Funk WD. The complete complement of C1q-domain-containing proteins in *Homo sapiens* Genomics 2005; 86 (1): 100-111.

[59] Huang S, Yuan S, Guo L, Yu Y, Li J, Wu T, Liu T, Yang M, Wu K, Liu H, Ge J, Yu Y, Huang H, Dong M, Yu C, Chen S, Xu A. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. Genome Res 2008; 18: 1112-1126.

[60] Azumi K, De Santis R, De Tomaso A, Rigoutsos I, Yoshizaki F, Pinto MR, Marino R, Shida K, Ikeda M, Ikeda M, Arai M, Inoue Y, Shimizu T, Satoh N, Rokhsar DS, Du Pasquier L, Kasahara M, Satake M, Nonaka M. Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "waiting for Godot". Immunogenetics 2003; 55 (8): 570-581.

[61] Chintala MM, Bushek D, Ford SE. Comparison of in vitro-cultured and wild-type *Perkinsus marinus*. II. Dosing methods and host response. Dis Aquat Org 2002; 51(3): 203-216.

[62] Prado-Alvarez M, Gestal C, Novoa B, Figueras A. Differentially expressed genes of the carpet Shell clam *Ruditapes decussatus* against *Perkinsus olseni*. Fish Shellfish Immunol 2009b; 26(1): 72-83.

[63] Zhang SM, Adema CM, Kepler TB, Loker ES. Diversification of Ig superfamily genes in an invertebrate. Science 2004; 305: 251-254.

[64] Fearon DT, Locksley RM. The instructive role of innate immunity in the acquired immune response. Science 1996; 272(5258): 50-53.

[65] Hoffmann JA, Kafatos FC, Janeway CA, Ezekowitz RAB. Phylogenetic perspectives in innate immunity. Science 1999; 284: 1313–1318.

[66] Angata T. Molecular diversity and evolution of the Siglec family of cell-surface lectins. Mol Diversity 2006; 10:555-566.

[67] Kürn U, Sommer F, Hemmrich G, Bosch, TCG, Khalturin K. Allorecognition in urochordates: identification of a highly variable complement receptor-like protein expressed in follicle cells of Ciona. Dev Comp Immunol 2007; 31(4): 360-371.

[68] Pollastri G, Martin AJM, Mooney C, Vullo A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information". BMC Bioinformatics 2007; 8:201.

Figure and table captions

Figure 1. Complete ORF and deduced amino acid sequence of *M. galloprovincialis* C1q (MgC1q). The initial and stop codons as well as the splicing sites are marked in bold. The predicted signal peptide is underlined, and the probable cleavage site (▲) pointed out. The intron is showed in italics, and the globular C1q domain is marked in grey.

Figure 2. Alignment of C1q-domains from MgC1q, MgSBL, and other C1qDCs from different species. Each C1q domain exhibits a ten-stranded β-sandwich, numbered B-1 to B-10. ▲: Indicates the eight conserved residues in human C1qDC proteins. ▲: Indicates the five highly conserved residues in the C1q family. Amino acid residues that are highly conserved (75% at least) are shaded in dark, while similar amino acids are shaded in gray.

Figure 3. Neighbor-Joining tree of the C1q domains shows the phylogenetic relationships of MGC00284 and MGC00322 (●), and the C1qDC sequences of some representative species. *: protein sequences translated from ESTs, Δ: TNF sequences from Cnidaria used as out-group. Branches corresponding to partitions reproduced in less than 50% of bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (10,000 replicates) are shown next to the branches.

Figure 4. Multiple alignment of the nonredundant dataset of the translated ESTs from the MGC00284 cluster. The sites under positive selection identified by “sitewise likelihood-ratio” (*SLR*) analysis are shaded (these sites are in the coil and exposed regions) (A)

Secondary structure prediction by Porter [46]: H = helix, E = strand. C = remainder. The underlined positions represent β-strand secondary structures according to the crystal structure of the human adiponectin C1q domain (PDB 1c28); (B) Relative Solvent Accessibility prediction by PaleAle [68]: B=completely hidden b=partly hidden, e=partly exposed E=completely exposed.

Figure 5. A. Quantitative expression of MgC1q during the ontogeny of mussels through the larval developmental stages. Results are presented as mean ± SD. Bars represent the expression of MgC1q relative to 18S transcript levels. (L): larval stages; (D): days; (m): months. **B.** Quantitative expression of MgC1q analyzed in different tissues of naïve adult mussels. Results are presented as mean ± SD. Bars represent the expression of MgC1q relative to 18S transcript levels.

Figure 6. Quantitative expression of MgC1q analyzed in adult mussels injected with Gram positive (*M. lysodeikticus*) and Gram negative (*V. anguillarum*) bacteria after 1, 3, 6, 24, and 72 h post infection. Results are presented as mean ± SD. Data represent the fold increase in expression of MgC1q relative to 18S transcript level of infected mussels, referred to not infected controls. **A.** Quantitative expression of MgC1q in digestive gland and muscle of adult mussels injected with Gram positive (*M. lysodeikticus*) and Gram negative (*V. anguillarum*) bacteria **B.** Quantitative expression of MgC1q in hemocytes of adult mussels injected with Gram positive bacteria (*M. lysodeikticus*). **C.** Quantitative expression of MgC1q analyzed in hemocytes of adult mussels injected with Gram negative bacteria (*V. anguillarum*).

Supplemental Figure 1. Multiple alignment of genomic exon sequences of the different MgC1q clones from 7 individual mussels. The parsimony-informative sites in the aligned box are highlighted in yellow and indels in red.

Supplemental Figure 2. Multiple alignment of genomic intron sequences of the different MgC1q clones from 8 individual mussels. The parsimony-informative sites in the aligned box are highlighted in yellow and indels in red.

Tables

Table 1. Polymorphisms and test for neutrality in MgC1q (MGC00284). S, number of segregating sites; pS, number of segregating sites per site; π, average number of nucleotide differences per site; θ_W, nucleotide diversity based on the proportion of segregating sites; D_T, Tajima's D; $D_{F\&L}$, Fu and Li’s D; E, Zhu, Gao and Tytgat test; H, Fay and Wu test. Numbers in parentheses are p values. Significant results ($p \leq 5\%$) are indicated by '*'.

Table 1

Number of alleles	70
S	109
pS	0.21499
π	0.041772
θ_W	0.041772
D_T	-1.509 (0.029)*
$D_{F\&L}$	-5.799 (0.000)*
E	-1.794 (0.005)*
H	-6.782 (0.345)

Figure 1

```
actttggtagattagttaggttttcgcaagaatagatagctgaacgtattatgaggatga    60
                                                   M  R  M        3
atgttaattggatttctgttttgactgttatggtactggtttccagtaagtttgatattg   120
N  V  N  W  I  S  V  L  T  V  M  V  L  V  S                      18
ctgtattgataatgcatatatctttttttgaaaataatacgagctgaaaggctatcaaatc   180
aattttaattgacgtataatttttaattactactctaaaaggataacacgcatacacaga   240
caggttgtgtaggtacagtataaataagaaagaatgcaagattttcctcaatgagacaac   300
aattcaacgacattcaaaaagtctcaaaaggcacgccgaggtcaacgtagaatcttcaat   360
ttctcaataacccgattcatatctttattctcacatgtgaaatcagatttttccatatca   420
tgaacttgtcaattgtcatatgtcaccattaaacttaatctgcggattatattctttgta   480
caaaactatgacgacccatacatattcattcaaattagattttttttcttacagaatctac   540
                                                     K  S  T     21
ctgcctcattgcatttctgcctatatgagcgaaagcaaagctggacagagcaattctat   600
  C▲ L  I  A  F  S  A  Y  M  S  E  S  K  A  G  Q  S  N  S  I     41
tattgatggatcaaccctgatcttcgataaagtagaaataaattcagggtcagattacag   660
  I  D  G  S  T  L  I  F  D  K  V  E  I  N  S  G  S  D  Y  S     61
cgtcttcacgggaaagttttctgttccatcttccggcatttatgcgttcacttggaccat   720
  V  F  T  G  K  F  S  V  P  S  S  G  I  Y  A  F  T  W  T  I     81
ctgcgtagattcacgttatacaaatccccctggaagattaaactacggggagtatggaac   780
  C  V  D  S  R  Y  T  N  P  P  G  R  L  N  Y  G  E  Y  G  T    101
agaattaatgatgggcagcacaaaaatcggtgttctacatacagatacagagacaaagta   840
  E  L  M  M  G  S  T  K  I  G  V  L  H  T  D  T  E  T  K  Y    121
cgacgacgcatgctccactggatttgtcatcagatatgtatcatcaggcaatcaagttta   900
  D  D  A  C  S  T  G  F  V  I  R  Y  V  S  S  G  N  Q  V  Y    141
cgtcagaaataactacgcccaccaaggcaaacttctaagcaaggaaagtcagaccagaac   960
  V  R  N  N  Y  A  H  Q  G  K  L  L  S  K  E  S  Q  T  R  T    161
aactttctctggatggaaaatgcaataaaaaggcaggttcaacggaacaaagttcgccag  1020
  T  F  S  G  W  K  M  Q  -                                     169
atctaaaatgtaatattcaaaatacaatttgaagatgg                        1058
```

Figure 2

Figure 2

```
                                      B-1                               B-2      B-3    B-4        B-5
                                     ------                            ------     ---    ----   -----------               -
M. galloprovincialis MGP00284 24/144 AFSAYMSES- -KAGQSNSI- --IDGSTLIF DKVETNSGSD YSVFTGKFSV PSSGIYAFTW TICVDSRYKN PPGRLNYGEY
M. galloprovincialis MGP00322 50/181 AFTAWLSKT- -TTLGT---- ----NHAVVY NKLLLNKGNA YNVHTGHFTA PVRGLYLLSA SIMASPR--- --------KH
C. farreri ABS50435 28/153           AFSAGLTHT- -QPLTS---- ----GETVRF DEVFTNIGGG YDLSSGNFKA PKGGVYAFTI HGYDANT--- --------DK
C. hortensis CAD83837 26/150         AFTAVLDKN- -LVLEN---- ----GDTLIP NKILINYGGG YNDKTGIFTA PKSGIYHLGV HAQTSLQ--- --------SN
H. pomatia ABF00124 25/149           AFSAAIDRE- -QTFDA---- ----NQVVIY DIVITNHGNA YDNSTGLFTA PVDGMYSFQL NLLTIKE--- --------KE
H. sapiens NP 000482 115/250         AFSATRTI-- NVPLRR---- --D--QTIRF DHVITNMNNN YEPRSGKFTC KVPGLYYFTY H----AS--- -------SRG
H. sapiens NP 004788 106/242         AFSVGLE--- -TYVT----- --IPNMPIRF TKIFYNQQNH YDGSTGKFHC NIPGLYYFAY H----ITV-- -------YMK
H. sapiens AAQ89429 72/105           AFAAV----R SHHHEPAGET GNGTSGAIYF DQVLVNEGGG FDRASGSFVA PVRGVYSFRF H---VVKV-- ------YNRQ
M. edulis Q6UQ16 109/235             AFSAELTHP- IENIAA---- ----EEIAHF DKVRVNSGHA YHADTGKFVA PEEGFFYFSV TICT-KR--- --------DS
L. japonicum BAD22833 119/237        AFSAKVDA-- YK--PA---- --A-GSPVTF NVIITNPQ-P LQPRHGKFTC AHPGVYYFTL Y----AHP-- -------SEQ
C. gigas CX069305 470/78*            SFYASMTTV- -IPIQG---- ---PHQRIVY DEIKTNNGNG YNESTEIFTV PESGTYVFTW TTFSYVE--- -----EY--V
A. irradians CB416625 112/444*       AFSATIQSD- MTPIAS---- ----NEVIIF DNVVTSVGGG YDPFIGRFQA PVAGYYHFVV SIMS-YS--- --------SH
P. fucata AAZ76258 35/155            TFTNI----K SGH------- ------IFVF DRTNTNHGNH YDSSNGNFTA PRDGPYVFTW T---VVT--- ------QDKY
R. decussatus EY255091 512/120*      GFHAYLSGP- -RCYHT---- ----GNPIKF DVASVNIGNR YSVHTGVYVV ATPGLYFFTW TVAAEEK--- -----NW--F
                                      ▲                             ▲      ▲            ▲      ▲ ▲

                                                B-6                 B-7           B-8       B-9                      B-10
                                                ------         ------------ -----------     -------                 ------
M. galloprovincialis MGP00284 24/144 GTELMMGSTK IG-VLHTDTE T-TYDDACS- -----TGFVI RYVSSGNQVY VR-N--NYAH QGKLLSKESQ TRTTFSGWKM
M. galloprovincialis MGP00322 50/181 -VALSMVRNG IN-FDALYGD SRNVNKYSSQ -----SRSYT VMLQRGDLVW MRTF--PGYE GQVLVGYSHA LYNSFSGVLL
C. farreri ABS50435 28/153           AMWIELKKND EL-LVSIS-- GYNSHSSAGN -----SVIIY LSAGENVHVQ AR-----PNQ SFSLFGKVDQ VYATFSGYMI
C. hortensis CAD83837 26/150         -LWLALYHND NY-VFSIY-- GRQTEYSDGG -----ANAAI LPLKKGDKVH VK-----ARD KSSLLGRPDN IYTTFTGFRL
H. pomatia ABF00124 25/149           -GWLELVHNG QL-KVSVY-- AKQDSTYDSS -----SNSVI IKMKEGDRVN VR-----AHK KSGLFGRDDE LYNTFSGHFL
H. sapiens NP 000482 115/250         NLCVNLMRGR ERAQKVVTFC DYAYNTFQ-- ---VTTGGMV LKLEQGENVF LQ--ATDK-N SLLGMEG-AN S--IFSGFLL
H. sapiens NP 004788 106/242         DVKVSLFKKD KA---MLFTY DQYQENNVD- ---QASGSVL LHLEVGDQVW LQVYGEGERN GLYADN-DND S--TFTGFLL
H. sapiens AAQ89429 72/105           TVQVSLMLNT WP-VISAFAN D----PDVT- -REAATSSVL LPLDPGDRVS LRL-----RR GNLLGG---W KYSSFSGFLI
M. edulis Q6UQ16 109/235             ILEMALHVND HD-EMIIHAD AEHLELG-CA -----SNSEI VHLQKGDHVE VVKH--GADG VPPFYIHTMS ---TFTGFML
L. japonicum BAD22833 119/237        HLVIMFMKNG KVIS---KVC G--EKYNA-- ----ISGSVV VSLKAGDEVW LELEA-PY-I NFYFNVH-RD S--VFSGYIL
C. gigas CX069305 470/78*            RLEIVVA--- -G-TIFGGTL SDTQETGDAD ---TETGIIV VNAQEGDQVF VRTQ--LRGP GNGRLYVEFE CASTFSGWKI
A. irradians CB416625 112/444*       FIEIELVRDG TL-LCRARAS QTYNAMGACA -----SN--- IHLNVGSDVW VRHR--SSSG D---YI---- ----------
P. fucata AAZ76258 35/155            SINVELFRNN QR-VSRAQAD G----QPGP- QNSSGSNTVV LDLKTGDVVN LRCGNWMNQD QRQLFG---D TYSTFSG---
R. decussatus EY255091 512/120*      VSDLMVN--- -G-VLRGNTM TDSYGNGKGN GIHPATGFVL VNVKRGDHVY IR-F--VTGH -GCIIKSNRM TRSTFSGWRL
                                                                                                                      ▲ ▲
```

Figure 3

Figure 3.

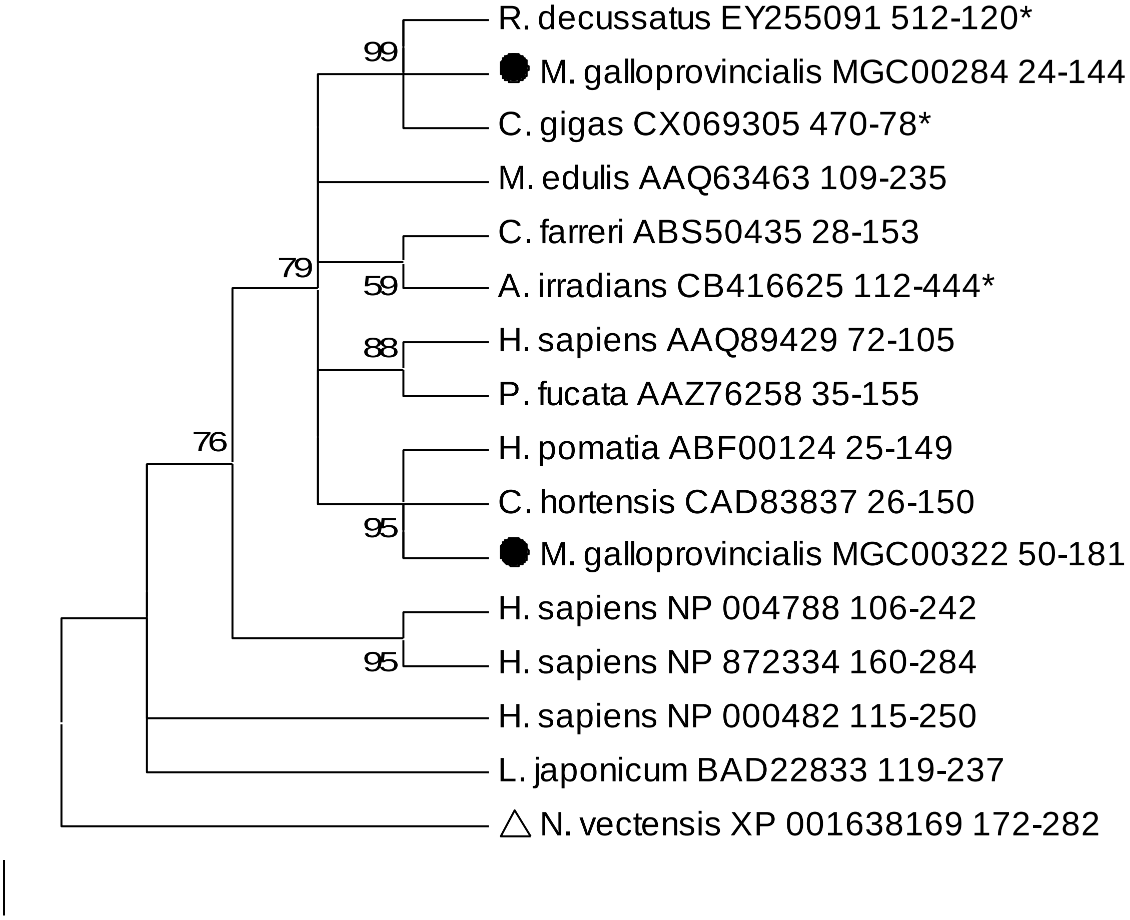

99
79
59
88
76
95
95
R. decussatus EY255091 512-120*
● M. galloprovincialis MGC00284 24-144
C. gigas CX069305 470-78*
M. edulis AAQ63463 109-235
C. farreri ABS50435 28-153
A. irradians CB416625 112-444*
H. sapiens AAQ89429 72-105
P. fucata AAZ76258 35-155
H. pomatia ABF00124 25-149
C. hortensis CAD83837 26-150
● M. galloprovincialis MGC00322 50-181
H. sapiens NP 004788 106-242
H. sapiens NP 872334 160-284
H. sapiens NP 000482 115-250
L. japonicum BAD22833 119-237
△ N. vectensis XP 001638169 172-282

Figure 4

Figure 4

```
Hae0401A12 MRMNVNWISVLTVMVLVSKSTCLIAFSAYMSESKAGQSNSIIDGSTLIFDKVETNSGSDYSVFTGKFSVPSSGIYAFTWTICVDSRYKNPPGRLNYGEYGTELMMGSTKIGVLHTDTETTYDDACSTGFVIRYVSSGNQVYVRNNYAHQGKLLSKESQTRTTFSGWKMQ
Hae0405H11 ....................S..................................................................................................................................................
Hae0501D01 .............................................A...........................................................................................................................
Hae0404G06 .................................................................................................K..............................................D........................
Hae0304B12 ..............................................................................................................................................G..........................
Hae0102B03 .......................................................................................T...............................K.................................................
Hae0504F01 ............L..........................................................................T...............................K.................................................
Hae0407C06 I...............................................L......................................P...............................K.................................................
Hae0503E10 I.....R................................................................................T...............................K.................................................
Hae0404G04 I......................................................................................T...............................K.................................................
Hae0405D12 I......................................................................................T...............................K..............T..................................
Hae0403H10 I......................................................................................T...............................K..................................R..............
Nor0164M01 I..............................K.......................................................T...............................K.................................................
Hae0406E03 I......................................................................................T...............................K..................R..............................
Hae0108E02 .......................................................................................T...............................K.................................................
Hae0406E11 .......................................N.............I........V..R....Y................T....K..........................K.................................................
Hae0105H03 .......................................N.............I........V..R....Y.........T......T....K..........................K.................................................
Hae0104B04 ......................................DN.............I........V..R....Y................T....K..........................K.................................................
Hae0305H08 ................A......................N.............I........V..R....Y................T....K..........................K.................................................
Hae0205A10 .............T.........................N.............I........V..R....Y................T....K..........................K.................................................
Hae0104A05 .........................S.............N.............I........V..R....Y................T....K..........................K.................................................
Hae0406B03 .......................................N.............I........V..R....Y.....L..........T....K..........................K........................................A........
Hae0407G12 .......................................N...E.........I........V..R....Y................T....K..........................K.................................................
Hae0305F04 I............................L.........N.............I........V..R....Y................T....K..........................K.................................................
Hae0304G11 ........P..I...........................N.............I........V..R....Y................T....K..........................K.........A.......................................
Hae0502F05 ...........I...........................N.............I........V..R....Y..V.............T....K..........................K.................................................
Nor0110D11 ...........I...........................N.............I........V..R....Y................T....K..........................K.......................................G.........
Nor0162K22 ...........I...........................N.............I........V..R....Y........I.......T....K..........................K.................................................
Hae0402H05 ...........I...........................N.............I........V..R....Y................T....K..........................K.................................................
Hae0501C10 .......T...I...........................N.............I........V..R....Y................T....K..........................K.................................................
Hae0501A08 ...........I...........................N.....A.......I........V..R....Y................T....K..........................K.................................................
Hae0107D05 V......................................N.............I........V..R....Y..........R.....T....K..........................K........................S........................
Hae0504G12 ...........I...........................N.............I........V..R....Y..........R.....T....K..........................K.................................................
Nor0161E22 .......................................N.............I........V..R....Y................T....K....K.....................K.................................................
Hae0306A04 .................................................................R.T.................T...............................K.................................................
Hae0504D10 .....................................................A...........R.T.................T...............................K.................................................
Hae0407G02 ..........S......................................................R.T.................T...............................K...............................................T.
Hae0402G04 .....S...........................................................R.T.................T....G..........................K....Y............................................
Hae0305C09 ..................E........................R.........I................Y................T...............................K.................................................
Hae0302G06 ..................E........................R.........I................Y...........A....T...............................K.................................................
Hae0306B09 ..................E........................R.........I...Y.............................T...............................K.................................................
Nor0163B19 ..................E.............G......Y...R.........I..................................T...............................K.................................................
Hae0107D11 ..................E............D..........E..........I......................................S.........................RE.................................................
Hae0207C01 .....D............E............D..........E..........I......................................S.........................RE.................................................
Nor0144L22 ..................E.........H..D..........E..........I......................................S.........................RE................................................R
Hae0403B11 ..................E............D..........E..........I......................................S.........................RE................................................R
Hae0303E05 ..................E............D..........A..........I......................................S..............K..........RK.................................................
Hae0303C04 ..................E............D..........A..........I......................................S..............K........A.RK.................................................
Nor0142F21 ..................E............D..........A..........I......................................S..............K..........RK........S........................................
Nor0152C04 ..................E........V...D..........A..........I......................................S..............K..........RK.................................................

A                                 CCHHHHHHHHHHHCCCCCCCCCCEECCCEEEEECCCCEECCCCEEECCCCEEEEEEEEEEEEECCCCCCCEEEECCCEEEEEECCEEEEEEEEECCCCCCCCEEEEEEEEEECCCEEEEEECCCCCCEEECCCCCCCEEEEEECCC
B                                 EbbBBBbBeEeEEEEEEEEEEEebBeBeebbbbEbEbBebEeBeBeBEbeBBBBBBBbBbebEEbEEEEEEbEEEEebBbBbBEEEeBeBBBbEbEEEEEebbEBBBBBbeBEEEbeBBBbBeebEEebBbbeEEEebBBBBBBEEE
```

Figure 5

Figure 5

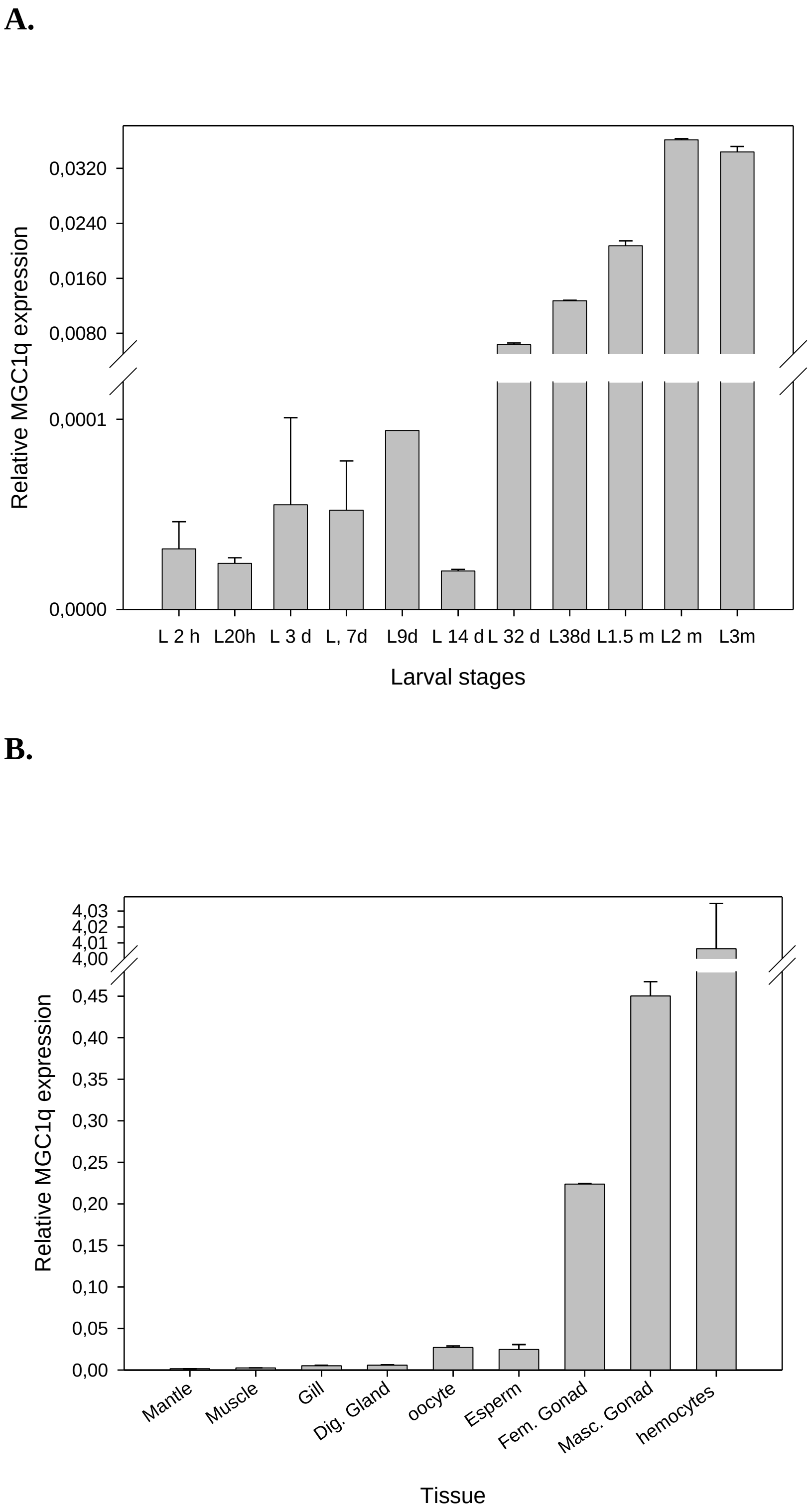
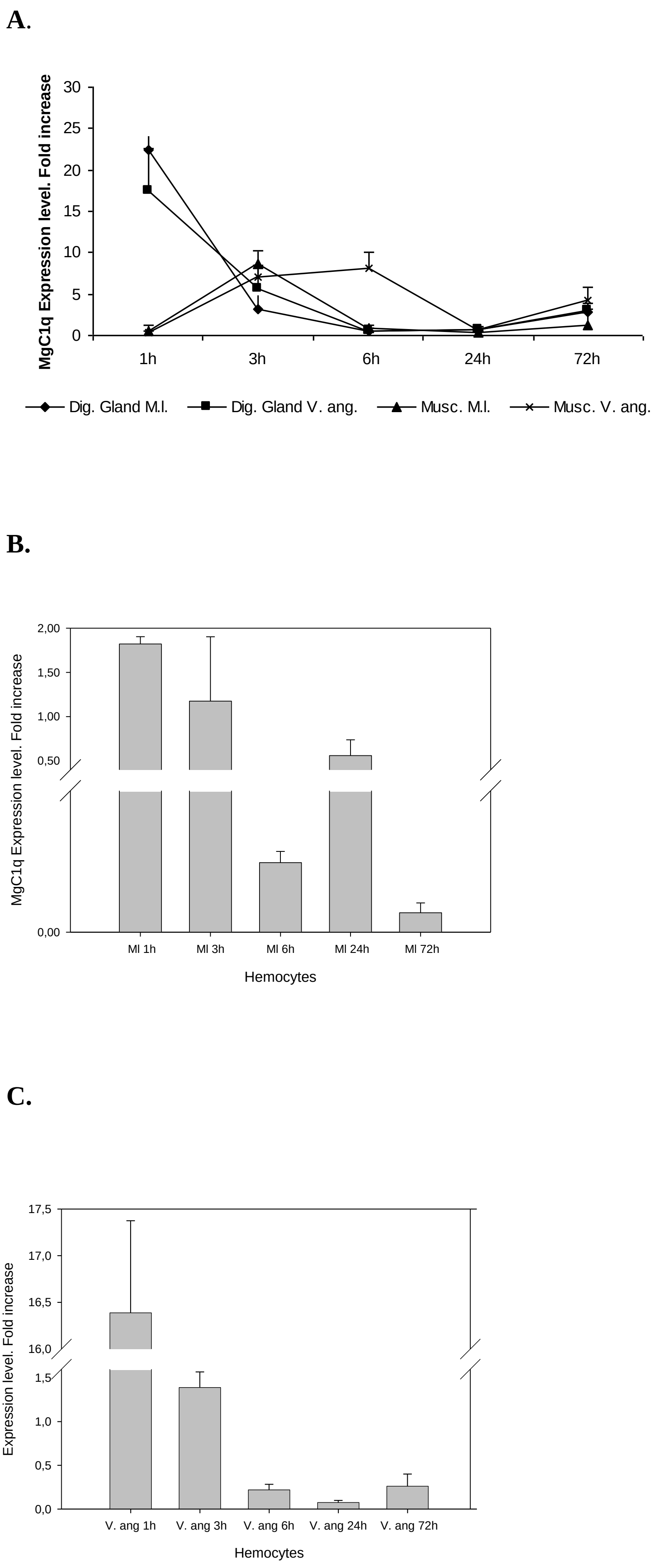


Figure 6

Figure 6



Supplementary Figure 1

Click here to download Supplementary Material for online publication: sup Fig 1 NEW-2.doc

Supplementary Figure 2

Click here to download Supplementary Material for online publication: sup Fig 2 NEW.doc