Genome Biology and Evolution Advance Access published July 21, 2015 doi:10.1093/gbe/evv133

GENOME BIOLOGY AND EVOLUTION

SMBE

RESEARCH ARTICLE

Identification and characterization of a novel family of cysteine-rich peptides (MgCRP-I) from *Mytilus galloprovincialis*

Marco Gerdol[1], Nicolas Puillandre[2], Gianluca De Moro[1], Corrado Guarnaccia[3], Marianna Lucafò[1], Monica Benincasa[1], Ventislav Zlatev[3], Chiara Manfrin[1], Valentina Torboli[1], Piero Giulio Giulianini[1], Gianni Sava[1], Paola Venier[4], Alberto Pallavicini[1*]

[1] Department of Life Sciences, University of Trieste, Via Giorgieri 5, 34127 Trieste, Italy

[2]Muséum National d'Histoire Naturelle, Département Systématique et Evolution, ISyEB Institut (UMR 7205 CNRS/UPMC/MNHN/EPHE), 43, Rue Cuvier, 75231 Paris, France

[3]Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Padriciano 99, 34149 Trieste, Italy

[4] Department of Biology, University of Padova, Via Ugo Bassi 58/B, 35131 Padova, Italy

*Author for Correspondence: Alberto Pallavicini, Department of Life Science, University of Trieste, Via Giorgieri 5, 34127 Trieste, Italy, telephone number: +39 0405588736, FAX: +39, e-mail address: pallavic@units.it

Data deposition

MgCRP-I nucleotide sequences were deposited at GenBank under the accession IDs KJ002647-KJ002676 and KR017759-KR017770.

GENOME BIOLOGY AND EVOLUTION

Abstract

We report the identification of a novel gene family (named MgCRP-I) encoding short secreted cysteine-rich peptides in the Mediterranean mussel *Mytilus galloprovincialis*. These peptides display a highly conserved pre-pro region and a hypervariable mature peptide comprising six invariant cysteine residues arranged in three intramolecular disulfide bridges. Although their cysteine pattern is similar to cysteines-rich neurotoxic peptides of distantly related protostomes such as cone snails and arachnids, the different organization of the disulfide bridges observed in synthetic peptides and phylogenetic analyses revealed MgCRP-I as a novel protein family. Genome- and transcriptome-wide searches for orthologous sequences in other bivalve species indicated the unique presence of this gene family in *Mytilus* spp.

Like many antimicrobial peptides and neurotoxins, MgCRP-I peptides are produced as pre-propeptides, usually have a net positive charge and likely derive from similar evolutionary mechanisms, i.e. gene duplication and positive selection within the mature peptide region; however, synthetic MgCRP-I peptides did not display significant toxicity in cultured mammalian cells, insecticidal, antimicrobial or antifungal activities. The functional role of MgCRP-I peptides in mussel physiology still remains puzzling.

Keywords

Toxin; antimicrobial peptide; bivalve mollusc; mussel; transcriptome.

Introduction

Marine ecosystems are characterized by an astonishing species diversity, with over 2 million different eukaryotic species belonging to various *phyla* estimated to compose the marine fauna (Mora, et al. 2011). Thus, marine organisms and environments can be regarded as a virtually unlimited source of bioactive compounds, either produced through complex biochemical synthetic reactions or gene-encoded peptides (Mayer, et al. 2011).

Nowadays, computer-assisted data mining coupled with the advent of next-generation sequencing (NGS) technologies allows the *in silico* identification of bioactive molecules also in non-model marine organisms (Li, et al. 2011; Sperstad, et al. 2011). The quick increase of bivalve transcriptome datasets (Suárez-Ulloa, et al. 2013b) and the recent genome sequencing of the oysters *Crassostrea gigas* (Zhang, et al. 2012) and *Pinctada fucata* (Takeuchi, et al. 2012) further broadens the horizons of genetic and genomic studies in bivalve mollusks. Due to their relevance as sea food and sentinel organisms, significant RNA sequencing efforts, both with 454 and Illumina technologies, have been performed on *Mytilus* spp. (Bassim, et al. 2014; Craft, et al. 2010; Freer, et al. 2014; Gerdol, et al. 2014; González, et al. 2015; Philipp, et al. 2012; Romiguier, et al. 2014; Suárez-Ulloa, et al. 2013a). Moreover, a recently released unrefined genome of the Mediterranean mussel *Mytilus galloprovincialis* further extends the molecular data available for this species (Nguyen, et al. 2014). The bioinformatic analysis of the mussel data has already contributed to the discovery of important immune-related molecules, including pathogen-recognition receptors, signaling intermediates and antimicrobial peptides (AMPs) (Gerdol, et al. 2012; Gerdol, et al. 2011; Gerdol and Venier ; Rosani, et al. 2011; Toubiana, et al. 2013; Toubiana, et al. 2014). As reported in this paper, large-scale bioinformatic analyses can also drive the discovery of novel gene families encoding peptides with unique chemico-physical properties and/or sequence patterns.

Actually, cysteine-rich peptides (CRPs) encompass a large and widespread group of secreted bioactive molecules, heterogeneous in primary sequence and structural arrangement, with different

functional roles and present in almost all living organisms, from bacteria to fungi, animals and plants (Gruber, et al. 2007; Marshall, et al. 2011; Taylor, et al. 2008). Invertebrate CRPs are particularly abundant and they have been frequently related to the immune defense against potential pathogens (Mitta, et al. 2000b). According to the number of cysteine residues and their arrangement in the tridimensional space, many families of cysteine-rich AMPs have been described in invertebrates, as for instance in crustaceans (Bartlett, et al. 2002; Destoumieux, et al. 1997), insects (Bulet and Stöcklin 2005) and arachnids (Ehret-Sabatier, et al. 1996; Fogaça, et al. 2004). In *Mytilus* spp., different families of cysteine-rich AMPs have been progressively discovered starting from the mid '90s. Peptides similar to arthropod defensins were purified from active fractions of hemolymph almost contemporarily in *M. edulis* and *M. galloprovincialis* (Charlet, et al. 1996; Hubert, et al. 1996), together with different novel AMPs which whose structure and biological activities were characterized in the following years. Those included mytilins (Mitta, et al. 2000a), myticins (Mitta, et al. 1999) and the strictly antifungal mytimycins. Only very recently other AMP families were described in mussel, either by cloning from hemolymph cDNAs libraries, such as in the case of myticusins (Liao, et al. 2013), or by detection in high throughput sequencing datasets, in the case of mytimacins and big defensins (Gerdol, et al. 2012).

Other evolutionarily related CRPs are animal venom components which possess neurotoxic properties, since they can selectively block various types of ion channels for predation or defense (Froy and Gurevitz 2004; Rodríguez de la Vega and Possani 2005). Notably, spider and scorpion venoms contain an extraordinary mixture of cysteine-rich peptides whose complexity has been only recently fully appreciated by 'omics approaches (Ma, et al. 2009; Zhang, et al. 2010). Even within the Mollusca phylum some species have developed a lethal venom arsenal to be used for predation: marine gastropods of the genus *Conus* indeed use a modified radula as a sting to inject and paralyze their prey with a powerful venom cocktail, mostly of peptidic nature (Olivera, et al. 2012). Due to their biological properties, many CRPs have been studied to guide the development of new drugs

for therapeutic applications in both human and veterinary medicine (Adams, et al. 1999; Otero-González, et al. 2010; Saez, et al. 2010).

Despite having physico-chemical properties similar to cysteine-rich AMPs and toxins, certain animal CRPs lack the expected activities and are instead involved in diverse functions: among these, Kunitz-type (Ranasinghe and McManus 2013) and Kazal-type (Rimphanitchayakit and Tassanakajon 2010) proteinase inhibitors represent two widespread groups.

The abundance and diversity of the CRPs described in protostomes is remarkable and, given the poor genomic knowledge of many taxonomic groups, a large part of these peptides probably still remain to be uncovered. In this paper, we report the application of a genome- and transcriptome-scale approach to the identification of sequences encoding novel cysteine-rich peptides from the Mediterranean mussel *M. galloprovincialis*. In agreement to nomenclature criteria reported elsewhere (Gerdol and Venier), we present the new MgCRP-I family, characterized by a conserved pre-pro region and an highly variable mature peptide with six conserved cysteine residues organized in the consensus $C(X_{3-6})C(X_{1-7})CC(X_{3-4})C(x_{3-5})C$. We investigated the organization and evolution of mussel MgCRP-I genes and pseudogenes, as well as the main features and possible functional roles of the encoded peptides.

Materials and methods

Identification of MgCRP-I sequences in mussel transcriptomes

The *M. galloprovincialis* Illumina transcriptomes available at the NCBI Sequence Reads Archive (retrieved in February, 2015) were assembled with Trinity v.2014-07-17 (Grabherr, et al. 2011) and with the CLC Genomics Workbench 7.5 (CLC Bio, Aarhus, Denmark), using default parameters. Following translation of the assembled contigs into the 6 possible reading frames with EMBOSS TranSeq (Rice, et al. 2000) we investigated the virtual mussel proteins for the presence of the C-C-CC-C-C signature, allowing a spacing between cysteine residues of one to ten amino acids, using a Perl script developed in-house (available upon request to the corresponding author). Matching sequences were aligned with MUSCLE (Edgar 2004) to generate a HMMER v3.0 profile (Eddy 2011) which was then used to retrieve partial-matching cases within the assembly (e-value cutoff $1x10^{-5}$). The procedure was re-iterated until no additional matches could be retrieved. Sequences showing an identity higher than 95% at the nucleotide level were considered as redundant and collapsed in a single consensus sequence, unless they were confirmed by genomic evidence (see section below). With just two exceptions (see the discussion section), all the sequences retrieved matched the presence of at least one $C(X_{3-6})C(X_{1-7})CC(X_{3-4})C(x_{3-5})C$ motif.

Identification of MgCRP-I genes in the mussel genome

The *M. galloprovincialis* genomic contigs (Nguyen, et al. 2014) were downloaded from GenBank and scanned for the presence of MgCRP-I genes as follows: (i) genes were identified based on BLASTn identity (Altschul, et al. 1990) to the previously identified MgCRP-I transcripts (e-value threshold of $1x10^{-5}$); (ii) genomic scaffolds were translated into the six possible reading frames with the EMBOSS Transeq tool (Rice, et al. 2000) and novel MgCRP-I loci were identified with HMMERv 3.0 (Eddy 2011).

The genes identified were manually annotated with mRNA and CDS traces, based on: (i) MUSCLE alignment between genomic contigs and the corresponding assembled transcripts, whenever available; (ii) mapping of the available *M. galloprovincialis* sequencing reads (see above), with the CLC Genomics Workbench *large gap mapping* tool; (iii) refinement of splice site positions with Genie (Reese, et al. 1997). An example of the results of the annotation procedure is shown in **Supplementary Figure 1 (Supplementary material online)**. Results obtained from the genome and transcriptome analyses were compared and redundant results (identity percentage higher than 95%) were removed, unless multiple gene copies were confirmed in the mussel genome (e.g. the presence of paralogous genes was tolerated).

MgCRP-I protein sequence analysis

Protein translations of mussel genes and transcripts identified with the strategy mentioned above were further analyzed as follows: the presence of a signal peptide was detected with SignalP 4.0 (Petersen, et al. 2011), and discriminated from transmembrane domains with Phobius (Käll, et al. 2004). Potential sites of post-translational proprotein convertase cleavage were identified with ProP 1.0 (Duckert, et al. 2004). Possible post-translational C-terminal cleavage sites by carboxypeptidase E or by peptidylglycine, α-amidating monooxygenase were detected with ELM (Dinkel, et al. 2011). The subcellular localization was predicted (for full-length peptides only) with TargetP 1.1 (Emanuelsson, et al. 2007). Isoelectric point and molecular weight of the predicted mature peptides was calculated at ExPASy (http://web.expasy.org/compute_pi/). Structural homologies with proteins deposited in the RCSB Protein DataBank database were investigated by Phyre2 (Kelley and Sternberg 2009).

The probabilities of codon bias for the six conserved cysteines and for the arginine residue responsible of the post-translational pro region cleavage were calculated assuming a binomial distribution, based on the codon usage inferred from the *M. galloprovincialis* transcriptome (Gerdol, et al. 2014) and using the tool *cusp* included in the EMBOSS package (Rice, et al. 2000).

We used PAML 4.7 (Yang 2007) and the graphical interface PAMLX 1.2 (Xu and Yang 2013) to test whether some sites in the codon-based alignments of MgCRP-I nucleotide sequences were under positive selection. In detail, only full-length coding sequences were processed and two site models were compared: M1, which assumes that the d_N/d_S ratio along the sequence ranges from 0 to 1 (purifying selection to neutral drift), and M2, which assumes that a few sites have a d_N/d_S ratio (i.e., $\omega > 1$; positive selection). The likelihoods of the two models were compared using a likelihood ratio test (LRT) with a χ^2 distribution, with 2 degrees of freedom. The Empirical Bayes approach was used to calculate the posterior probabilities (PP) for site classes. Positive selection was concluded at PP > 0.95.

Comparative genomics analyses

The NGS Illumina transcriptome data available for 71 bivalve species were downloaded from the Sequence Read Archive (SRA). The full list and the corresponding Bioproject accession IDs are shown in **Supplementary Table S1, supplementary material online**. The bivalve sequence datasets were independently *de novo* assembled with the CLC Genomic Workbench 7.5 (CLC Bio, Aarhus, Denmark). All transcriptomes were translated into the six possible open reading frames with EMBOSS Transeq (Rice, et al. 2000) and significant similarity with MgCRP-I proteins was assessed with BLASTp (e-value threshold 0.01) and HMMER v 3.0 using the protein profile mentioned above (p-value threshold 0.01).

The complete UniProtKB/Swiss-Prot protein sequence database and the whole set of peptides predicted from the fully sequenced genomes of *C. gigas* (release 9) (Zhang, et al. 2012) and *P. fucata* (v 1.0) (Takeuchi, et al. 2012) were screened for the presence of the C-C-CC-C-C pattern with a custom Perl script, without any constraint about the spacing between cysteine residues. Only full-length sequences shorter than 100 amino acids and showing a signal peptide by SignalP 4.0 (Petersen, et al. 2011) were selected. Sequences bearing more than 7 cysteine residues within the mature region were considered as characterized by more complex disulfide arrays and

discarded. The possible presence of mis-annotated CRP-I-like genes in *C. gigas* and *P. fucata* was evaluated by performing the same analyses on the genomic scaffolds translated into the 6 possible reading frames with EMBOSS TranSeq.

Phylogeny and evolutionary tests

We used all the available MgCRP-I proteins, their orthologous sequences identified in *Mytilus edulis* (MeCRP-I), *Mytilus californianus* (McCRP-I) and *Mytilus trossulus* (MtCRP-I), and the positive hits resulting from the data mining, to infer the phylogenetic relationships among sequences bearing a similar cysteine signature. Given the high sequence diversity, only the signal peptides, as predicted using SignalP 4.0 (Petersen, et al. 2011), were retained to facilitate sequence alignment. Following alignment with Muscle (Edgar 2004) and manual refinement, maximum likelihood analyses were performed with RaxML (Stamatakis 2006) as implemented on the CYPRES Portal (www.phylo.org/portal2), using the RAxML-HPC2 on TG Tool. The robustness of the nodes was assessed with a bootstrapping procedure of 100 replicates. Because the evolutionary relationships of the peptides included in the analysis with other peptides were unknown, no outgroup was considered and a mid-point rooting strategy was applied. A similar analysis was performed with a dataset that included only the *M. galloprovincialis* CRP-I peptides. Partial MgCRP-I peptides with an incomplete signal peptide were excluded from the analyses.

Peptide synthesis, oxidative folding and disulfide mapping

The peptides MgCRP-I 7 (26 aa) and MgCRP-I 9 (26 aa) were selected for solid phase peptide synthesis (SPPS) for their primary sequence features: a low number of hydrophobic residues, especially if not consecutive, facilitates synthesis and improves solubility during purification and folding, proline acts as secondary structure breaker during SPPS and the presence of aromatic amino acids allows easy UV quantitation. The two peptides were synthesized according standard solid-phase Fmoc chemistry using 4 equivalents of HCTU/Fmoc-Xaa-OH/DIEA

(0.95/1.00/1.90) with respect to the resin loading (Tentagel S-Trt, 0.2mmol/g (Sigma-Aldrich, St. Louis, MO)). The synthesis was semi-automatically performed with a customized Gilson Aspec XL peptide synthesizer (Middleton, WI) on a 0.05 mmol scale. Cysteines were manually added as N-α-Fmoc-S-trityl-L-cysteine pentafluorophenyl ester in order to minimize racemization. After cleavage from the resin the peptides were precipitated with diethylether, washed and freeze-dried. The peptides reduced by TCEP treatment were purified by RP-HPLC on a semipreparative Zorbax 300SB-C18 9.4x250mm column (Agilent, Santa Clara, CA) using a gradient from A (0.1% TFA in water) to B (0.1%TFA/60% acetonitrile in water) in 100 min at 4 ml/min. The calculated K* (retention factor) is 4.12 assuming a shape selectivity factor (S) for the peptides of 0.25*Mw0.5.

Peptide fractions from the semipreparative RP-HPLC were checked by electrospray mass spectrometry (amaZonSL iontrap, Bruker (Billerica, MA)) and fractions with at least 95% purity were quantified by UV absorbance at 280 nm and immediately diluted at 0.1 mg/ml in either of the following refolding buffers: (i) RefoldA: 0.2 M Tris-HCl, 2 mM EDTA, 10 mM GSH, 1 mM GSSG, pH 8, previously degassed with argon bubbling; (ii) RefoldB: 50 mM NaOAc, 1 mM EDTA, 1 mM GSH, 0.1 mM GSSG, 2M $(NH_4)_2SO_4$, pH 7.7. The oxidative refolding proceeded for 18 h at 4°C, was quenched by TFA addition and finally checked by LC-MS analysis.

All proteolysis reactions were carried out at 37 °C for 18-48 h in sodium acetate buffer (100 mM, pH 5.5) containing 1 M GuHCl and 5 mM $CaCl_2$. The purified MgCRP-I 7 peptide (60 µg) was dissolved in 90 µl of buffer and trypsin (3 µg) was added. A second aliquot of MgCRP-I 7 (60µg in 90µl) was incubated for 48 h at 37 °C in the presence of chymotrypsin (6 µg). Digestions of MgCRP-I 9 with trypsin and chymotrypsin were carried out in the same conditions. The digestions were quenched using formic acid (1% final) and the proteolytic fragments were fractionated by RP-HPLC (column Jupiter C18, 1x50mm, Phenomenex (Torrance, CA) using a gradient from water/0.1% formic acid to 60% acetonitrile and analyzed by LC-MS/MS (amaZonSL, Bruker (Billerica, MA)).

Gene expression analysis

The expression levels of selected MgCRP-I genes were evaluated in samples representing hemolymph, digestive gland, inner mantle, mantle rim, gills, foot and posterior adductor muscle. Total RNA was extracted from the tissues of 30 adult specimens (5-7 cm shell length) collected from the Gulf of Trieste, Italy, homogenized in equal quantity in Trizol® (Life Technologies, Carlsbad, CA) according to the manufacturer's protocol. RNA quality was assessed by electrophoresis on denaturing agarose gel and its quantity was estimated by UV-spectrophotometry. cDNAs were prepared using a qScript™ cDNA Synthesis Kit (Quanta BioSciences Inc., Gaithersburg, MD) according to the manufacturer's instructions. Primer pairs were designed to obtain the specific PCR amplicons (**Table 1**), with the exception of the primer pairs co-amplifying the paralogous sequences MgCRP-I 3/25 and MgCRP-I 10/26. The 15 µL PCR reaction mix comprised 7.5 µL of SsoAdvanced™ SYBR® Green Supermix (Bio-rad, Hercules, CA), 0.3 µL of each of the two 10 µM primers and 2 µL of a 1:20 cDNA dilution.

The following thermal profile was used for qPCR amplification in a C1000 thermal cycler (Bio-Rad, Hercules, CA): an initial denaturation step at 95° C for 3', followed by 40 cycles at 95° for 5" and 55° for 30". The products of amplification were analyzed with a 65°/95° C melting curve. The expression of the selected genes was calculated with the delta Ct method; Ct values were corrected based on primer pairs PCR efficiencies using Lin-RegPCR (Ramakers, et al. 2003) and expression values were normalized using the elongation factor EF-1 as a housekeeping gene. Results are shown as the mean with standard deviation of three technical replicates.

Cytotoxicity assays

Human colorectal carcinoma (HT-29), human neuroblastoma (SHSY5Y) and breast cancer (MDAMB231) cell lines were used for the cytotoxicity assays. HT-29 was maintained in RPMI-

1640 and MDAMB231 was maintained in Dulbecco's Modified Eagle's Medium (DMEM): the culture medium was supplemented with 10% (v/v) fetal bovine serum (FBS), penicillin (100 U/mL), streptomycin (100 µg/mL), and L-glutamine 2 mM. SHSY5Y was cultured in DMEM medium supplemented with penicillin (100 U/mL), streptomycin (100 µg/mL), L-glutamine 2 mM and with 10% heat-inactivated fetal bovine serum. Cells were grown at 37 °C in a 95 % air and 5 % CO_2 humidified incubator.

HT-29, SHSY5Y and MDAMB231 were harvested by trypsinization and plated into 96-well culture plates at a density of approximately 1.5×10^4 cells per well. After 24 hrs of incubation, different concentrations of MgCRP-I 7 and 9 (10, 1, 0.1 and 0.01 µM) dissolved in culture medium were added to each well. Then, the samples were incubated 24 hrs at 37 °C in the humidified atmosphere (5 % CO_2). The colorimetric 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) assay was performed to assess the metabolic activity of cells treated as described above. Aliquots of 20 µl stock MTT (5 mg/mL) were added to each well, and cells were then incubated for 4 hrs at 37 °C. Cells were lysed with isopropanol HCl 0.04 N. Absorbance was measured at 540 and 630 nm using a microplate reader (Automated Microplate Reader EL311, BIOTEK® Instruments, Vermont, USA). All measurements were done in six technical replicates, and three independent experiments were carried out.

Insecticidal test

The synthetic peptides MgCRP-I 7 and 9 were dissolved in PBS and injected with a sterile syringe in *Zophobas morio* larvae (~50 mm long, weighting ~500 mg). The control group (n = 10) was injected with a volume of 50 µl PBS. Two experimental groups of larvae for each peptide (n = 10) were injected with 30 and 300 µg peptide/Kg body weight respectively, for an injection volume of 50 µl. Larvae were monitored for signs of neurotoxicity for 48 hours, including lack of movement, twitching and death. During the experimental time course, larvae were not fed and kept

at room temperature. The median lethal dose (LD50) for the two mussel peptides was calculated according to Tedford and colleagues (2001).

Bacterial/fungal strains and minimum inhibitory concentration (MIC assay)

The growth inhibitory effect of MgCRP-I 7 and 9 was tested on *Escherichia coli* ATCC 25922, *Staphylococcus aureus* ATCC 25923, two strains of *Candida albicans* (ATCC 90029 and a clinical isolate), four strains of *Cryptococcus neoformans* (ATCC 90112, ATCC 52816, ATCC 52817 and a clinical isolate), and two strains of filamentous fungi (a clinical isolate of *Aspergillus fumigatus* and *Aspergillus brasiliensis* ATCC 16404). The bacterial inoculum was incubated overnight in Mueller-Hinton Broth (MHB, Difco) at 37°C with shaking. For the assays, the overnight bacterial cultures were diluted 1:30 in fresh MHB and incubated at 37°C with shaking for approximately two hours to obtain a mid-logarithmic phase bacterial culture. Fungi were grown on Sabouraud agar (Difco) plates at 30°C for 48 hours. Fungal suspensions were prepared by picking and suspending five colonies in 5 ml of sterile PBS. The turbidity of the bacterial or fungal suspensions was measured at 600 nm and was adjusted to obtain the appropriate inoculum according to previously derived curves relating the number of colony forming units (CFUs) with absorbance.

Filamentous fungi were grown on Sabouraud agar slants at 30°C for 7 days. The fungal colonies were then covered with 3 mL of PBS and gently scraped with a sterile pipette. The resulting suspensions were transferred to sterile tubes, and heavy particles were allowed to settle. The turbidity of the conidial spore suspensions was measured at 600 nm and was adjusted in Sabouraud broth to obtain an appropriate inoculum.

The antimicrobial activity was evaluated by the broth micro-dilution susceptibility assay performed according to the guidelines of the Clinical and Laboratory Standards Institute (CLSI) and previously described (Benincasa, et al. 2010; Benincasa, et al. 2004). Briefly, two-fold serial dilutions of MgCRP-I 7 and MgCRP-I 9 were prepared in 96-well microplates in the appropriate

medium, to a final volume of 50 µl. Fifty µl of bacterial suspension in MHB, or fungal suspension in Sabouraud, were added to each well to a final concentration of 1-5x10^5 cells/ml for bacteria, and 5x10^4 cells/ml for fungi. Bacterial and fungal samples were then incubated at 37°C for 24h or 30°C for 48h, respectively. The MIC (Minimum Inhibitory Concentration) was taken as the lowest concentration of peptide resulting in the complete inhibition of visible growth after incubation. All tests were performed in triplicate.

Results and discussion

MgCRP-I sequence features

Overall, we identified 67 different MgCRP-I sequences (**Table 2**). More in detail, 48 sequences could be identified in over 2 million genomic contig sequences, which provide a preliminary view of the mussel genome (Nguyen, et al. 2014). Twenty-three of them were also detected as expressed transcripts in publicly available RNA-seq data. In addition, 19 expressed sequences with no match in the genomic contigs were identified, likely corresponding to genes located in genomic regions which are still not covered by the assembly. Overall, 41 sequences can be considered of full length, as the entire CDS from the initial ATG to the STOP codon was represented. The remaining partial sequences (**Table 2**) were lacking either the 5' or the 3' end, due to truncated genomic contigs or low read coverage from RNA-seq data. In addition, BLAST and HMMER approaches revealed at least six pseudogenes with an ORF interrupted by nonsense or frameshift mutations and which lacked expression in RNA-seq experiments (**supplementary table S5, supplementary material online**).

The MgCRP-I peptides are secreted pro-peptides characterized by two features: a conserved pre-pro region and the presence of at least one conserved cysteine array C-C-CC-C-C. In detail, all the members of this family display an unambiguous N-terminal signal peptide cleavage site, followed by a ~15 residues long pro-region ending with a highly conserved dibasic cleavage site for proprotein convertases (KR or, more rarely, RR). While the signal peptide and pro-peptide regions show low sequence variability, the C-terminal region of MgCRP-I corresponding to the putative mature peptide, appears hypervariable (**Figure 1**). The six invariant cysteine residues involved into the formation of intramolecular disulfide bridges are embedded within this highly variable region. The two central cysteine residues (Cys3 and Cys4) are directly linked with a peptidic bond. As a result, the consensus of these peptides can be defined as $C(X_{3-6})C(X_{1-7})CC(X_{3-4})C(x_{3-5})C$ (**Figure 1**

and **2**). We also detected a limited number of protein-coding genes sharing significant sequence similarity with MgCRP-I peptides but which lacked the expected cysteine array (these sequences will be described in detail in the sections below); these, together with non-coding pseudogenes, should be defined as *Mg-CRP-I-like* sequences.

Most MgCRP-I peptides present a short C-terminal extension and, after the 6th cysteine, they often display dibasic amino-acidic motifs which might be the target of post-translational cleavage by carboxypeptidase E, one of the most common modifications observed in neurotoxic peptides from invertebrates such as scorpion venoms (Xiong, et al. 1997) and conotoxins (Fan, et al. 2003; Wang, et al. 2003). Secreted cysteine-rich peptides are among the molecules undergoing the largest amount of different post-translational modifications, as demonstrated by the case of conotoxins (Bergeron, et al. 2013; Craig, et al. 1999) and the case of MgCRP-I peptides could be similar; given the difficulty of obtaining purified peptides from mussel tissues due to their low expression levels, in the absence of proteomic studies we had to rely on *in silico* prediction for the identification of the most likely modification sites.

Based on the predicted proteolytic cleavages, and with a few exceptions, the virtual MgCRP-I mature peptides are 25-38 amino acids long with estimated molecular weight of 2.5-4 KDa. Almost invariably, the mature peptides have a basic isoelectric point (mostly between 8 and 9.5), indicative of positive net charge at physiological pH, which might be balanced by the presence of conserved negatively charged residues in the pro-region (**Figure 1**); this feature might be important for the biological activity of MgCRP-I peptides, as it is maintained in all sequences despite their remarkable sequence variability.

In addition to the standard pre-pro-peptide organization described above, several peculiar transcripts, named "multi-MgCRP-I" encoding precursors characterized by multiple cysteine-rich modules were also identified (**Table 2**). The modules, ranging from 2 to 4 in number, are structurally close to each other, as each Cys6 of the N-terminal module is separated by just 2-4 residues from the Cys1 of the following one. Cysteine-rich domains of multi-MgCRP-I do not show

any peculiarity compared to those of regular, mono-domain peptides (see **supplementary figure S3, supplementary material online**) and the different domains within the same sequence are likely derived by the duplication of a single original module. Although the functional significance of the multi-MgCRP-I peptides is still unknown, the maintenance of cysteine pattern and isoelectric points within each single domain (data not shown) suggest that these long precursors might be post-translationally cleaved into smaller functional peptides. In this case, such a process would be an interesting strategy adopted to achieve the co-expression of different variants, in a similar fashion to other invertebrate AMPs characterized by several sequential tandem repeats of conserved motifs (Casteels-Josson, et al. 1993; Destoumieux-Garzón, et al. 2009; Ratzka, et al. 2012; Rayaprolu, et al. 2010).

Structure of MgCRP-I genes

Despite the low average size of the assembled genomics contigs of *M. galloprovincialis*, we could identify eleven MgCRP-I gene regions corresponding to a full-length coding sequence, from the initial ATG to the STOP codon (MgCRP-I 13, 14, 18, 24, 28, 40, 45, multi-MgCRP-I 2, 11, 12 and 13) and several partial matches, either to the 3' or to the 5' region of the CDS (**Table 2**).

The structure of MgCRP-I genes is conserved, with four exons and three introns (**Figure 2**). The first exon, which could only be annotated in five sequences thanks to the alignment with RNA-seq data, includes part of the 5'UTR region. In most cases, the second exon encodes the first 17 amino acids of the precursor protein, thus comprising most of the signal peptide. A phase-1 intron separates the second and the third exon. The third exon is ~100 nt long and covers the signal peptide cleavage site and most of the propeptide region. The open reading frame is interrupted by a phase-2 intron, which separates the third and the fourth exon. The last exon invariably comprises the final 10 nucleotides of the pro- region, including the highly conserved dibasic precursor cleavage site, the complete cysteine-rich C-terminal region and the entire 3'UTR region.

While the modular organization of the multi-MgCRP-I precursor proteins finds striking similarities to other invertebrate AMP-related cysteine-rich proteins (i.e. the Lepidoptera X-Tox family), their genomic organization is remarkably different: indeed, while the defensin-like motifs in X-Tox are encoded by separate exons (d'Alençon, et al. 2013), all cysteine-rich motifs in multi-MgCRP-I precursors are encoded within a single exon. A schematic representation of the structural organization of MgCRP-I genes and encoded precursor proteins is shown in **Figure 2**.

Gene duplication and positive selection are driving the evolution of the MgCRP-I gene family

The combined genomic/transcriptomic analyses indicate the presence of at least 67 different potentially functional CRP-I loci in the genome of the Mediterranean mussel. Owing to the preliminary nature of the released mussel genome (Nguyen, et al. 2014), this has to be considered as a conservative estimate. The phylogenetic analysis of the CRP-I signal peptide regions (**Figure 3**), evidenced the existence of several highly similar paralogous genes, highlighting the important role of gene duplication events in the evolution of the MgCRP-I gene family, which in some cases appear to have occurred very recently (**supplementary figure S4, supplementary material online**). A number of MgCRP-I pseudogenes with frame-shift or missense mutations were identified in the mussel genome (**supplementary table S5, supplementary material online**) and, at the same time, the frequent sequence truncations caused by the small size of the genomic contigs makes impossible to infer how many of the incomplete MgCRP-I loci are fully functional (see **Table 2**). For the same reason, the presence of common regulatory regions and transposable elements which could have driven the expansion of this gene family will be matter of future investigations.

However, gene duplication is not sufficient by itself to explain some peculiar features of MgCRP-I genes: indeed, the amino acid diversity of the peptide precursors is strikingly higher within the mature peptide region compared to the signal peptide and pro-region which are, in turn, highly conserved (**Figure 1**). This observation suggests an increased rate of mutations within the 4th

exon, and an accelerated evolutionary rate. The likelihood ratio tests we performed to assess this hypothesis strongly support positive selection of the MgCRP-I genes (p-value = $1.717x10^{-8}$). In fact, we could identify nine positively selected sites (PP > 0.95), all located within the mature peptide region, after the cleavage site of the pro-peptide (**Figure 1**). Five out of the six invariable cysteines engaged in disulfide bridges, buried within this hypervariable and positively selected region, undergo site-specific codon preservation (**Figure 4**). This peculiar phenomenon, which was also observed for the arginine responsible of the pro-peptide cleavage, is likely driven by the unique properties of these residues for the maintenance of the tridimensional structure and an efficient biosynthesis and folding of the mature peptide, as suggested for many other protein families, including conotoxins (Conticello, et al. 2001; Steiner, et al. 2013).

Although structurally important codons appear to be somehow protected from variation, the high selective pressure acting on the fourth exon (encoding the entire mature peptide) of MgCRP-I genes in some cases introduced mutations which resulted in the loss of cysteine residues (**supplementary figure S5, supplementary material online**). MgCRP-I 12 and MgCRP-I 23 represent two instructive examples, since the deduced proteins maintain the highly conserved signal peptide and pro-peptide regions, features clearly identifying them as CRP-I related sequences in a phylogenetic analysis though lacking the canonical cysteine array. Indeed, MgCRP-I 12 lacks four out of the six conserved cysteine residues and just the two adjacent residues Cys3 and Cys4 are retained; on the other hand, MgCRP-I 23 is an even more extreme case, as completely devoid of cysteines. Altogether, we propose to identify cases such as MgCRP-I 12 and 23 and the six non-coding pseudogenes we identified as MgCRP-I-like sequences, as they lack one of the two main distinctive features of the MgCRP-I family (a conserved signal peptide and the $C(X_{3-6})C(X_{1-7})CC(X_{3-4})C(X_{3-5})C$ array), but they retain significant similarity with known MgCRP-I sequences (detectable by HMMER or BLASTn with an e-value threshold of $1x10^{-5}$). These criteria will be important to identify further MgCRP-I-related loci once the mussel genome will be fully released.

CRP-I sequences are only found in the order Mytiloida

Following BLAST searches, the MgCRP-I peptides did not show significant sequence similarity with any other sequence deposited in public databases and the prediction of their tridimensional structure was considered unreliable by Phyre 2, due to the absence of models sharing sufficient sequence affinity within the PDB database. For this reason, we investigated the presence of MgCRP-I-like sequences in genomic and transcriptomic datasets which are increasingly available also for bivalve molluscs (Suárez-Ulloa, et al. 2013b).

Looking for short secreted peptides with a C-C-CC-C-C motif in the transcriptomes of 71 different species (see Materials and methods **and supplementary table S1, supplementary material online**), we could find this signature only in the mussels *M. edulis*, *M. trossulus*, and *M. californianus*. The first two species and *M. galloprovincialis* are widespread and genetically close to each other, as evidenced by the presence of natural hybrid populations (Beaumont, et al. 2004) whereas *M. californianus* is distributed in the Pacific coast of North America and is more distantly related to the other mussel ecotypes (Hilbish, et al. 2000). The full length CRP-I sequences identified, here named McCRP-I, MeCRP-I or MtCRP-I on the basis of the species name, are reported in **supplementary material online**.

No CRP-I-like sequence was detected in any of the other bivalve transcriptomes and we hypothesize their complete absence or no detectable expression in the analyzed bivalve species. Due to the high depth of NGS technologies, the latter hypothesis is unlikely and the genomic analysis of the oysters *C. gigas* and *P. fucata* strongly supports the absence of CRP-I-like genes, thus ruling out the possibility of a missed detection of poorly expressed transcripts. In addition, no evidence of CRP-I-like peptides was found in *Limnoperna fortunei*, *Perna viridis* and *Bathymodiolus platifrons*, the only three species among the over 50 different genera of the order Mytiloida, beside *Mytilus* spp., which have been subjected to Illumina RNA-sequencing so far.

Based on the available data, CRP-Is are certainly present in *Mytilus* spp. and appear to be absent in other mytiloids. Since the C-C-CC-C-C array is not common in bivalves and nothing

similar was found in the fully sequenced genomes of *C. gigas* (order Ostreoida) and *P. fucata* (order Pterioida), the CRP-I-like sequences, appear to have a narrow taxonomical distribution, comparable to that of other mussel cysteine-rich peptides (i.e. mytilins, myticins and mytimycins which cannot be found outside Mytiloida) and can be therefore considered as a taxonomically restricted gene (TRG) family (Khalturin, et al. 2009).

Occurrence of the C-C-CC-C-C array in protein databases and relationship with CRP-I peptides

Large-scale bioinformatic analyses revealed the presence of a CRP-I-like cysteine pattern mostly in animals and, among them, almost exclusively in invertebrates (see **Table 3**). More in detail, within the UniProtKB/Swiss-Prot database we found 452 peptides (279 if considering non-redundant peptides based on a 95% sequence identity criterion) mostly belonging to invertebrate animals: cone snails, turrids and terebrids (grouped as Conoidea in **Figure 5**), spiders, scorpions, pancrustaceans (with a single horseshoe crab and 35 insect sequences) and nematodes. Almost the totality of these peptides display a neurotoxic activity due to their high affinity to ion channels, like in the case of conotoxins, turritoxins and augertoxins (Aguilar, et al. 2009; Imperial, et al. 2003; Terlau and Olivera 2004) and peptides produced in the venom gland of spiders (Zhang, et al. 2010) and scorpions (Ma, et al. 2009). With the exception of a wasp toxin, all the entries from insects were related to bombyxin, a prothoracicotropic hormone involved in morphogenesis (Nijhout and Grunert 2002).

No CRP-I-like cysteine pattern was found in chordates, with the exception of 6 peptides (5 intestinal trefoil factors, with unclear function, and veswaprin, a snake antimicrobial peptide). A limited number of such peptides were found in Fungi, with the positive hits corresponding to uncharacterized proteins, and in green plants, with all cases representing antimicrobial peptides. The CRP-I-like array was also found in Baculoviruses (in a viral family of conotoxin-like peptides)

(Eldridge, et al. 1992). The full list of these peptides with their taxonomic origin and reported function are shown in **supplementary table 2, supplementary material online**.

Given the great sequence diversity, likely dependent on a fast evolutionary rate of molecular substitutions, the relationships among the MgCRP-I peptides (**Figure 3)** and with other CRPs (**Figure 5**) remain unresolved. Nevertheless, several features could be underlined by the phylogenetic analyses. All MgCRP-I sequences and the orthologous sequences from *M. edulis*, *M. trossulus* and *M. californianus* clustered together in a single clade, highlighting that the conservation of the signal peptide is a relevant criteria for the identification of CRP-I protein precursors. Most of the other CRP sequences clustered in distinct groups which included peptides from the same taxon (Pancrustacea, Conoidea, Spiders, Scorpions, Nematodes, Chordates, Plants, Fungi and Viruses), even if: *(i)* peptides from a single taxon were found in different clusters (e.g. conopeptides typically cluster in different groups that correspond to different superfamilies (Kaas, et al. 2010; Puillandre, et al. 2012)); *(ii)* some peptides, characterized by a long branch in the tree, clustered in a group mostly composed by peptides from another taxon (as an example, two conopeptides clustered within the group of mussel CRP-I peptides, with long branches, suggesting that they may belong to completely different structural classes).

Overall, no distinct traits can unequivocally link the evolutionary history of CRP-I peptides with those of other protein families characterized by the same cysteine array. As CRP-I-like peptides are absent in bivalves other than mussels, the only other molluscan group where the C-C-CC-C-C array is present is represented by Conoidea (Gastropoda). Nevertheless the absence of this molecular motif in the fully sequenced genomes of other gastropods (*Aplysia californica*, *Lottia gigantea* and *Biomphalaria glabrata*) suggests that it might have been acquired independently in these two molluscan groups. Even though the large sequence divergence prevents definitive conclusions, the study of the disulfide bonds topology in the MgCRP-I synthetic peptides provided further support to this hypothesis, as reported in the next section.

Oxidative refolding and disulfide bond topology of synthetic MgCRP-I peptides

Optimization of the oxidative folding yield for peptides with disulfide bridges is still an empiric exercise. Various parameters affect folding yields such as temperature, additives, redox couples, peptide concentration and duration of the folding reaction (Bulaj 2005; Bulaj and Olivera 2008). We synthetized the MgCRP-I 7 and MgCRP 9 peptides and their purified fractions were subjected to oxidative folding reactions in the presence of redox reagents GSH and GSSG (10:1) at 4°C buffered at pH 8 (RefoldA, see Materials and methods). This protocol, with slight modifications, has been commonly used in the refolding of disulfide rich proteins and, for example, it yielded good amounts (~90%) of the cystine knot peptide Huwentoxin-IV in its native structure (Deng, et al. 2013) and also in the synthesis of ω-conotoxin MVIIC (~50%). We tested also a high salt refolding mixture (RefoldB) which had shown improved yields of the same conotoxin (Kubo, et al. 1996).

In our hands each peptide refolding mixture (RefoldA) produced one major component, with purity ~85% and ~75% for MgCRP-I 7 and MgCRP-I 9, respectively, as determined by RP-HPLC peak area integration. Refold B gave similar results but with slightly lower yields (data not shown). According to ESI-MS, the molecular mass of folded MgCRP-I 7 and folded MgCRP-I 9 were in good agreement with those of the fully oxidized products (see **supplementary figure S3 and S4, supplementary material online**): 3,160.0 (calc. mono. 3,160.2) and 3,090.1 (calc. mono. 3,090.2), respectively. The evidence of one dominant product in the refolding of both peptides is consistent with the assumption of native conformation but, on the other hand, no comparison is currently possible between the synthetic products and the native counterparts eventually present in *Mytilus*, in particular due to the very low expression of MgCRP-I gene products in all tissues in physiological conditions (see section below).

We performed a disulfide connectivity prediction using the DiANNA (DiAminoacid Neural Network Application) Web Server of the Boston college (Ferrè and Clote 2005); the algorithm predicted a 1-2, 3-4, 5-6 topology for MgCRP-I 9 and a 1-4, 2-6, 3-5 topology for MgCRP-I 7 but

with a very high score (0.76 on a maximum of 1) in favor of a Cys14-Cys15 disulfide. At this point we experimentally determined the disulfide bond geometry using enzymatic fragmentation (trypsin and chymotrypsin) and LC/MS/MS. The analysis of both peptides revealed that the cysteine connectivity follows the nearest-neighbor pattern (1-2, 3-4, 5-6), namely Cys3-Cys8, Cys14-Cys15 and Cys20-Cys25 (see **supplementary tables S3 and S4, supplementary material online**). In a published detailed disulfide classification based on SwissProt and PFam databases the topology 1-2, 3-4, 5-6 is largely represented and contains a very heterogeneous ensemble of protein families (Gupta, et al. 2004); notably, the vicinal disulfide bond present in our MgCRP-I peptides and formed between the side chains of adjacent cysteines (Cys14-Cys15) represents a rare structural element. The vicinal disulfide, due to its intrinsic constrained nature, is usually described to be accompanied by the formation of a tight turn of the protein backbone (Carugo, et al. 2003); additionally, the oxidized and reduced states of this bond present very different structural features suggesting a possible role as conformational switch (Carugo, et al. 2003). At the present time, we do not know the significance that this vicinal bond could have on the activity of the MgCRP-I peptides but the observed 1-2, 3-4, 5-6 topology is distinctively different from the 1-4, 2-5, 3-6 topology common to knottins, which comprise conopeptides and most of the other peptides represented in **Figure 5** with an experimentally determined tridimensional structure (Hartig, et al. 2005). Further studies will be aimed in the future at the purification of native peptides to confirm the experimental results obtained concerning the folding of MgCRP-I 7 and 9 synthetic peptides.

MgCRP-I transcript levels

The number of sequences related to the MgCRP-I family identified in the many mussel transcriptome datasets analyzed was extremely low (**Table 4**) and suggests a very limited basal expression of these genes in different tissues under physiological conditions. More in detail, no evidence of MgCRP-Is was found in Sanger sequencing-based EST collections, with the exception of a single *M. edulis* sequence detected in a SSH library (digestive gland of mussels exposed to

styrene). The number of CRP-I sequences detected in the pyrosequencing-based datasets increased, even though in many cases MgCRP-I transcripts could not be detected. Finally, the analysis of Illumina sequencing-based transcriptomes clearly pointed out the high sequencing depth necessary to detect MgCRP-I messenger RNAs, which can be estimated to cumulatively contribute to less than 0.01% (but often to even less than 0.001%) of the total gene expression in most tissues.

To better evaluate the MgCRP-I tissue-specificity, we analyzed the expression levels of 17 MgCRP-I transcripts by quantitative PCR (qPCR) in different tissues (hemolymph, digestive gland, inner mantle, mantle rim, foot, posterior adductor muscle and gills) of a pool of 30 naïve adult mussels (*M. galloprovincialis*). We found large variability among the expression profiles of individual CRP-I sequences, with the overall expression levels being almost invariably very low in all tissues. However, three tissues emerged as main sites of MgCRP-I expression, namely the digestive gland, the inner mantle and the mantle rim. Several MgCRP-I displayed, at least to some extent, a certain degree of tissue specificity (**Figure 6**). In most cases these genes were not expressed at all in hemolymph, foot, gills and posterior adductor muscle, indicating that these are not the primary sites of production of MgCRP-I peptides, which is consistent with RNA-seq data (**Table 4**).

Overall, the gene expression data leave room to different hypotheses which need to be tested in future experiments: *(i)* the expression of these peptides may be induced by still unknown specific stimuli; *(ii)* MgCRP-I are expressed by a low number of highly specialized cells and therefore the global contribution to mRNAs extracted from macro-tissues is low; *(iii)* MgCRP-I are not expressed in adult individuals, but they play an important role in the early developmental stages (but this hypothesis seems to be disproved by the analysis of *M. edulis* larvae RNA-seq data, see **Table 4**).

MgCRP-I synthetic peptides do not show any significant cytotoxic, insecticidal, antifungal and antimicrobial activity

In an attempt to characterize the biological activity of the synthetic peptides MgCRP-I 7 and MgCRP-I 9, we evaluated their cytotoxicity on human tumor cell lines and insect larvae, and their antimicrobial activity on the bacteria *E. coli* and *S. aureus*, and on fungal strains of *C. albicans*, *C. neoformans*, *A. fumigatus* and *A. brasiliensis* (see Materials and methods). The MTT assays indicated that both synthetic peptides were not cytotoxic on the HT-29, SHSY5Y and MDAMB231 cell lines up to 10 μM concentration. Similarly, no insecticidal effect was observed in *Z. morio* larvae 48 hours after the injection of 300 μg peptide/Kg body weight, a quantity much higher than those determining visible neurotoxic effects, or even death, for other invertebrate toxins (Yang, et al. 2012; Zhong, et al. 2014). Finally, the antimicrobial activity assay evidenced that both MgCRP-I 7 and 9 did not show any effect on the selected bacterial and fungal strains at concentrations up to 32 μM.

Although these results indicate that MgCRP-I synthetic peptides did not display antimicrobial or cytotoxic activity in the tested conditions, their involvement in defense processes cannot be ruled out. In fact, post-translational modifications might occur in mussel cells but this can hardly be investigated due to the low expression levels of MgCRP-I genes which, in turn, makes the purification of native peptides difficult. Hence, the absence of biological effects could depend on a variety of modifications not present in the synthetic MgCRP-Is but often reported as fundamental for the antimicrobial or toxic activity of short cysteine-rich peptides (Bergeron, et al. 2013; Buczek, et al. 2005a; Buczek, et al. 2005b; Guder, et al. 2000). In addition, as we have previously stated, given the difficulty in purifying peptides expressed at such low levels from tissue extracts, we cannot certify that the folding observed for synthetic peptides is identical to that of native peptides, even though the fact that one dominant product was obtained in the refolding of both peptides is consistent with this assumption.

These considerations are important in perspective and, although the characterization of the activity of native MgCRP-I peptides is beyond the scope of this paper, this will be an important task to be accomplished in future studies.

Conclusions

Thanks to an exploratory bioinformatics approach applied to the NGS sequencing data, we could identify a novel family of cysteine rich peptides, named MgCRP-I, which appears to be exclusively present in Mytiloida, an order of marine filter-feeding mussels. The MgCRP-I gene family and the encoded peptides share a number of structural and evolutionary traits in common with other families of CRPs, which almost invariably have an antimicrobial or toxic function. These marked similarities initially suggested that MgCRP-I peptides could have similar biological functions, thus making them intriguing targets for possible future biotechnological and pharmacological applications. However, all the tests performed on two synthetic MgCRP-I peptides led to inconclusive results, leaving their biological role still puzzling. In addition, the biological targets (both at the molecular and at the species level) of MgCRP-I peptides are still unknown and the events triggering the expression of these molecules are still elusive. In absence of further indications, these questions remain unsolved.

Overall, we have provided a preliminary overview on MgCRP-I peptides, which is intended as starting point for further investigations on their possible action on prokaryotic or eukaryotic cells. Our work also highlights the possibility of identifying previously uncharacterized, potentially bioactive, peptides from whole genomes and transcriptomes of non-model organisms without any previous knowledge about their primary sequence, an experimental approach which could speed up the discovery or the design of novel molecules with potential biotechnological applications. Due to their still limited genomic knowledge, marine invertebrates in particular represent a virtually unlimited and almost unexplored source of novel bioactive compounds.

Acknowledgements

This work was supported by BIVALIFE (FP7-KBBE-2010-4) and PRIN 2010-11 (20109XZEPR).

Figure legends

Figure 1: Sequence variability of MgCRP-I sequences; variability index W is plotted in the upper panel, while the sequence consensus, obtained with Weblogo (http://weblogo.berkeley.edu) is shown in the lower panel. Only positions covered by at least 50% sequences in the global alignment of MgCRP-I peptides are shown. Sites under positive selection are indicated by an asterisk.

Figure 2. Exon/intron structure of the complete coding regions of the MgCRP-I 13, 14, 28, 45 and multi-MgCRP-I 2 genes (panel A) and corresponding organization of the encoded peptide precursors (panel B). The positions of the signal peptide, pro-region and mature peptide regions are highlighted and each cysteine-rich module is marked by a box.

Figure 3: Maximum likelihood tree obtained with the MgCRP-I peptides based on the alignment of the signal peptide region only. Only Booststraps values > 75 are shown. Some sequences were not considered in this analysis as their N-terminal region was incomplete (see **Table 2**). Arrows indicate two *MgCRP-I-like* peptides with a disrupted cysteine array (MgCRP-I 12 and 23), marking an unconventional mature region.

Figure 4: codon usage for the Arg residue responsible of the pro-peptide cleavage site and for the 6 cysteine residues engaged in disulfide bridges, calculated on *M. galloprovincialis* MgCRP-I peptides. The probabilities of finding the observed codon biases were calculated assuming a binomial distribution and the codon usage (a priori probabilities) inferred from the transcriptome

published by Gerdol et al. 2014 (75.1%-24.9% for TGT-TGC, encoding Cys, 49.3%-16.3%-13.3%-12.5%-4.7%-3.8% for AGA-AGG-CGA-CGT-CGG-CGC, encoding Arg). Significant ($p < 0.05$) and highly significant ($p < 0.01$) deviations from the expected distributions ($p < 0.01$) are marked by * and ** respectively. NS = not significant.

Figure 5: Maximum likelihood tree obtained with the signal peptide of MgCRP-I, the orthologous sequences from other mussel species and all the CRPs mined from UniProtKB/Swiss-Prot (see Materials and methods). Peptides from pancrustaceans, conoideans, spiders, scorpions, nematodes, chordates, plants, fungi, viruses and mussels are shown. Mussel CRP-I peptides are highlighted in a grey background.

Figure 6: gene expression of 17 selected MgCRP-I genes in six tissues (HE: hemolymph; DG: digestive gland; IM: inner mantle; MR: mantle rim; FO: foot; GI: gills; AM: posterior adductor muscle); primers for MgCRP-I 3 also target MgCRP-I 25, primers for MgCRP-I 10 also target MgCRP-I 26. Bars represent the expression relative to EF-1 alpha; results are mean ± standard deviation of three replicates. MgCRP-I sequences are divided into three panels based on their expression level: panel A – genes with maximum relative expression value comprised between 0.05 and 0.25; panel B – genes with maximum relative expression value comprised between 0.004 and 0.01; panel C - with maximum relative expression value lower than 0.003. Panel D shows a schematic representation of a *M. galloprovincialis* anatomical features, highlighting the sampled tissues.

References

Adams DJ, Alewood PF, Craik DJ, Drinkwater RD, Lewis RJ 1999. Conotoxins and their potential pharmaceutical applications. Drug Develop Res. 46: 219-234. doi: http://dx.doi.org/10.1002/(SICI)1098-2299(199903/04)46:3/4<219::AID-DDR7>3.0.CO;2-S

Aguilar MB, de la Rosa RAC, Falcón A, Olivera BM, Heimer de la Cotera EP 2009. Peptide pal9a from the venom of the turrid snail *Polystira albida* from the Gulf of Mexico: Purification, characterization, and comparison with P-conotoxin-like (framework IX) conoidean peptides. Peptides. 30: 467-476. doi: http://dx.doi.org/10.1016/j.peptides.2008.09.016

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic local alignment search tool. J Mol Biol. 215: 403-410. doi: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Bartlett TC, et al. 2002. Crustins, homologues of an 11.5-kDa antibacterial peptide, from two species of penaeid shrimp, *Litopenaeus vannamei* and *Litopenaeus setiferus*. Mar Biotechnol. 4: 278-293. doi: http://dx.doi.org/10.1007/s10126-002-0020-2

Bassim S, Genard B, Gauthier-Clerc S, Moraga D, Tremblay R 2014. Ontogeny of bivalve immunity: assessing the potential of next-generation sequencing techniques. Reviews in Aquaculture. 6: 1-21. doi: http://dx.doi.org/10.1111/raq.12064

Beaumont AR, Turner G, Wood AR, Skibinski DOF 2004. Hybridisations between *Mytilus edulis* and *Mytilus galloprovincialis* and performance of pure species and hybrid veliger larvae at different temperatures. J Exp Mar Biol Ecol. 302: 177-188. doi: http://dx.doi.org/10.1016%2Fj.jembe.2003.10.009

Benincasa M, et al. 2010. Antifungal activity of Amphotericin B conjugated to carbon nanotubes. ACS Nano. 5: 199-208. doi: http://dx.doi.org/10.1021/nn1023522

Benincasa M, et al. 2004. Antimicrobial activity of Bac7 fragments against drug-resistant clinical isolates. Peptides. 25: 2055-2061. doi: http://dx.doi.org/10.1016/j.peptides.2004.08.004

Bergeron ZL, et al. 2013. A 'conovenomic' analysis of the milked venom from the mollusk-hunting cone snail *Conus textile*—The pharmacological importance of post-translational modifications. Peptides. 49: 145-158. doi: http://dx.doi.org/10.1016/j.peptides.2013.09.004

Buczek O, Bulaj G, Olivera BM 2005a. Conotoxins and the posttranslational modification of secreted gene products. Cell Mol Life Sci. 62: 3067-3079. doi: http://dx.doi.org/10.1007/s00018-005-5283-0

Buczek O, Yoshikami D, Bulaj G, Jimenez EC, Olivera BM 2005b. Post-translational amino acid isomerization: a functionally important d-amino acid in an excitatory peptide. J Biol Chem. 280: 4247-4253. doi: http://dx.doi.org/10.1074/jbc.M405835200

Bulaj G 2005. Formation of disulfide bonds in proteins and peptides. Biotechnology Advances 23: 87-92. doi: http://dx.doi.org/10.1016/j.biotechadv.2004.09.002

Bulaj G, Olivera BM 2008. Folding of conotoxins: Formation of the native disulfide bridges during chemical synthesis and biosynthesis of *Conus* peptides. Antioxid Redox Sign. 10: 141-155. doi: http://dx.doi.org/10.1089/ars.2007.1856

Bulet P, Stöcklin R 2005. Insect antimicrobial peptides: Structures, properties and gene regulation. Protein Peptide Lett. 12: 3-11. doi: http://dx.doi.org/10.2174/0929866053406011

Carugo O, et al. 2003. Vicinal disulfide turns. Protein Eng.16: 637-639. doi: http://dx.doi.org/10.1093/protein/gzg088

Casteels-Josson K, Capaci T, Casteels P, Tempst P 1993. Apidaecin multipeptide precursor structure: a putative mechanism for amplification of the insect antibacterial response. EMBO J. 12: 1569-1578.

Charlet M, et al. 1996. Innate immunity: Isolation of several cysteine-rich antimicrobial peptides from the blood of a mollusc, *Mytilus edulis*. J Biol Chem. 271: 21808-21813. doi: http://dx.doi.org/10.1074/jbc.271.36.21808

Conticello SG, et al. 2001. Mechanisms for evolving hypervariability: The case of conopeptides. Mol Biol Evol. 18: 120-131.

Craft JA, et al. 2010. Pyrosequencing of *Mytilus galloprovincialis* cDNAs: Tissue-specific expression patterns. PLoS ONE. 5. doi: http://dx.doi.org/10.1371/journal.pone.0008875

Craig AG, Bandyopadhyay P, Olivera BM 1999. Post-translationally modified neuropeptides from *Conus* venoms. Eur J Biochem. 264: 271-275. doi: http://dx.doi.org/10.1046/j.1432-1327.1999.00624.x

d'Alençon E, et al. 2013. Evolutionary history of x-tox genes in three lepidopteran species: Origin, evolution of primary and secondary structure and alternative splicing, generating a repertoire of immune-related proteins. Insect Biochem Molec. 43: 54-64. doi: http://dx.doi.org/10.1016/j.ibmb.2012.10.012

Deng M, et al. 2013. Synthesis and biological characterization of synthetic analogs of Huwentoxin-IV (Mu-theraphotoxin-Hh2a), a neuronal tetrodotoxin-sensitive sodium channel inhibitor. Toxicon. 71: 57-65. doi: http://dx.doi.org/10.1016/j.toxicon.2013.05.015

Destoumieux-Garzón D, et al. 2009. Spodoptera frugiperda X-Tox protein, an immune related defensin rosary, has lost the function of ancestral defensins. PLoS ONE 4: e6795. doi: http://dx.doi.org/10.1371/journal.pone.0006795

Destoumieux D, et al. 1997. Penaeidins, a new family of antimicrobial peptides isolated from the shrimp *Penaeus vannamei* (Decapoda). J Biol Chem. 272: 28398-28406. doi: http://dx.doi.org/10.1074/jbc.272.45.28398

Dinkel H, et al. 2011. ELM—the database of eukaryotic linear motifs. Nucleic Acids Res. 40: D242-51. doi: http://dx.doi.org/10.1093/nar/gkr1064

Duckert P, Brunak S, Blom N 2004. Prediction of proprotein convertase cleavage sites. Protein Eng Des Sel. 17: 107-112. doi: http://dx.doi.org/10.1093/protein/gzh013

Eddy SR 2011. Accelerated profile HMM searches. PLoS Comput Biol. 7: e1002195. doi: http://dx.doi.org/10.1371/journal.pcbi.1002195

Edgar RC 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792-1797. doi: http://dx.doi.org/10.1093/nar/gkh340

Ehret-Sabatier L, et al. 1996. Characterization of novel cysteine-rich antimicrobial peptides from scorpion blood. J Biol Chem. 271: 29537-29544. doi: http://dx.doi.org/10.1074/jbc.271.47.29537

Eldridge R, Li Y, Miller LK 1992. Characterization of a baculovirus gene encoding a small conotoxinlike polypeptide. J Virol. 66: 6563-6571.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2: 953-971. doi: http://dx.doi.org/10.1038/nprot.2007.131

Fan C-X, et al. 2003. A novel conotoxin from *Conus betulinus*, κ-BtX, unique in cysteine pattern and in function as a specific BK channel modulator. J Biol Chem. 278: 12624-12633. doi: http://dx.doi.org/10.1074/jbc.M210200200

Ferrè F, Clote P 2005. DiANNA: A web server for disulfide connectivity prediction. Nucleic Acids Res.33: W230-W232. doi: http://dx.doi.org/10.1093/nar/gki412

Fogaça AC, et al. 2004. Cysteine-rich antimicrobial peptides of the cattle tick *Boophilus microplus*: Isolation, structural characterization and tissue expression profile. Dev Comp Immunol. 28: 191-200. doi: http://dx.doi.org/10.1016/j.dci.2003.08.001

Freer A, Bridgett S, Jiang J, Cusack M 2014. Biomineral proteins from *Mytilus edulis* mantle tissue transcriptome. Mar Biotechnol. 16: 34-45. doi: http://dx.doi.org/10.1007/s10126-013-9516-1

Froy O, Gurevitz M 2004. Arthropod defensins illuminate the divergence of scorpion neurotoxins. J Pept Sci. 10: 714-718. doi: http://dx.doi.org/10.1002/psc.578

Gerdol M, 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 Research Notes 7: 722. doi: http://dx.doi.org/10.1186/1756-0500-7-722

Gerdol M, De Moro G, Manfrin C, Venier P, Pallavicini A 2012. Big defensins and mytimacins, new AMP families of the Mediterranean mussel *Mytilus galloprovincialis*. Dev Comp Immunol. 36: 390-399. doi: http://dx.doi.org/10.1016/j.dci.2011.08.003

Gerdol M, 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: http://dx.doi.org/10.1016/j.dci.2011.01.018

Gerdol M, Venier P An updated molecular basis for mussel immunity. Fish Shellfish Immun (in press). doi: http://dx.doi.org/10.1016/j.fsi.2015.02.013

González VL, et al. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Procl Bio Sci (in press). doi: http://dx.doi.org/10.1098/rspb.2014.2332

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652. doi: http://dx.doi.org/10.1038/nbt.1883

Gruber CW, Čemažar M, Anderson MA, Craik DJ 2007. Insecticidal plant cyclotides and related cystine knot toxins. Toxicon. 49: 561-575. doi: http://dx.doi.org/10.1016/j.toxicon.2006.11.018

Guder A, Wiedemann I, Sahl HG 2000. Posttranslationally modified bacteriocins - The lantibiotics. Biopolymers - Peptide Science Section. 55: 62-73. doi: http://dx.doi.org/10.1002/1097-0282(2000)55:1%3C62::AID-BIP60%3E3.0.CO;2-Y

Gupta A, Van Vlijmen HWT, Singh J 2004. A classification of disulfide patterns and its relationship to protein structure and function. Protein Sci. 13: 2045-2058. doi: http://dx.doi.org/10.1110/ps.04613004

Hartig GRS, Tran TT, Smythe ML 2005. Intramolecular disulphide bond arrangements in nonhomologous proteins. Protein Sci. 14: 474-482. doi: http://dx.doi.org/10.1110/ps.04923305

Hilbish TJ, et al. 2000. Origin of the antitropical distribution pattern in marine mussels (*Mytilus* spp.): routes and timing of transequatorial migration. Mar Biol. 136: 69-77. doi: http://dx.doi.org/10.1007/s002270050010

Hubert F, Noël T, Roch P 1996. A member of the arthropod defensin family from edible mediterranean mussels (*Mytilus galloprovincialis*). Eur J Biochem. 240: 302-306. doi: http://dx.doi.org/10.1111/j.1432-1033.1996.0302h.x

Imperial JS, et al. 2003. The augertoxins: Biochemical characterization of venom components from the toxoglossate gastropod *Terebra subulata*. Toxicon. 42: 391-398. doi: http://dx.doi.org/10.1016/S0041-0101(03)00169-7

Kaas Q, Westermann JC, Craik DJ 2010. Conopeptide characterization and classifications: An analysis using ConoServer. Toxicon. 55: 1491-1509. doi: http://dx.doi.org/10.1016/j.toxicon.2010.03.002

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

Kelley LA, Sternberg MJ 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 4: 363-371. doi: http://dx.doi.org/10.1038/nprot.2009.2

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25: 404-413. doi: http://dx.doi.org/10.1016/j.tig.2009.07.006

Kubo S, Chino N, Kimura T, Sakakibara S 1996. Oxidative folding of ω-conotoxin MVIIC: Effects of temperature and salt. Biopolymers. 38: 733-744. doi: http://dx.doi.org/10.1002/(SICI)1097-0282(199606)38:6<733::AID-BIP5>3.0.CO;2-S

Li H, Parisi MG, Parrinello N, Cammarata M, Roch P 2011. Molluscan antimicrobial peptides, a review from activity-based evidences to computer-assisted sequences. Invert Surviv J. 8: 85-97.

Liao Z, et al. 2013. Molecular characterization of a novel antimicrobial peptide from *Mytilus coruscus*. Fish Shellfish Immun. 34: 610-616. doi: http://dx.doi.org/10.1016/j.fsi.2012.11.030

Ma Y, et al. 2009. Transcriptome analysis of the venom gland of the scorpion *Scorpiops jendeki*: Implication for the evolution of the scorpion venom arsenal. BMC Genomics. 10: 290. doi: http://dx.doi.org/10.1186/1471-2164-10-290

Marshall E, Costa LM, Gutierrez-Marcos J 2011. Cysteine-Rich Peptides (CRPs) mediate diverse aspects of cell-cell communication in plant reproduction and development. J ExpBot. 62: 1677-1686. doi: http://dx.doi.org/10.1093/jxb/err002

Mayer AMS, Rodríguez AD, Berlinck RGS, Fusetani N 2011. Marine pharmacology in 2007-8: Marine compounds with antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiprotozoal, antituberculosis, and antiviral activities; Affecting the immune and nervous system, and other miscellaneous mechanisms of action. Comp Biochem Phys C. 153: 191-222. doi: http://dx.doi.org/10.3390%2Fmd11072510

Mitta G, Hubert F, Noël T, Roch P 1999. Myticin, a novel cysteine-rich antimicrobial peptide isolated from haemocytes and plasma of the mussel *Mytilus galloprovincialis*. Eur J Biochem. 265: 71-78. doi: http://dx.doi.org/10.1046/j.1432-1327.1999.00654.x

Mitta G, Vandenbulcke F, Hubert F, Salzet M, Roch P 2000a. Involvement of mytilins in mussel antimicrobial defense. J Biol Chem. 275: 12954-12962. doi: http://dx.doi.org/10.1074/jbc.275.17.12954

Mitta G, et al. 2000b. Differential distribution and defence involvement of antimicrobial peptides in mussel. J Cell Sci. 113: 2759-2769.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B 2011. How Many Species Are There on Earth and in the Ocean? PLoS Biol. 9: e1001127. doi: http://dx.doi.org/10.1371/journal.pbio.1001127

Nguyen TTT, Hayes BJ, Ingram BA 2014. Genetic parameters and response to selection in blue mussel (*Mytilus galloprovincialis*) using a SNP-based pedigree. Aquaculture. 420–421: 295-301. doi: http://dx.doi.org/10.1016/j.aquaculture.2013.11.021

Nijhout HF, Grunert LW 2002. Bombyxin is a growth factor for wing imaginal disks in lepidoptera. Proc Natl Acad Sci USA. 99: 15446-15450. doi: http://dx.doi.org/10.1073/pnas.242548399

Olivera BM, et al. 2012. Adaptive radiation of venomous marine snail lineages and the accelerated evolution of venom peptide genes. Ann NY Acad Sci. 1267: 61-70. doi: http://dx.doi.org/10.1111/j.1749-6632.2012.06603.x

Otero-González AJ, et al. 2010. Antimicrobial peptides from marine invertebrates as a new frontier for microbial infection control. FASEB J. 24: 1320-1334. doi: http://dx.doi.org/10.1096/fj.09-143388

Petersen TN, Brunak S, von Heijne G, Nielsen H 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8: 785-786. doi: http://dx.doi.org/10.1038/nmeth.1701

Philipp EER, et al. 2012. Massively parallel rna sequencing identifies a complex immune gene repertoire in the lophotrochozoan *Mytilus edulis*. PLoS ONE. 7: e33091. doi: http://dx.doi.org/10.1371/journal.pone.0033091

Puillandre N, Koua D, Favreau P, Olivera BM, Stöcklin R 2012. Molecular phylogeny, classification and evolution of conopeptides. J Mol Evol. 74: 297-309. doi: http://dx.doi.org/10.1007/s00239-012-9507-2

Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339: 62-66. doi: http://dx.doi.org/10.1016/S0304-3940(02)01423-4

Ranasinghe S, McManus DP 2013. Structure and function of invertebrate Kunitz serine protease inhibitors. Dev Comp Immunol. 39: 219-227. doi: http://dx.doi.org/10.1016/j.dci.2012.10.005

Ratzka C, et al. 2012. Molecular Characterization of antimicrobial peptide genes of the carpenter ant *Camponotus floridanus*. PLoS ONE 7: e43036. doi: http://dx.doi.org/10.1371/journal.pone.0043036

Rayaprolu S, Wang Y, Kanost MR, Hartson S, Jiang H 2010. Functional analysis of four processing products from multiple precursors encoded by a lebocin-related gene from *Manduca sexta*. Dev Comp Immunol. 34: 638-647. doi: http://dx.doi.org/10.1016/j.dci.2010.01.008

Reese MG, Eeckman FH, Kulp D, Haussler D 1997. Improved Splice Site Detection in Genie. J Comput Biol. 4: 311-323. doi: http://dx.doi.org/10.1089/cmb.1997.4.311

Rice P, Longden I, Bleasby A 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276-277. doi: http://dx.doi.org/10.1016/S0168-9525(00)02024-2

Rimphanitchayakit V, Tassanakajon A 2010. Structure and function of invertebrate Kazal-type serine proteinase inhibitors. Dev Comp Immunol. 34: 377-386. doi: http://dx.doi.org/10.1016/j.dci.2009.12.004

Rodríguez de la Vega RC, Possani LD 2005. On the evolution of invertebrate defensins. Trends Genet. 21: 330-332. doi: http://dx.doi.org/10.1016/j.tig.2005.03.009

Romiguier J, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature. 515: 261-263. doi: http://dx.doi.org/10.1038/nature13685

Rosani U, et al. 2011. Massively parallel amplicon sequencing reveals isotype-specific variability of antimicrobial peptide transcripts in *Mytilus galloprovincialis*. PLoS ONE. 6: e26680. doi: http://dx.doi.org/10.1371%2Fjournal.pone.0026680

Saez NJ, et al. 2010. Spider-venom peptides as therapeutics. Toxins. 2: 2851-2871. doi: http://dx.doi.org/10.3390/toxins2122851

Sperstad SV, et al. 2011. Antimicrobial peptides from marine invertebrates: Challenges and perspectives in marine antimicrobial peptide discovery. Biotechnol Adv. 29: 519-530. doi: http://dx.doi.org/10.1016/j.biotechadv.2011.05.021

Stamatakis A 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22: 2688-2690. doi: http://dx.doi.org/10.1093/bioinformatics/btl446

Steiner AM, Bulaj G, Puillandre N 2013. On the importance of oxidative folding in the evolution of conotoxins: cysteine codon preservation through gene duplication and adaptation. PLoS ONE. 8: e78456. doi: http://dx.doi.org/10.1371/journal.pone.0078456

Suárez-Ulloa V, et al. 2013a. The CHROMEVALOA database: A resource for the evaluation of Okadaic Acid contamination in the marine environment based on the chromatin-associated transcriptome of the mussel *Mytilus galloprovincialis*. Mar Drugs. 11: 830-841. doi: http://dx.doi.org/10.3390/md11030830

Suárez-Ulloa V, et al. 2013b. Bivalve omics: state of the art and potential applications for the biomonitoring of harmful marine compounds. Mar Drugs. 11: 4370-4389. doi: http://dx.doi.org/10.3390/md11114370

Takeuchi T, et al. 2012. Draft genome of the pearl oyster *Pinctada fucata*: A platform for understanding bivalve biology. DNA Res. 19: 117-130. doi: http://dx.doi.org/10.1093/dnares/dss005

Taylor K, Barran PE, Dorin JR 2008. Structure-activity relationships in β-defensin peptides. Biopolymers - Peptide Science Section. 90: 1-7. doi: http://dx.doi.org/10.1002/bip.20900

Tedford HW, Fletcher JI, King GF 2001. Functional significance of the β-Hairpin in the insecticidal neurotoxin ω-Atracotoxin-Hv1a. J Biol Chem. 276: 26568-26576. doi: http://dx.doi.org/10.1074/jbc.M102199200

Terlau H, Olivera BM 2004. Conus Venoms: A Rich Source of Novel Ion Channel-Targeted Peptides. Physiol Rev. 84: 41-68. doi: http://dx.doi.org/10.1152/physrev.00020.2003

Toubiana M, et al. 2013. Toll-like receptors and MyD88 adaptors in *Mytilus*: Complete cds and gene expression levels. Dev Comp Immunol. 40: 158-166. doi: http://dx.doi.org/10.1016/j.dci.2013.02.006

Toubiana M, et al. 2014. Toll signal transduction pathway in bivalves: Complete cds of intermediate elements and related gene transcription levels in hemocytes of immune stimulated *Mytilus galloprovincialis*. Dev Comp Immunol. 45: 300-312. doi: http://dx.doi.org/10.1016/j.dci.2014.03.021

Wang C-Z, Jiang H, Ou Z-L, Chen J-S, Chi C-W 2003. cDNA cloning of two A-superfamily conotoxins from *Conus striatus*. Toxicon. 42: 613-619. doi: http://dx.doi.org/10.1016/j.toxicon.2003.08.005

Xiong Y-M, Ling M-H, Wang D-C, Chi C-W 1997. The cDNA and genomic DNA sequences of a mammalian neurotoxin from the scorpion *Buthus martensii* Karsch. Toxicon. 35: 1025-1031. doi: http://dx.doi.org/10.1016/s0041-0101(96)00224-3

Xu B, Yang Z 2013. pamlX: A Graphical User Interface for PAML. Mol Biol Evol. doi: http://dx.doi.org/10.1093/molbev/mst179

Yang S, et al. 2012. Chemical punch packed in venoms makes centipedes excellent predators. Mol Cell Proteomics. 11: 640-650. doi: http://dx.doi.org/10.1074/mcp.M112.018853

Yang Z 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24: 1586-1591. doi: http://dx.doi.org/10.1093/molbev/msm088

Zhang G, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 490: 49-54. doi: http://dx.doi.org/10.1038/nature11413

Zhang Y, et al. 2010. Transcriptome analysis of the venom glands of the Chinese wolf spider *Lycosa singoriensis*. Zoology. 113: 10-18. doi: http://dx.doi.org/10.1016/j.zool.2009.04.001

Zhong Y, et al. 2014. A novel neurotoxin from venom of the spider, *Brachypelma albopilosum*. PLoS ONE. 9: e110221. doi: http://dx.doi.org/10.1371/journal.pone.0110221

Table 1: primers designed for assessing the tissue-specific expression levels of MgCRP-I genes by real-time PCR.

Primer name	Primer sequence
MgCRP-I 1 for	TGTGTGTTGTTGGTCGTCGT
MgCRP-I 1 rev	GTAACCGGAACGACAAAAGC
MgCRP-I 2 for	AGCCTCAAGTAAGAAGTAAAACAGA
MgCRP-I 2 rev	CAGCTTMTTCTACCGCATCC
MgCRP-I 3/25 for	GACAAAGTGAACTAAAGCATTTCA
MgCRP-I 3/25 rev	CTCCGTTTTCTCCAAAGCTG
MgCRP-I 4 for	CATGGCACATGAMGAAATGC
MgCRP-I 4 rev	TTAGCCACCATAGCGTTTGC
MgCRP-I 5 for	TGGATAAAAGGTGACCCACAG
MgCRP-I 5 rev	TCTTCCAGCATTTCGTCCTT
MgCRP-I 6 for	AAYATGGCGAAGGAAGACAT
MgCRP-I 6 rev	AAGTTCAGTCGCGCCTACAT
MgCRP-I 7 for	GTTGGAGTCAACATGGCAAA
MgCRP-I 7 rev	GCGCATGCATTTTCTGTAAG
MgCRP-I 8 for	GCATTTGCTTATAGTGTTGCAGA
MgCRP-I 8 rev	TKCAAATGATGGATGGCTAA
MgCRP-I 9 for	GCTTTTTGTTTGTTTGGTAGCC
MgCRP-I 9 rev	CGAACACATCTTCTGTATGAGCA
MgCRP-I 10/26 for	GGCACATGAAGAAATGTTCG
MgCRP-I 10/26 rev	CCTGCATACGCCAAAACAT
MgCRP-I 11 for	TAAACCCCTTGTTCGGTCAC
MgCRP-I 11 rev	AGTGTGACGGATGCAAACAA
MgCRP-I 14 for	AGCCTTCGTTGGAACTAGCA
MgCRP-I 14 rev	TCGAGCGAGATTGACATCTG
multi-MgCRP-I 1 for	CTGACGAAATGGTGGAGGAT
multi-MgCRP-I 1 rev	TACAGCATTGACGGCTGTTT
multi-MgCRP-I 2 for	GCAAACATGGCCAAAGAAGT
multi-MgCRP-I 2 rev	GTCACGGGTCTTTTTGCATT
multi-MgCRP-I 3 for	AAGAGCTCCTGCATGTGGAT
multi-MgCRP-I 3 rev	TCCTCCTCCCGTTCTCTTTT
EF-1 alpha for	CCTCCCACCATCAAGACCTA
EF-1 alpha rev	GGCTGGAGCAAGGTAACAA

Table 2: list of the MgCRP-I sequences identified in the present work. *complete sequence correspond to a full-length coding sequence, from the initial ATG to the STOP codon; **T: transcriptome; G: genome.

Sequence name	Status*	Evidence**	Genomic scaffold	Cysteine-rich domains
MytiCRP-I 1	complete	T	/	1
MytiCRP-I 2	complete	G, T	APJB011836849.1	1
MytiCRP-I 3	complete	T	/	1
MytiCRP-I 4	complete	G, T	APJB011405634.1	1
MytiCRP-I 5	complete	G, T	APJB010137175.1	1
MytiCRP-I 6	complete	T	/	1
MytiCRP-I 7	complete	T	/	1
MytiCRP-I 8	complete	G, T	APJB0118677191.1	1
MytiCRP-I 9	complete	G, T	APJB010390130.1	1
MytiCRP-I 10	complete	T	/	1
MytiCRP-I 11	incomplete	T	/	1
MytiCRP-I 12	complete	T	/	none
MytiCRP-I 13	complete	G, T	APJB010539225.1	1
MytiCRP-I 14	complete	G, T	APJB011981014.1	1
MytiCRP-I 15	complete	G, T	APJB012303283.1	1
MytiCRP-I 16	complete	T	/	1
MytiCRP-I 17	complete	T	/	1
MytiCRP-I 18	complete	G, T	APJB010096024.1	1
MytiCRP-I 19	complete	T	/	1
MytiCRP-I 20	incomplete	T	/	1
MytiCRP-I 21	complete	T	/	1
MytiCRP-I 22	complete	G, T	APJB010149223.1	1
MytiCRP-I 23	complete	T	/	none
MytiCRP-I 24	complete	G, T	APJB011420996.1	1
MytiCRP-I 25	complete	G, T	APJB011525896.1	1
MytiCRP-I 26	complete	G, T	APJB011451595.1	1
MytiCRP-I 27	incomplete	G	APJB010022937.1	1
MytiCRP-I 28	complete	G	APJB010019019.1	1
MytiCRP-I 29	incomplete	G, T	APJB010215939.1	1
MytiCRP-I 30	incomplete	G	APJB010309773.1	1
MytiCRP-I 31	complete	G, T	APJB010337167.1	1
MytiCRP-I 32	incomplete	G	APJB010405325.1	1
MytiCRP-I 33	incomplete	G	APJB010538560.1	1
MytiCRP-I 34	complete	G, T	APJB010602145.1	1
MytiCRP-I 35	incomplete	G	APJB010726714.1	1
MytiCRP-I 36	complete	G, T	APJB010858750.1	1
MytiCRP-I 37	incomplete	G	APJB011013544.1	1
MytiCRP-I 38	incomplete	G, T	APJB011377302.1	1
MytiCRP-I 39	incomplete	G, T	APJB011417411.1	1
MytiCRP-I 40	complete	G	APJB011602152.1	1

MytiCRP-I 41	incomplete	G	APJB01171896.1	1
MytiCRP-I 42	incomplete	G	APJB011833940.1	1
MytiCRP-I 43	incomplete	G	APJB011892489.1	1
MytiCRP-I 44	incomplete	G	APJB011902451.1	1
MytiCRP-I 45	complete	G, T	APJB012001676.1	1
MytiCRP-I 46	incomplete	G	APJB012002994.1	1
MytiCRP-I 47	incomplete	G	APJB012084462.1	1
MytiCRP-I 48	incomplete	G	APJB011591868.1	1
MytiCRP-I 49	incomplete	G	APJB011815456.1	1
MytiCRP-I 50	complete	T	/	1
MytiCRP-I 51	complete	T	/	1
MultiMytiCRP-I 1	complete	T	/	3
MultiMytiCRP-I 2	complete	G, T	APJB011508508.1	4
MultiMytiCRP-I 3	complete	G, T	APJB012209485.1	2
MultiMytiCRP-I 4	incomplete	T	/	2
MultiMytiCRP-I 5	complete	T	/	2
MultiMytiCRP-I 6	incomplete	G	APJB010388843.1	2 or more
MultiMytiCRP-I 7	incomplete	G	APJB010167718.1	2 or more
MultiMytiCRP-I 8	incomplete	G	APJB010303175.1	3 or more
MultiMytiCRP-I 9	incomplete	G	APJB010305277.1	4
MultiMytiCRP-I 10	incomplete	G	APJB010449694.1	4
MultiMytiCRP-I 11	complete	G	APJB010750334.1	2
MultiMytiCRP-I 12	complete	G	APJB011083975.1	2
MultiMytiCRP-I 13	complete	G	APJB011153262.1	2
MultiMytiCRP-I 14	complete	G, T	APJB011903515.1	4
MultiMytiCRP-I 15	incomplete	G	APJB011965594.1	2
MultiMytiCRP-I 16	complete	T	/	2

Table 3: Number of peptides with a MgCRP-I-like cysteine array found in the UniProtKB/Swiss-Prot protein sequence database, listed per taxonomic group (function is indicated if available).

Taxonomic group	Number of identified CRPs in UniprotKB*	Molecular function
Cone snails	163	conotoxins
Spiders	223	venom toxins
Fungi	2	uncharacterized
Viruses (Buculoviridae)	4	uncharaxterized
Insects	35	hormones/venom toxins
Green Plants	5	antimicrobial peptides
Scorpions	8	venom toxins
Terebrids	2	augertoxins
Horseshoe crabs	1	antimicrobial peptides
Nematodes	1	Insulin-like
Vertebrates	6	unnknown/antimicrobial peptides
Turrids	2	turritoxins

*Non-redundant positive matches based on threshold criteria of a 95% sequence identity.

Table 4: Number of CRP-I sequencing reads identified in the publicly available transcriptome datasets from *Mytilus* spp. (retrieved from NCBI SRA, February 2015).

Database	REF	tissue	sequencing strategy	total number of sequences	sequences related to MgCRP-I	%
M. galloprovincialis	(Venier, et al. 2009)	mixed tissues	Sanger	19,617	0	0
M. californianus	NA	mixed tissues	Sanger	42,354	0	0
M. coruscus	NA	foot	Sanger	719	0	0
M. galloprovincialis	(Craft, et al. 2010)	foot	454	31,227	0	0
M. galloprovincialis	(Craft, et al. 2010)	mantle	454	52,057	0	0
M. galloprovincialis	(Suárez-Ulloa, et al. 2013a)	digestive gland	454	2,206,478	0	0
M. trossulus	(Romiguier, et al. 2014)	mixed tissues	Illumina	~58 million	142	<0.001
M. galloprovincialis	NA	hemocytes	Illumina	~106 million	490	<0.001
M. galloprovincialis	NA	gills	Illumina	~52 million	182	<0.001
M. edulis	(Philipp, et al. 2012)	hemocytes	454	407,061	2	<0.001
M. edulis	(Bassim, et al. 2014)	larvae	Illumina	~295 million	3,423	0.001
M. californianus	(Romiguier, et al. 2014)	mixed tissues	Illumina	~78 million	644	0.001
M. edulis	(Philipp, et al. 2012)	mixed tissues	454	365,626	3	0.001
M. edulis	(Freer, et al. 2014)	mantle	454	494,391	8	0.002
M. galloprovincialis	(Craft, et al. 2010)	gill	454	58,271	1	0.002
M. galloprovincialis	NA	gills	Illumina	~120 million	3,103	0.003
M. edulis	(Philipp, et al. 2012)	digestive gland	454	1,112,061	30	0.003
M. galloprovincialis	(Gerdol, et al. 2014)	digestive gland	Illumina	~54 million	3,269	0.006
M. galloprovincialis	NA	posterior adductor muscle	Illumina	~103 million	9,429	0.009
M. galloprovincialis	(Romiguier, et al. 2014)	mixed tissues	Illumina	~108 million	11,695	0.011
M. edulis	(Philipp, et al. 2012)	inner mantle	454	323,482	46	0.014
M. edulis	(Romiguier, et al. 2014)	mixed tissues	Illumina	~103 million	15,086	0.015
M. edulis	(González, et al. 2015)	mantle/foot	Illumina	~49 million	8,492	0.017
M. edulis	NA	mixed tissues	Sanger	5,300	1	0.019
M. galloprovincialis	NA	mantle	Illumina	~108 million	30,802	0.029
M. galloprovincialis	(Craft, et al. 2010)	digestive gland	454	33,992	2	0.059
M. edulis	(Philipp, et al. 2012)	mantle rim	454	324,592	299	0.092

Figures

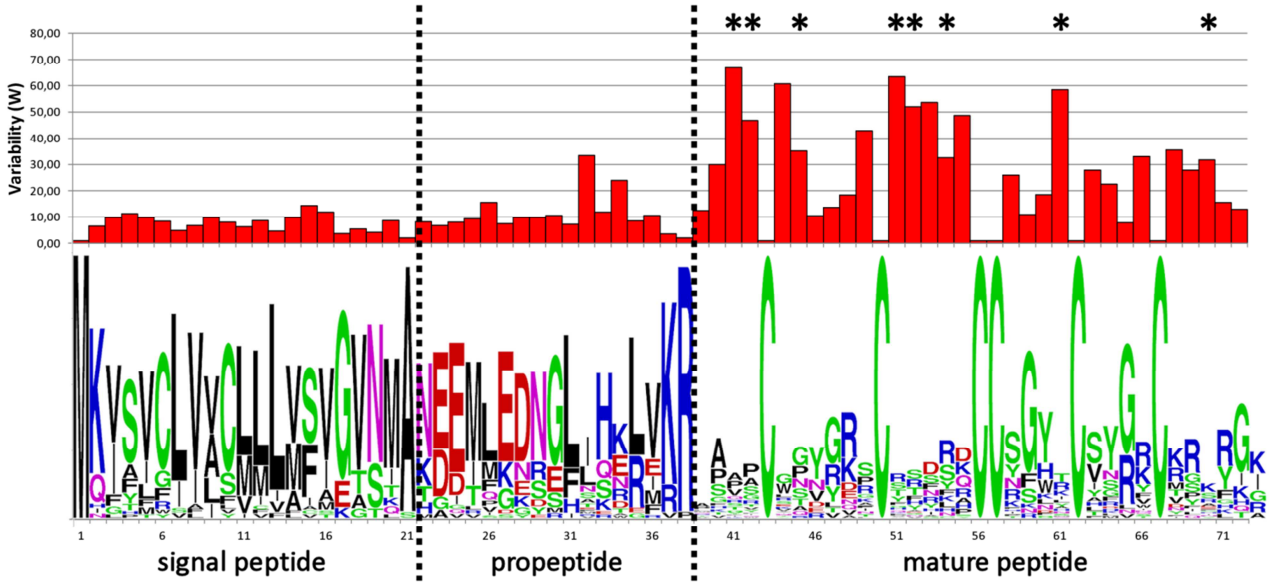

Figure 1

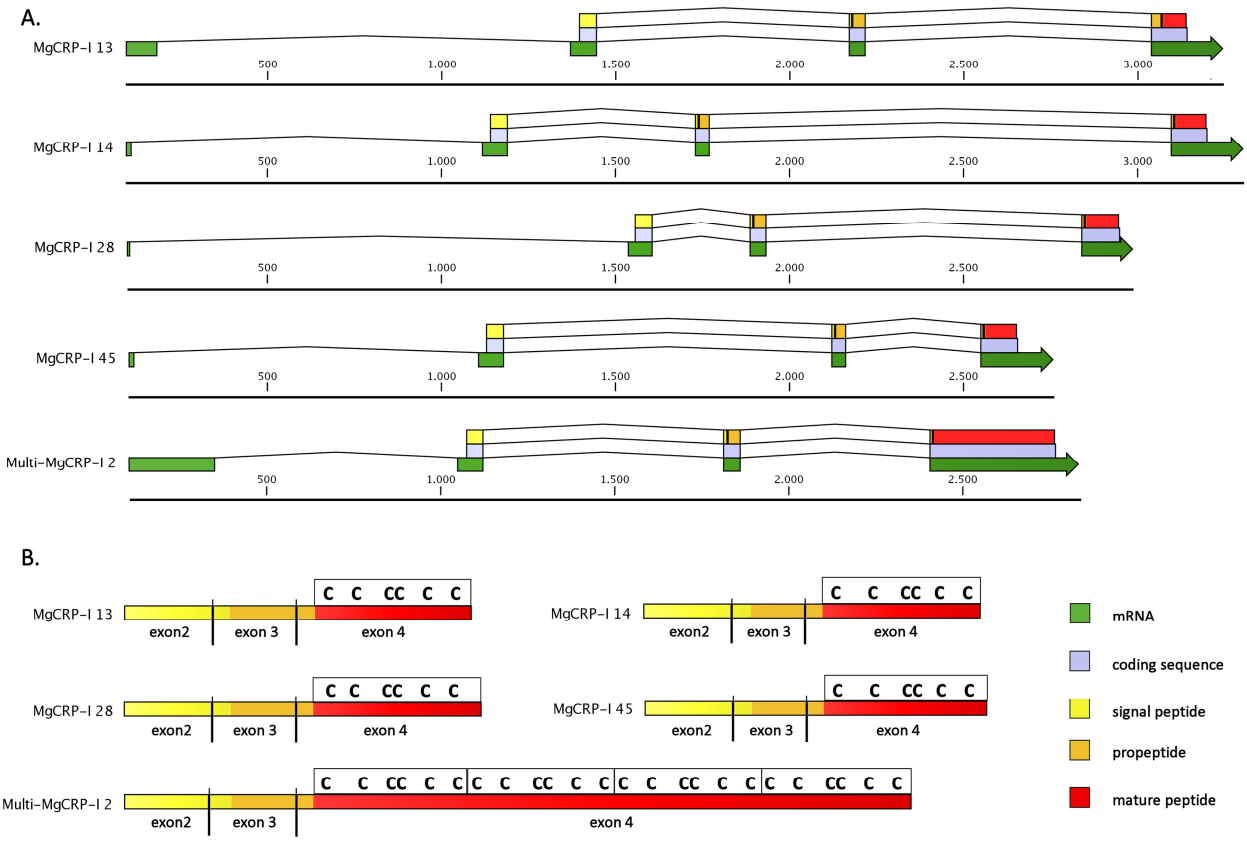

Figure 2

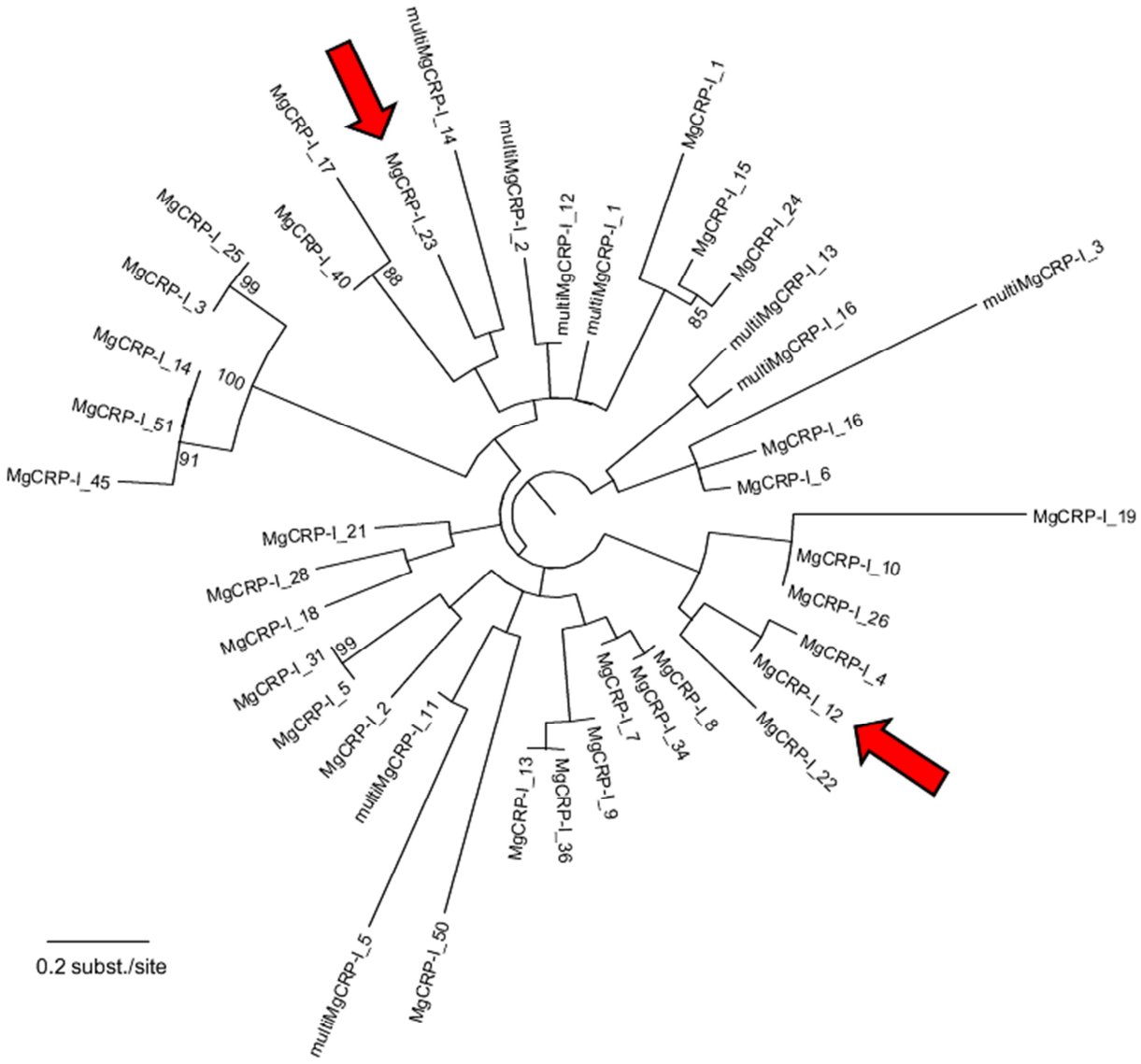

Figure 3

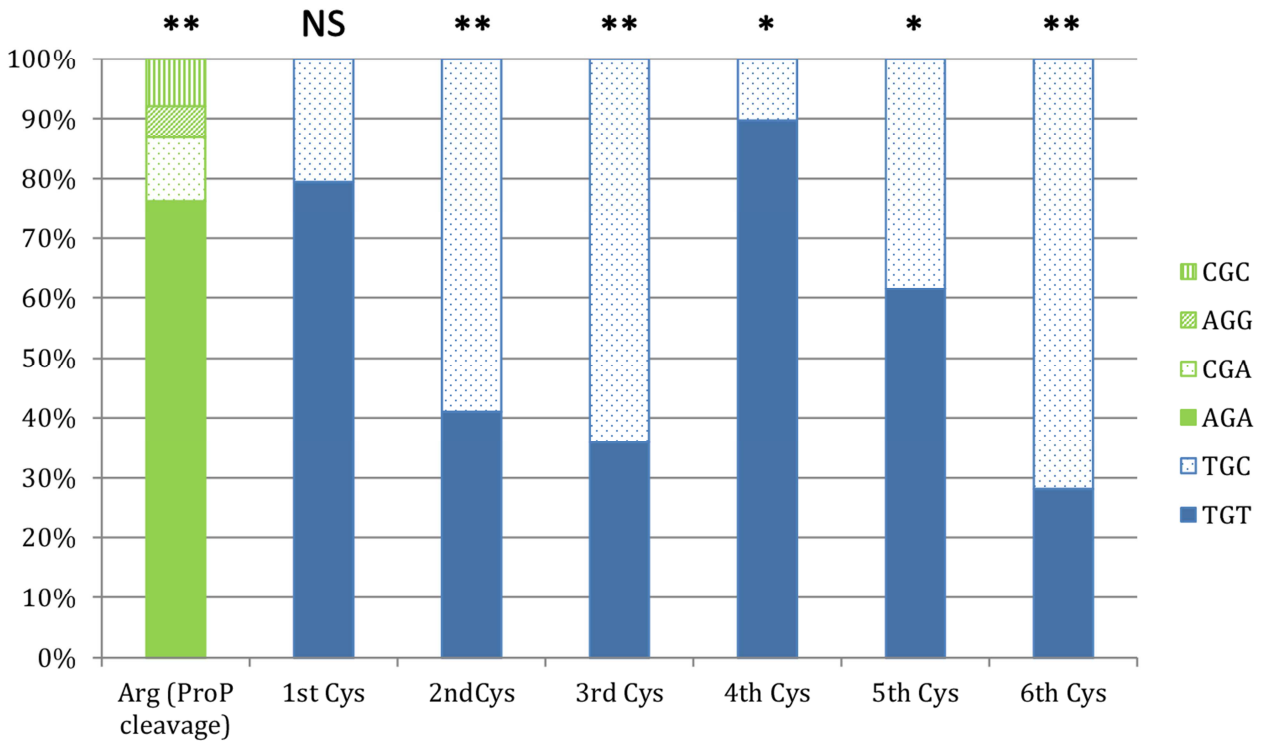

Figure 4

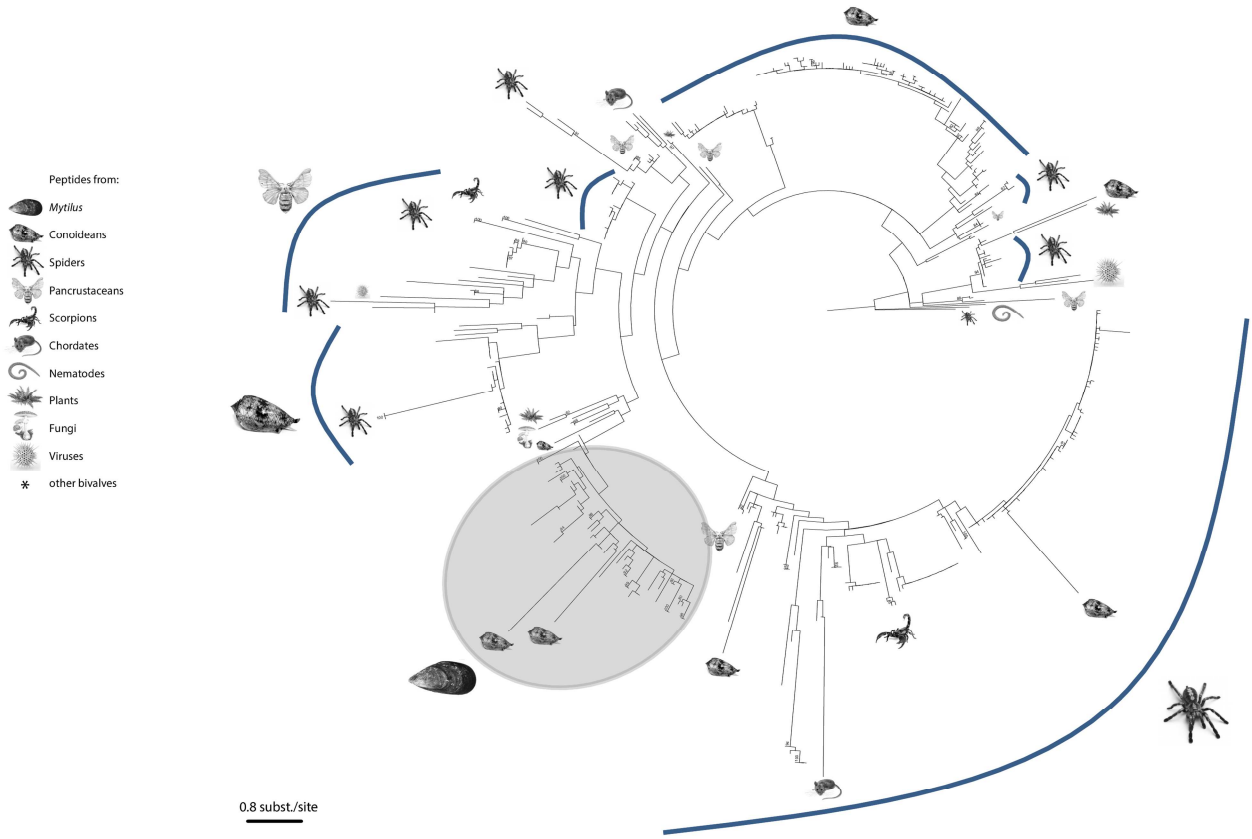

Figure 5

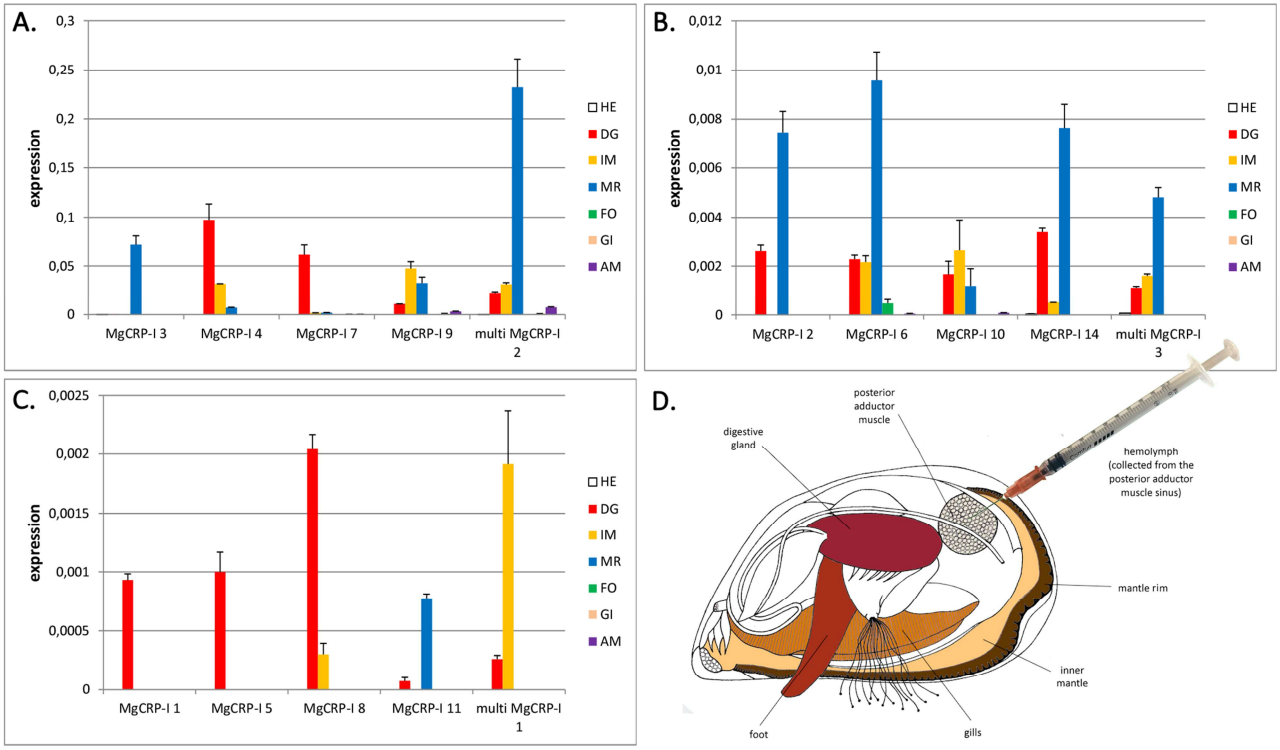

Figure 6