UNIVERSITA' DEGLI STUDI DI TRIESTE

XXIV CICLO DEL DOTTORATO DI RICERCA IN METODOLOGIE DI BIOMONITORAGGIO DELL'ALTERAZIONE AMBIENTALE

UN APPROCCIO TRASCRITTOMICO PER DELINEARE I MECCANISMI MOLECOLARI ALLA BASE DELLA RISPOSTA DEL MITILO *MYTILUS GALLOPROVINCIALIS* A PATOGENI E ALLA CONTAMINAZIONE DA BIOTOSSINE ALGALI

Settore scientifico-disciplinare BIO/18

DOTTORANDO

MARCO GERDOL

COORDINATORE

PROF. PIERLUIGI NIMIS

RELATORE

PROF. ALBERTO PALLAVICINI

ANNO ACCADEMICO 2010/2011

INDICE

INTRODUZIONE

Mytilus galloprovincialis e mitilicolura

Mytilus galloprovincialis (Lamarck, 1819), comunemente noto come mitilo Mediterraneo è, assieme a *Mytilus edulis* (Linneo, 1758), una delle due specie del genere *Mytilus* utilizzata a fini alimentari e soggette a mitilicoltura. Di taglia leggermente inferiore e diffuso alle medie latitudini, in acque più calde rispetto alla specie sorella, maggiormente adattatasi alle temperature più rigide dei mari del nord, il *M. galloprovincialis* è diffuso in tutto il Mar Mediterraneo e sulle coste atlantiche di Spagna e Portogallo fino ad arrivare in Bretagna, dove inizia un area di sovrapposizione con *M. edulis*, nella quale è possibile ritrovare anche ibridi tra le due specie [1]. Oltre a queste zone, probabilmente corrispondenti alle aree di diffusione originarie [2], *M. galloprovincialis* è presente anche sulle coste cinesi ed in limitate regioni di Sud Africa, costa occidentale degli Stati Uniti, Australia e Nuova Zelanda [3, 4], pur tuttavia senza mai raggiungere latitudini tropicali. Questa diffusione è con ogni probabilità da imputarsi al traslocazioni accidentali di esemplari in tempi recenti da parte di navi mercantili. Complessivamente, il genere *Mytilus* è largamente distribuito nelle acque temperate di entrambi gli emisferi del globo ed ha un ruolo di grande importanza nel funzionamento degli ecosistemi bentonici, fornendo un supporto per le comunità delle zone intertidali delle coste rocciose e contribuendo al trasferimento di materia organica e minerale [5]. Una rappresentazione schematica della distribuzione a livello mondiale di *M. galloprovincialis* e *M. edulis*, assieme agli altri principali ecotipi *M californianus* (Lamy, 1936 e Soot-Ryen, 1955) *M. trossolus* (Gould, 1850), è presentata in Figura 1.

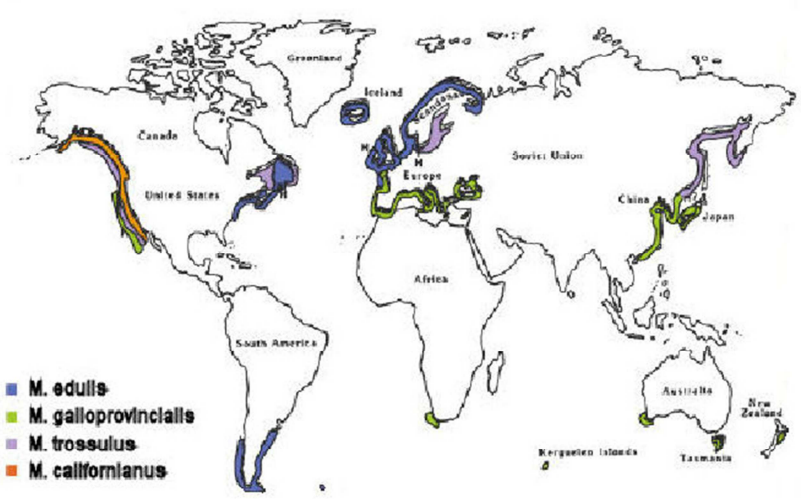

Figura 1: distribuzione mondiale dei principaliecotipi del genere *Mytilus*. Modificato da Gosling, 1992.

I mitili sono organismi sessili, che si nutrono per filtrazione attraverso due coppie di branchie composte da una serie di filamenti paralleli, che sono anche l'organo responsabile degli scambi gassosi. L'assorbimento dei nutrienti avviene poi nella ghiandola digestiva, un organo dal colore

scuro situato al centro del corpo dell'animale (vedi Figura 2). Il mantello è il tessuto a diretto contato con le due valve, che è il responsabile della costruzione del guscio ed è il punto focale della gametogenesi. La riproduzione avviene attraverso il rilascio diretto dei uova e sperma direttamente in acqua, dove la fecondazione ha luogo. Nel corso dei primi stadi di sviluppo , le larve sono provviste di ciglia e possono quindi muoversi attivamente, aumentando così, unitamente alle correnti marine, la propria dispersione. Nell'arco di 4 settimane e dipendentemente da svariati parametri ambientali, le larve vanno incontro a metamorfosi e sviluppano il piede, iniziando la ricerca per un substrato adatto per la vita adulta, finchè la produzione dei filamenti del bisso garantisce un saldo ancoraggio nell'ambiente dove la larva completerà le fasi finali della metamorfosi, divenendo del tutto simile ad un individuo maturo [6].
Un mitilo adulto, in condizioni normali, filtra all'incirca da 4 a 5 litri d'acqua all'ora, trattenendo più o meno il 90% delle particelle in essa contenute, purchè rientrino nelle dimensioni filtrabili. Oltre alle dimensioni, diversi studi recenti hanno dimostrato in alcune specie di bivalvi una preferenza basata anche sulla "qualità" delle particelle, visto che larve e piccoli organismi di forme particolarmente complesse non vengono filtrate [7]. Il mitilo è in grado di captare particelle con un diametro compreso tra i 2 ed i 5 micron e quindi in questo range sono inclusi batteri, numerosi organismi planctonici, particelle di argilla, uova e larve di un gran numero di specie e residui di specie animali e vegetali. L'ingresso di particelle di dimensioni maggiori è impedito dall'orlo del mantello. Il movimento dell'acqua all'interno delle valve è generato dal battito delle ciglia laterali delle branchie (nell'ordine di 2-5 pulazioni al secondo in condizioni normali). Il passaggio della corrente d'acqua tra i filamenti adiacenti delle amibranchie determina la filtrazione attraverso il setaccio delle ciglia latero-frontali, le quali si strovano i stretto contatto tra di loro.

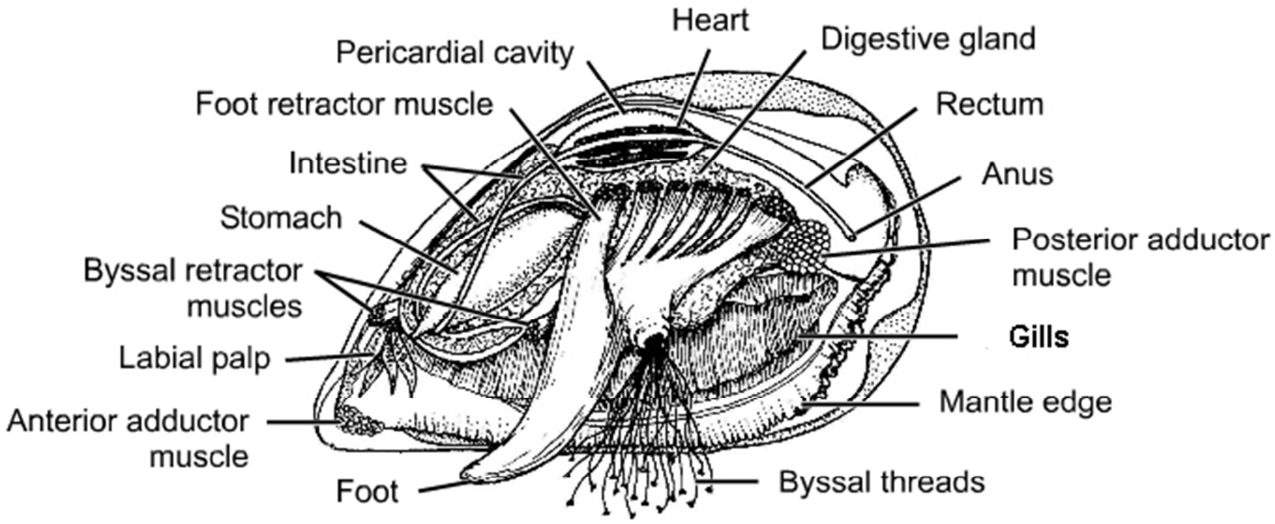

Figura 2: rappresentazione schematica dell'anatomia del mitilo.

I mitili sono organismi di grande rilevanza nel campo dell'acquacultura, con una produzione mondiale di circa un milione di tonnellate l'anno per quanto riguarda *M. galloprovincialis* e circa il doppio per quanto riguarda *M. edulis* (FAO, 2012, figura 3). A livello europeo, il 90% della produzione di molluschi è coperta da mitilo ed ostrica.

La mitilicoltura è un'attività tradizionale e consolidata in molte regioni italiane, rappresentando la principale attività di allevamento di specie acquatiche marine per quantitativi prodotti e per distribuzione sul territorio. La produzione è particolarmente concentrata nelle regioni dell'alto Adriatico (Emilia Romagna, Veneto e Friuli Venezia Giulia), con una produzione che sfiora le 40.000 tonnellate annue con un fatturato annuo che supera i 40 milioni di euro [8]. I metodi comunqmente utilizzati per la coltivazione dei mitili sono sostanzialmente due: l'allevamento su pali fissi e quello su filari galleggianti. Il primo consiste nella disposizione di pali di cemento, legno o metallo collegati tra loro da cavi ai quali vengono appese le reste, delle reti tubolari in materiale plastico contenenti i mitili. Il secondo si basa su corpi galleggianti ancorati al fondo, collegati da lunghi cavi a loro volta mantenuti in prossimità della superficie da una serie di galleggianti, ai quali vengono appese le reste. Al fine di aumentare la produttività, le mitilicolture sono sempre state tradizionalmente realizzate in corrispondenza di bacini chiusi o comunque protetti, vicini a foci di fiumi, di modo da garantire un continuo apporto di nutrienti, adatto a massimizzare la crescita di questi organismi filtratori. Tuttavia questa localizzazione espone i mitili a patogeni ed inquinanti, e questo problema si è fatto sentire in particolare negli ultimi anni, visto il continuo peggioramento della qualità delle acque fluviali. Inoltre, i mitili sono resistenti a molti patogeni che possono causare infezioni e malattie in altri molluschi e per questo motivo sono ritenuti essere dei "serbatoi" di malattie per altre specie [9]. Data la loro alimentazione per filtrazione, i mitili possono accumulare batteri, virus e tossine in seguito a fioriture algali. Proprio questi problemi, monitorati costantemente dalle autorità sanitarie competenti, sono alla base di frequenti chiusure delle mitilicolture in tutto il mondo, nel caso in cui i parametri rilevati superino i limiti previsti per legge.
E'anche per questo motivo che in tempi recenti si è osservata una tendenza allo spostamento delle coltivazioni verso mare aperto, allo scopo di ottenere un prodotto igienicamente conforme e di conseguenza di poter imitare le perdite economiche.

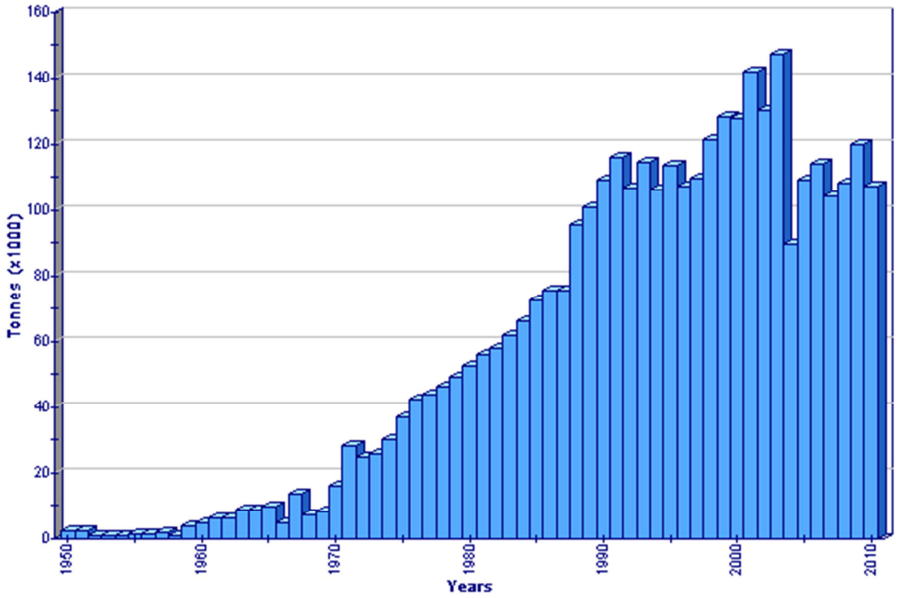

Figura 3: produzione mondiale di *M. galloprovincialis* dal 1950 ad oggi (FAO Fishery Statistics).

Paralytic Shellfish Poisoning (PSP)

Le dinoflagellate o dinoficee sono alghe prevalentemente unicellulari, generalmente provviste di due lunghi flagelli, che rappresentano un'importante componene del fitoplancton anche nel Mediterraneo, dove sono note oltre 300 specie [10].
La fioritura di queste specie algali è spesso assiociata al termine "marea rossa" (red tide), anche se in realtà il termine non è corretto visto che non necessariamente è effettiamente visibile un cambiamento del colore dell'acqua. Lo stesso termine è altresì frequentemente utilizzato per descrivere la fioritura di dinoflagellate tossiche, ma anche in questo caso il termine non è corretto dal momento che anche specie non tossiche possono causare tale fenomeno dal forte impatto visivo [11].
Per questo motivo, quando ci si riferisce a fioriture algali di specie tossiche è più consono riferirsi a "Harmful Algal Blooms" (HABs). La frequenza degli HAB registrati è incrementata notevolmente negli ultimi anni su scala mondiale; se ciò può essere da un lato spiegato con il maggior interesse maturato recentemente nei confronti del fenomeno, dall'altro può essere imputato all'eutrofizzazione delle acque (fenomeno causato dall'aumento dei nutrienti disponibili nel corpo d'acqua) e ad i cambiamenti climatici in atto [12].
Molte specie di dinoflagellate producono, almeno potenzialmente, tossine. Nonostante il ruolo fisiologico che ricoprono nelle dinoflagellate sia tuttora piuttosto oscuro, esse rappresentano un pericolo per la salute umana, ma possono colpire numerosi altri organismi, a partire da altri mammiferi, fino ad arrivare ad uccelli e pesci ed invertebrati marini filtratori. Sono proprio gli organismi filtratori però a rappresentare la causa principe di intossicazioni alimentari da biotossine algali, vista la capacità di accumulare le tossine ad alte concentrazioni nei propri tessuti, veicolandole così nell'intera catena alimentare [13].
Oltre alle ovvie conseguenze ecologiche, con la morte o la compromissione di nuomerosi organismi marini, le HAB hanno importanti ricadute anche in campo economico, dal momento che il turismo balneare, la pesca e la molluschicoltura risentono pesantemente del fenomeno. In particolare la sospensione delle attività di mitilicoltura, spesso protratta per diversi mesi l'anno, comporta ingenti perdite economiche nel settore. Risulta quindi chiaro che le HAB, oltre ad essere un rischio concreto per la salute umana, rappresentano anche un vero e proprio problema sociale.
Le sindromi causate dall'ingestione di molluschi filtratori contaminati con tossine prodotte da dinoflagellate sono variegate e vengono raggruppate in cinque categorie principali, sulla base dei sintomi osservati:

1. *Paralytic Shellfish Poisoning (PSP)*, che verrà descritta più nel dettaglio in seguito.
2. *Diarrhoeic Shellfish Poisoning (DSP)*, provocata dall'acido okadaico e dai suoi analoghi (le dinophysitossine, DTX) oppure dalle pectenotossine (PTX) o infine dalle yesotossine. Classicamente queste tossine provocano sindromi diarroiche.
3. *Azaspiracid Shellfish Poisoning (AZA)*, provocata da neurotossine liposolubili prodotte da *Protopteridinium crassipes* [14], con effetti molto simili alle tossine DSP.
4. *Neurologic Shellfish Poisoning (NSP)*, causata dalla brevitossine, prodotte dalla dinoflagellata *Gymnodinium breve*. Si riscontra, oltre che nell'uomo, in mammiferi ed uccelli marini, ma non nei molluschi. In situazioni particolari si possono formare degli areosol tossici, che causano sindromi asmatiche nell'uomo.

5. *Amnesic Shellfish Poisoning (ASP)*, provocata dall'acido domoico, prodotto però solo da alcune specie di diatomee [12]. E'caratterizzata da crampi addominali, vomito, disorientamento ed amnesie.

Per PSP si intendono le sindromi provocate dalla saxitossina (STX) e dai suoi analoghi (gonyautossine, GTX), prodotte da dinoflagellate del genere *Alexandrium* o da *Gymnodinium catenatum* [15, 16] (vedi figura 4). Le proprietà delle PSTs hanno origine nella loro alta affinità di legame ai canali del sodio voltaggio-dipendenti [17]. Il legame blocca i potenziali di azione, in modo simile alla tetradotossina. I sintomi dell'avvelenamento sono pertanto principalmente di natura neurologica ed includono debolezza, formicolio agli arti, stato confusionale, mancaza di coordinazione mototia e difficoltà nella parola. Anche se nella maggior parte dei casi il recupero è completo e senza particolari complicazioni, in circostanze eccezionali sono stati riportati casi letali con decessi avvenuti per paralisi respiratoria [18].

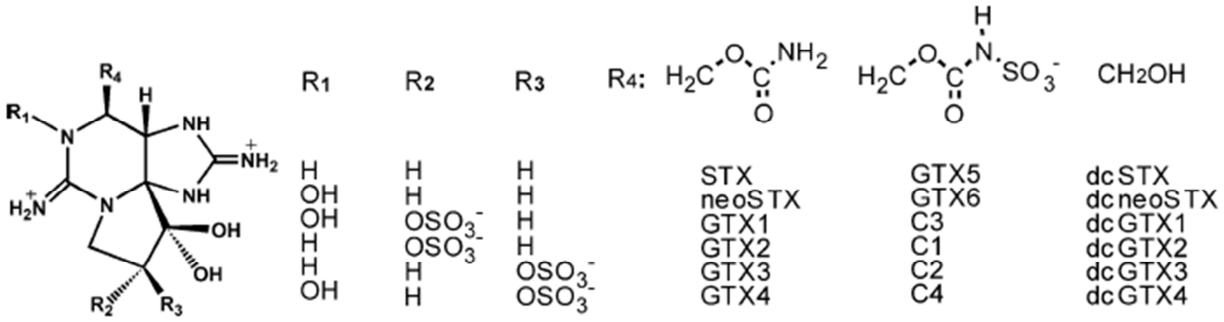

Figura 4: struttura delle tossine paralitiche (PSP). STX: saxitossina; GTX: gonyautossina; dc: decarbamoil-. Tratto ds Lin et al., 2004.

Il PSP è un fenomeno di livello mondiale, che riflette l'amplissima diffusione delle dinoflagellate produttrici delle tossine. Casi sono stati riportati sulle coste atlantiche europee, sia sulla costa occidentale che sulla cosa orientale dell'America del Nord, in in Sud America, in Giappone e nel Mediterraneo. Secondo HAEDAT (http://iodeweb6.vliz.be/haedat/), a partire dal 1987 oltre 800 eventi di PSP sono stati osservati in tutto il ondo (vedi figura 5).
Gli effetti del PSP sono spesso apprezzabili anche negli invertebrati marini, incluse diverse specie di molluschi bivalvi, nelle quali si possono osservare sensibili incrementi nel tasso di mortalità. Nonostante specie diverse mostrino differenti risposte comportamentali a fioriture algali PSP, è chiara la relazione esistente tra sensibilità e la capacità di nutrirsi continuativamente e di conseguenza di accumulare le tossine nei propri tessuti [19, 20]. Le specie maggiormenti sofferenti riducono drasticamente i tassi di filtrazione, limitando quindi l'accumulo, mentre al contrario le specie insensibili, tra le quali rientra anche *M. galloprovincialis*, non modificano le proprie abitudini filtratorie, arrivando ad accumuluare tossine paralitiche a livelli altissimi, specialmente nella ghiandola digestiva. In caso di fioriture eccezionali sono stati osservati accumuli di PSTs fino a 3000 µg STX eq $100g^{-1}$, concentrazioni elevatissime se comparate ai limiti di legge stabiliti dall'Unione Europea e dalla FDA (80 µg STX eq $100g^{-1}$), che possono risultare letali anche nel caso di ingestione di pochi grammi di cibo da parte dell'uomo[21].

Il motivo di tale resistenza agli effetti paralitici di STX ed analoghi è da ricercarsi con ogni probabilità in polimorfismi a livello del canale del sodio voltaggio-dipendente, che riducono drasticamente l'affinità di legame della tossina [22]. Nonostante ciò, sono stati riportati in alcuni casi effetti avversi anche in mitilo, con aumentata mortalità ed ingenti modificazioni istopatologiche [23, 24].

Se gli effetti delle tossine paralitiche sono stati studiati piuttosto approfonditamente in mitilo ed in altri bivalvi per quanto riguarda alterazioni fisiologiche, istologiche e comportamentali, gli effetti dell'accumulo sono stati soltanto marginalmente presi in considerazione da un punto di vista molecolare e pertanto eventuali meccanismi molecolari che portano al riconoscimento, al legame, all'accumulo ed alla depurazione dalle PSTs sono quasi del tutto ignoti.

Una delle tematiche prese in considerazione nel corso di questa tesi riguarda proprio la delucidazione dei meccanismi molecolari attuati da *M. galloprovincialis* per fronteggiare l'accumulo di PSTs. L'analisi è stata svolta attraverso RNA-seq da ghiandola digestiva, l'organo di maggior accumulo nel mitilo, permettendo la comparazione dei profili di espressione di individui cibati con ceppi tossici e non tossici dell'alga unicellulare dinoflagellata *Alexandrium minutum*.

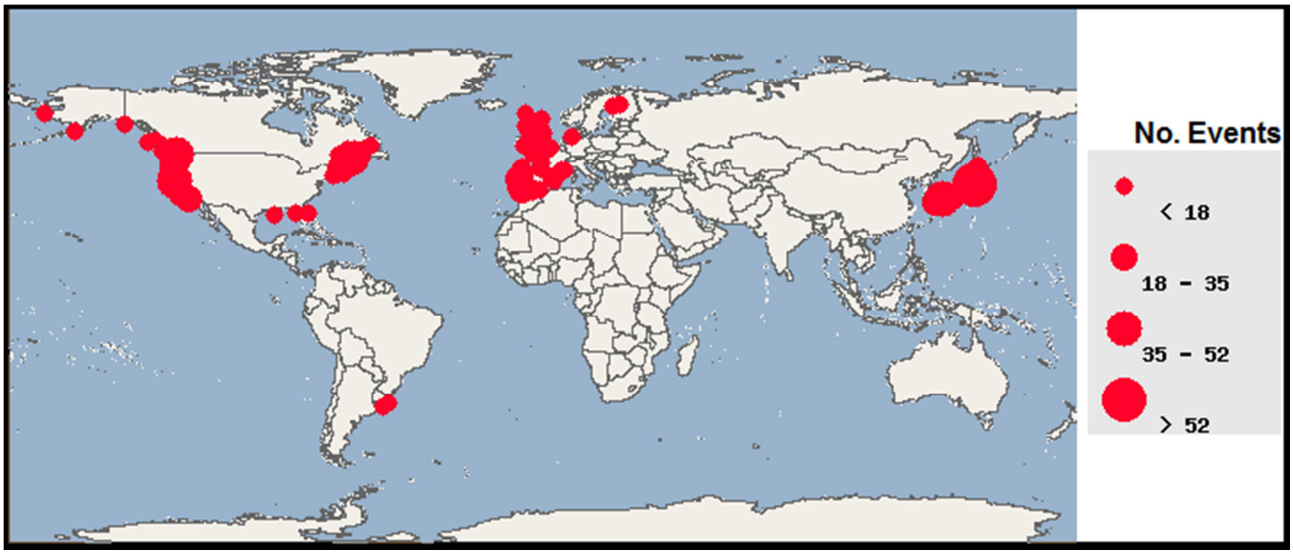

Figura 5: mappa di frequenza degli eventi di PSP registrati in HAEDAT (http://iodeweb6.vliz.be/haedat/).

Il sistema immunitario del mitilo

I mitili, così come tutti i molluschi bivalvi filtratori, ospitano tipicamente un'ampia comunità di organismi commensali, opportunisti e talvolta patogeni. La composizione di tali comunità, comprendenti batteri, virus ed endoparassiti pluricellulari, è estremamente variabile e condizionata da numerosi fattori ambientali, quali temperatura e salinità delle acque, oppura la concomitante presenza di inquinanti e l'associazione con altre specie bentoniche o planctoniche. In particolari condizioni, determinate componenti di questo microbioma possono proliferare fuori controllo, causando talvolta malattie o addirittura provocando la morte dell'ospite, specialmente nel caso delle larve e degli stadi giovanili [25].

I molluschi del genere Mytilus sembrano essere meno soggetti a tali problemi rispetto altri bivalvi dalla grande importanza commerciale, come ad esempio vongole ed ostriche [26, 27]. Data questa marcata tolleranza nei confronti di alcuni patogeni, i mitili possono agire da veri e propri serbatoi di agenti infettivi per altri organismi, incluso l'uomo [9].

I mitili, così come tutti gli invertebrati, sono privi di meccanismi di immunità adattativa e pertanto, per sopravvivere in ambienti in cui l'esposizione ad un'ampia gamma di potenziali agenti patogeni è continua, hanno sviluppato un sistema di difesa molto efficiente, rapido e robusto, basato esclusivamente su componenti innate del sistema immunitario.

In particolare gli emociti, cellule circolanti piuttosto eterogenee coinvolte in una serie di funzioni omeostatiche, sono le principali componenti dei sistemi di difesa cellulari dei molluschi, provvedendo all'uccisione deli patogeni attraverso lisi o fagocitosi. Unitamente agli emociti, una serie di elementi molecolari contribuiscono a rendere i meccanismi di difesa innati dei molluschi estremamente robusti: difatti solo l'azione concertata di lectine transmembrana o secrete nell'ambiente extracellulare, di recettori Toll-like, di enzimi idrolitici, cascate proteolitiche e peptidi antimicrobici è in grado di garantire una risposta così efficace ma allo stesso tempo così specifica e finemente regolata [28].

L'analisi dei dati trascrittomici di mitilo a disposizione ha permesso di identificare un numero elevatissimo di componenti di questi sistemi, organizzati in network altamente interconnessi di vie di segnalazione e di regolazione ([28], vedi Figura 6).

Tuttavia il primo requisito per l'attivazione di una cascata di vie di difesa a valle è la presenza di validi e specifici sistemi di riconoscimento per i "Pathogen Associated Molecular Patterns" (PAMPs), collettivamente noti come "Pathogen Recognition Receptors" (PRPs). Il ruolo di numerose molecole "lectin-like", ad altissima plasticità e potenziale di legame, nel riconoscimento specifico di PAMPs nei molluschi è già stato messo in luce, evidenziando come C-type lectins, FREPs e proteine contenenti domini C1q siano utilizzate in modo estensivo dai molluschi bivalvi a questi fini [29-31].

Uno degli aspetti approfonditi in questa tesi è proprio il ruolo della vastissima famiglia di proteine C1qDC in mitilo. Sulla base dei dati trascrittomici disponibili al momento, essa sembra rappresentare di gran lunga la famiglia di proteine secrete più ampia in *M. galloprovincialis*. Data la notevolissima espansione di questa famiglia genica avvenuta, come evidenziato da uno dei lavori sperimentali allegati, specificamente nella classe Bivalvia, la regolazione positiva della loro espressione in presenza di batteri patogeni e le spiccate proprietà di legame del dominio gC1q, è stato possibile ipotizzare che le proteine C1qDC ricoprano, nei molluschi bivalvi, un ruolo fondamentale come PRPs.

Un altro punto esaminato in dettaglio in questa tesi riguarda i peptidi antimicrobici (AMPs), altri importanti componenti del sistema immunitario umorale. Diffusi in tutti gli orgnismi viventi e caratterizzati da un'incredibile variabilità di sequenza, di struttura e di meccanismi di azione, questi piccoli peptidi sono generalmente accomunati dalla natura anfipatica e cationica. La maggior parte degli AMP finora descritti nei molluschi appartiene alla classe degli AMP ricchi in cisteine [32], la cui struttura tridimensionale è fortemente stabilizzata da ponti disolfuro intramolecolari, anche se anche altri peptidi antimicrobici con altre proprietà sono stati descritti sporadicamente [33].
Tuttavia, tutti gli AMP di mitilo noti appartengono a questa famiglia e sono accomunati dall'essere espressi negli emociti ed accumulati in granuli, per poi essere rilasciati nella forma attiva ed entrare in azione in seguito al riconoscimento della cellula estranea da parte dell'emocita, provocandone la lisi. Più nel dettaglio, gli AMP descritti in mitilo appartengono alle famiglie delle defensine, delle mitiline, delle miticine e delle mitimicine.
In questa tesi è riportata l'identificazione, a partire dal database di sequenze ottenuto grazie allo studio di RNA-seq, di due ulteriori gruppi di AMPs, le big defensine e le mitimicine. In uno dei dovumenti allegati ne sono descritte le caratteristiche salienti e viene presa in esame la storia evolutiva di queste due famiglie di AMP sulla base di studi comparativi.

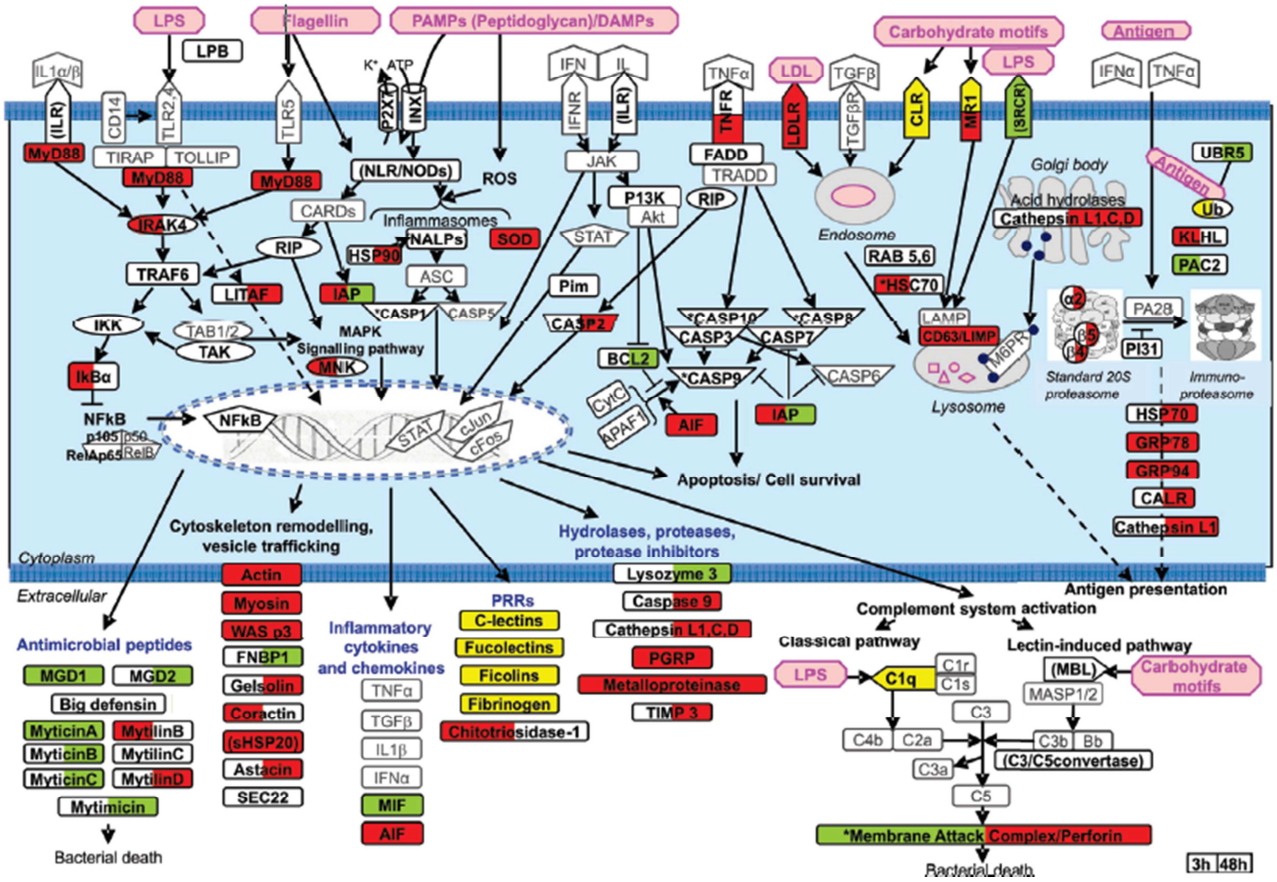

Figura 6: Modello della risposta del mitilo ad iniezione con *Vibrio*. Da Venier et al., 2011.

Le tecnologie di sequenziamento di nuova generazione e la loro applicazione allo studio di trascrittomi

La tecnica di sequenziamento di DNA che più comunemente utilizzata è stata ideata da Frederick Sanger nel 1977 [34]. Si tratta di un metodo enzimatico che si basa sull'utilizzo di dideossinucleotidi marcati che, una volta incorporati nel filamento di DNA, comportano l'interruzione della catena. La marcatura permette il successivo riconoscimento dei frammenti così generati e quindi la ricostruzione dell'intera sequenza di DNA di interesse. Originariamente la marcatura era di tipo radioattivo ed il riconoscimento era effettuato su lastra fotografica; naturalmente nel corso degli oltre trent'anni di storia del sequenziamento Sanger sono state prodotte numerose innovazioni tecniche, che hanno portato all'introduzione di ddNTPs marcati fluorescentemente ed all'automazione del riconoscimento. Nonostante questo metodo sia piuttosto accurato e permetta il sequenziamento di frammenti lunghi diverse centinaia di paia di basi, richiede costi non indifferenti, soprattutto per applicazioni rivolte su scala genomica o trascrittomica. Inoltre questo metodo richiede numerosi passaggi prima del sequenziamento vero e proprio, che di fatto ne limitano la processività.

Negli ultimi anni sono state sviluppate le cosiddette tecniche di sequenziamento di "seconda generazione" o di "nuova generazione" (NGS, abbreviato dal termine, inglese, "Next Generation Sequencing"), che hanno di fatto superato le limitazioni del metodo Sanger e, per molte applicazioni, lo hanno completamente sostituito visti i costi ridotti e l'elevatissima processività (motivo per il quale tali metodologie sono anche note con il termine "high throughput sequencing"). La prima piattaforma di NGS commercializzata è stata l'FLX Genome Sequencer della 454 Life Sciences (Branfrd, Connecticut, USA) nel 2005, ora proprietà del gruppo Roche (Svizzera), seguita nel 2007 dal Genome Analyzer I della Illumina (San Diego, California, USA) e dal SOLiD della Applied Biosystems (Foster City, California, USA). Queste tre piattaforme dominano attualmente il panorama del sequenziamento, anche se si stanno affacciando sul mercato nuove piattaforme di "terza generazione", ancora in fase di test e perfezionamento.

Queste tre piattaforme, pur presentando differenze sostanziali, sono accumunate da una cosiddetta "amplificazione su fase solida", durante la quale singole molecole di DNA vengono immobilizzate in determinate posizioni su un supporto solido ed amplificate tramite PCR. Le differenze principali tra le tre piattaforme risiedono proprio nel metodo attraverso il quale le singole basi incorporate nella catena in crescita vengono riconosciute, ma una descrizione in dettaglio della chimica alla base delle tre tecnice esula da questa tesi; pertanto, per approfondimenti in questo senso si rimanda a specifiche review sull'argomento [35]. E'un dato di fatto che, dopo i primi anni in cui le tecnologie Illumina e 454 si sono divise più o meno equamente il mercato del sequenziamento di nuova generazione, in tempi più recenti le piattaforme Illumina siano emerse come "gold standard" nel settore (figura 7), anche se per particolari applicazioni il pirosequenziamento 454 è ancora preferibile, data la maggior lunghezza dei frammenti sequenziati.

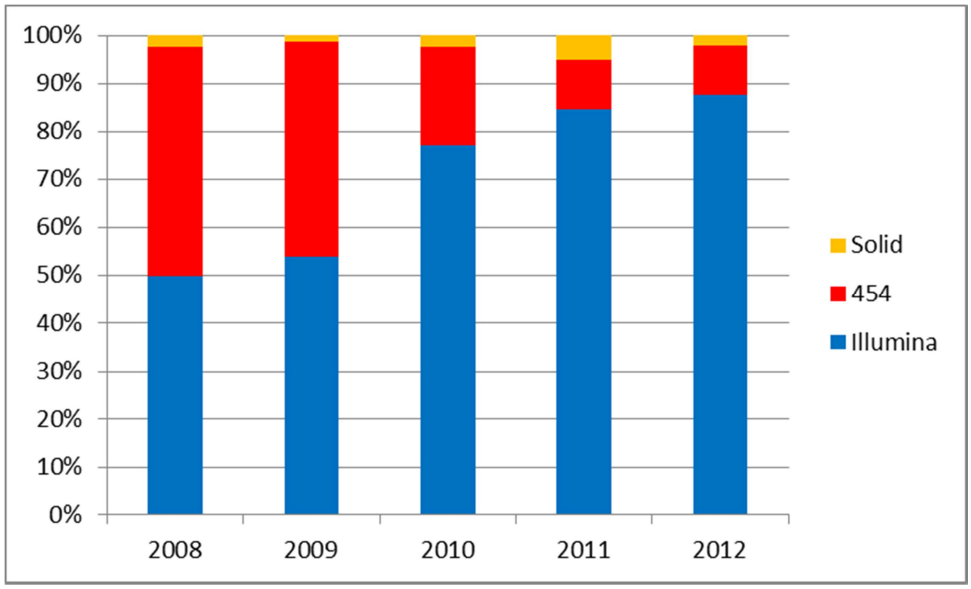

Figura 7: Confronto dell'utilizzo delle 3 principali piattaforme di sequenziamento di nuova generazione (ABI Solid, 454 Life Sciences ed Illumina) nel corso degli ultimi 5 anni (dati ricavati dagli esperimeni depositati negli archivi SRA).

Come già detto, le tecnologie di nuova generazione hanno ridotto enormemente i costi ed i tempi di sequenziamento [36]. Una singola piastra Illumina nei più moderni strumenti HiSeq 2000 genera 600 Gb di dati di sequenza a fronte di costi limitatissimi rispetto alle tecniche Sanger. In effetti questa improvvisa disponibilità a prezzi ragionevoli di tecnologie di sequenziamento a processività così elevata è stata recepita piuttosto rapidamente dalla comunità scientifica, come si evince chiaramente dal numero di esperimenti depositati ogni anno nel database appositamente dedicato "Short Read Archive" (SRA), in crescita esponenziale (figura 8).

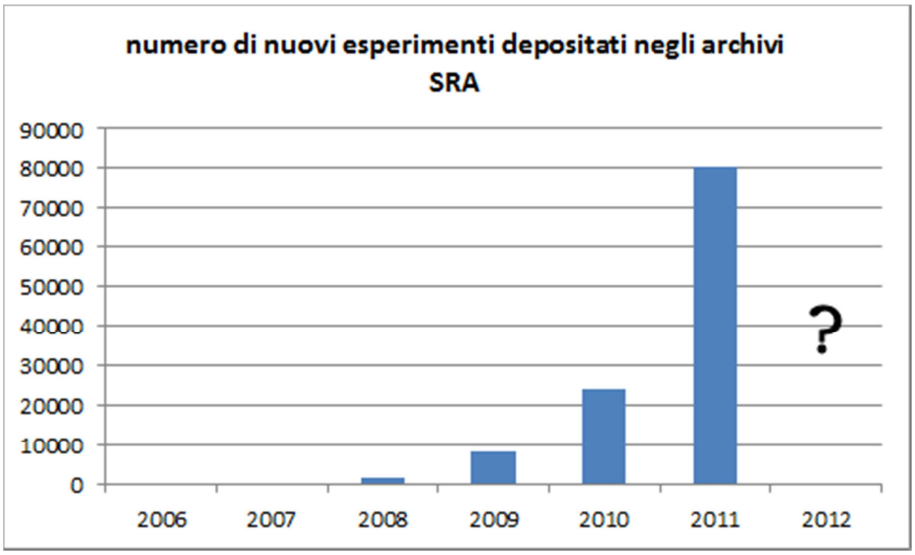

Figura 8: numero di nuovi esperimenti depositati annualmente negli archivi SRA (Short Reads Archive). Si sta assistendo ad una crescita esponenziale dei dati generati con tecnologie di sequenziamento di nuova generazione.

L'enorme mole di dati ottenibili ed i costi ridotti, unite alla possibilità di sequenziare un campione anche in totale assenza di qualsiasi conoscenza molecolare pregressa di un organismo rendono queste metodiche ideali per particolari applicazioni, tra le quali una di particolare rilievo è il "Whole Transcriptome Shotgun Sequencing" (WTSS) o più semplicemente RNA-seq, un vero e proprio strumento rivoluzionario nel campo della trascrittomica [37]. Questa tecnica permette, attraverso il sequenziamento del cDNA di un campione, di ottenere informazioni riguardo al mRNA espresso, consentendo, unitamente allo sviluppo di software ed algoritmi dedicati, l'assemblaggio *de novo* di trascrittomi e la misura molto accurata del livello di espressione di un numero elevatissimo di geni.
Per questo motivo le tecniche NGS hanno ormai quasi completamente soppiantato i sequenziamenti Sanger per quanto riguarda la generazione di una base di dati trascrittomici in organismi non modello. In figura 9 risulta evidente come il numero di Expressed Sequence Tags (ESTs) depositate generate con le classiche tecniche Sanger abbia subito un netto declino negli ultimi anni, parallelamente all'ascesa delle tecniche NGS (figura8).

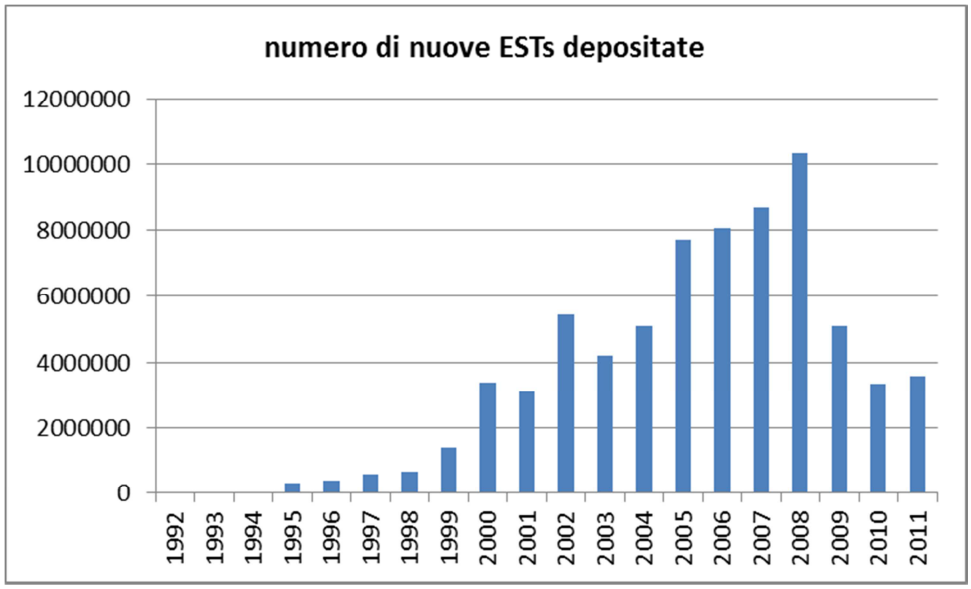

Figura 9: numero di nuove expressed sequence tags (ESTs) depositate annualmente nel database dbEST. Appare evidente che, dopo un incremento costante durato oltre un quindicennio, con il picco massimo toccato nel 2008, si è assistito ad un vistoso declino nel numero di sequenze depositate, in esatta corrispondenza dell'avvento delle tecniche di sequenziamento di nuova generazione.

Anche nel caso specifico dei Molluschi, phylum secondo tra gli invertebrati solo agli Artropodi per numero di specie, l'avvento del NGS ha rapresentato una svolta. Nonostante la notevole importanza di questo phylum, ad oggi un unico genoma è stato sequenziato, quello del Gastropode *Lottia Gigantea* [38] e la totalità dei dati genetici disponibili per ampie classi quali Bivalvi e Cafalopodi risedevano fino a pochi anni fa in sforzi di sequenziamento di poche decine di migliaia di ESTs.
In Tabella 1 sono riportati tutti gli studi di RNA-seq condotti o attualmente in fase di realizzazione per quanto riguarda i Molluschi. L'accessibilità di queste tecniche ha reso recentemente possibile l'applicazione anche a classi poco studiate sulle quali non era disponibile alcuna conoscenza genetica pregressa, permettendo interessanti studi filogenomici che hanno chiarificato le complesse relazioni evolutive tra le classi di Molluschi [39, 40].
L'RNA-seq è sempre più spesso utilizzato anche per studi di espressione genica. Fin dagli anni'90 gli strumenti d'eccellenza per analisi di questo tipo su larga scala sono stati i microarray a DNA. Basati sull'ibridazione competitiva di cDNA marcati con sonde altamente specifiche immobilizzate su un supporto, i microarray permettono l'analisi simultanea dell'espressione di migliaia di geni. Questa tecnica, sebbene sia stata migliorata notevolmente nel corso degli anni, mantiene alcune imperfezioni che condizionano a volte l'ottenimento di risultati riproducibili e altamente attendibili.
In primo luogo, basandosi sulla rilevazione di segnali di fluorescenza risultanti da ibridazione competitiva, i livelli di espressione osservati non possono essere quantificati in modo assoluto, ma soltanto in modo relativo. In secondo luogo gli unici geni la cui espressione può essere monitorata sono quelli la cui sequenza (perlomeno parziale) è già nota. Di fatto, il disegno stesso dei microarray si è sempre basato su precedenti sforzi di sequenziamento di ESTs (oppure, ove disponibili, su genomi sequenziati).

Questi svantaggi sono potenzialmente superabili con un approccio basato sul NGS, dal momento che il numero di letture di sequenziamento ottenute per singolo trascritto può essere convertito in una conta digitale che indica il livello di espressione assoluto. D'altra parte la fase di assemblaggio *de novo* che viene di norma utilizzata prima di procedere all'analisi di espressione vera e propria permette di avere una panoramica completa dell'intero complemento dei trascritti espressi in un determinato tessuto, superando la limitazione imposta dalle sonde spottate su un microarray. Per questo motivo, nonostante i microarray rappresentino ancora un valido strumento in molti casi, l'RNA-seq li sta rapidamente soppiantando come strumento principe per analisi di espressione su scala dell'intero trascrittoma, specialemnte in organismi non modello [37].

In questa tesi si descrive proprio l'applicazione di tecnologie di sequenziamento di nuova generazione al mitilo Mediterraneo *M. galloprovincialis*, che ha consentito lo sviluppo di un database di trascritti espressi in mitilo attraverso assemblaggio *de novo*. Il database così costruito è servito come riferimento per la valutazione di differenze di espressione in individui sottoposti a stimoli diversi (nel caso specifico all'accumulo di tossine paralitiche prodotte da ceppi tossici di *Alexandrium minutum*). Inoltre il database creato rappresenta una risorsa inestimabile per l'identificazione e lo studio di determinate famiglie geniche di fondamentale importanza nei processi di immunità innata di questo organismo e certamente potrà essere la base di ulteriori studi rivolti ad altri argomenti di interesse.

Organism	Class	Tissue	NGS platform	Data output	Year	Paper
Mytilus galloprovincialis	Bivalvia	digestive gland	Illumina	7,4 Gb	2012	Gerdol et al. 2012 (in preparation)
Haliotis diversicolor	Bivalvia	early developmental stages	454	?	2012	\
Ruditapes philippinarum	Bivalvia	larvae	Illumina	2,5 Gb	2012	\
Pomacea canaliculata	Gastropoda	mixed tissues	Illumina	2,3 Gb	2012	\
Ruditapes Philippinarum	Bivalvia	gonad	Illumina	6,4 Gb	2011	Ghiselli et al., 2011
Radix balthica	Gastropoda	mixed tissues	Illumina	1,7 Gb	2011	Feldmeyer et al., 2011
Octopus vulgaris	Cephalopoda	whole body	Illumina	2,4 Gb	2011	Smith et al., 2011
Littorina littorea	Bivalvia	whole body	454	56,9 Mb	2011	Smith et al., 2011
Neomeniomorpha sp 1	Aplacophora	whole body	Illumina	7,3 Gb	2011	Smith et al., 2011
Neomenia megatrapezata	Aplacophora	whole body	Illumina	6,1 Gb	2011	Smith et al., 2011
Gadila tolmiei	Scaphopoda	whole body	Illumina	7,9 Gb	2011	Smith et al., 2011
Ennucula tenuis	Protobranchia	whole body	Illumina	8,1 Gb	2011	Smith et al., 2011
Solemya velum	Bivalvia	whole body	Illumina	6,9 Gb	2011	Smith et al., 2011
Yoldia limatula	Bivalvia	whole body	454	50,5 Mb	2011	Smith et al., 2011
Siphonaria pectinata	Gastropoda	whole body	454	91,2 Mb	2011	Smith et al., 2011
Antalis entalis	Scaphopoda	whole body	454	43,3 Mb	2011	Smith et al., 2011
Perotrochus lucaya	Gastropoda	whole body	454	55,9 Mb	2011	Smith et al., 2011
Nautilus pompilius	Cephalopoda	whole body	454	59,7 Mb	2011	Smith et al., 2011
Lingula anatina	Brachiopoda	whole body	454	38,8 Mb	2011	Smith et al., 2011
Laevipilina hyalina	Monoplacophora	whole body	454	43,4 Mb	2011	Smith et al., 2011
Chaetopleura apiculata	Polyplacophora	whole body	454	84,6 Mb	2011	Smith et al., 2011
Octopus vulgaris	Cephalopoda	whole body	454	2 Mb	2011	Kocot et al., 2011
Nautilus pompilius	Cephalopoda	whole body	454	0,3 Mb	2011	Kocot et al., 2011
Loligo paelei	Cephalopoda	whole body	454	0,5 Mb	2011	Kocot et al., 2011
Tritonia diomedia	Gastropoda	whole body	454	0,5 Mb	2011	Kocot et al., 2011
Antalis vulgaris	Scaphopoda	whole body	454	1,5 Mb	2011	Kocot et al., 2011
Nucula nitidosa	Protobranchia	whole body	454	2,3 Mb	2011	Kocot et al., 2011
Melanoides tuberculatus	Gastropoda	whole body	454	2,3 Mb	2011	Kocot et al., 2011
Solemya velum	Bivalvia	whole body	454	2,2 Mb	2011	Kocot et al., 2011
Theodoxus fluvatilis	Gastropoda	whole body	454	1,9 Mb	2011	Kocot et al., 2011
Scutopus ventrolienatus	Aplacophora	whole body	454	0,3 Mb	2011	Kocot et al., 2011
Wirenia argentea	Aplacophora	whole body	454	0,6 Mb	2011	Kocot et al., 2011
Hanleya sp.	Polyplacophora	whole body	454	0,5 Mb	2011	Kocot et al., 2011
Helisoma trivolvis	Gastropoda	whole body	454	0,4 Mb	2011	Kocot et al., 2011
Hermissenda crassicornis	Gastropoda	whole body	454	0,2 Mb	2011	Kocot et al., 2011
Lymnaea stagnalis	Gastropoda	whole body	454	0,6 Mb	2011	Kocot et al., 2011
Pleurobranchaea californica	Gastropoda	whole body	454	0,8 Mb	2011	Kocot et al., 2011
Ruditapes philippinarum	Bivalvia	whole body	454	285,1 Mb	2011	Milan et al., 2011
Mizuhopecten yessoensis	Bivalvia	mixed tissues	454	433,7 Mb	2011	Hou et al., 2011
Pinctada fucata	Bivalvia	larvae	454	854,7 Mb	2011	\
Aplysia californica	Gastropoda	mixed tissues	Illumina	74,8 Gb	2011	\
Meretrix meretrix	Bivalvia	?	454	?	2011	\
Crassostrea gigas	Bivalvia	?	Solid	4,2 Gb	2011	\
Crassostrea gigas	Bivalvia	larvae	454	798,7 Mb	2011	\
Bankia setacea	Bivalvia	?	Solid	12,6 Gb	2011	\
Alasmidonta varicosa	Bivalvia	?	454	312,7 Mb	2011	\
Haliotis midae	Gastropoda	whole body	Illumina	1,1 Gb	2011	Franchini et al., 2011
Mytilus galloprovincialis	Bivalvia	mixed tissues	454	34,4 Mb	2010	Craft et al., 2010
Nucella lapullus	Gastropoda	?	454	350,9 Mb	2010	\
Placobranchus ocellatus	Gastropoda	whole body	454	595,3 Mb	2010	\
Elysia timida	Gastropoda	whole body	455	615,4 Mb	2010	\
Laternula elliptica	Bivalvia	mantle	454	234,5 Mb	2010	Clark et al., 2010
Bathymodiolus azoricus	Bivalvia	gill	454	237,5 Mb	2010	Bettencourt et al., 2010
Ilyanassa obsoleta	Gastropoda	embryo	454	698,7 Mb	2010	Lambert et al., 2010
Littorina saxatilis	Gastropoda	whole body	454	170 Mb	2010	Galindo et al., 2010
Pinctada margaritifera	Bivalvia	mantle	454	75 Mb	2010	Joubert et al., 2010
Crepidula fornicata	Gastropoda	early developmental stages	454	695.7 Mb	2010	Henry et al., 2010
Strombus gigas	Gastropoda	testis	454	74 Mb	2010	Spade et al., 2010

Tabella 1: Lista riassuntiva degli studi trascrittomici basati su sequenziamento di nuova generazione finora effettuati su specie di molluschi. Come si può notare, il loro numero è in crescita esponenziale e tendenzialmente i lavori più recenti si stanno orientando su piattaforme di sequenziamento Illumina.

PRODUZIONE SCIENTIFICA DEL CANDIDATO

In questa sezione sono presentati i testi integrali originali (in lingiua inglese) dei tre lavori prodotti e seguiti in prima persona dal candidato nel corso di questo dottorato di ricerca. Ciascuno dei tre studi è introdotto da una breve introduzione in lingua italiana, che presenta la tematica e descrive brevemente i risultati salienti ottenuti.
Inoltre sono presentati i brevi riassunti di ulteriori due lavori pubblicati a cui il candidato ha contribuito.
Di seguito la lista dei lavori allegati:

Gerdol M, De Moro G, Manfrin C, Milandri A, Riccardi E, Beran A, Venier P, Pallavicini A: **RNA-seq and de novo digestive gland transcriptome assembly of the mussel *Mytilus galloprovincialis* provide insights on mussel response to paralytic shellfish poisoning**. *Manuscript in preparation* 2012.

Gerdol M, Manfrin C, De Moro G, Figueras A, Novoa B, Venier P, Pallavicini A: **The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A widespread and diverse family of immune-related molecules**. *Developmental & Comparative Immunology* 2011, **35**(6):635-643.

Gerdol M, De Moro G, Manfrin C, Venier P, Pallavicini A: **Big defensins and mytimacins, new AMP families of the Mediterranean mussel *Mytilus galloprovincialis***. *Developmental & Comparative Immunology* 2012, **36**(2):390-399.

Manfrin C, Dreos R, Battistella S, Beran A, Gerdol M, Varotto L, Lanfranchi G, Venier P, Pallavicini A: **Mediterranean Mussel Gene Expression Profile Induced by Okadaic Acid Exposure**. *Environmental Science & Technology* 2010, **44**(21):8276-8283.

Domeneghetti S, Manfrin C, Varotto L, Rosani U, Gerdol M, De Moro G, Pallavicini A, P. V: **How gene expression profiles disclose vital processes and immune responses in *Mytilus* spp.** *ISJ - Invertebrate Survival Journal* 2011, **8**(2).

RNA-seq and *de novo* digestive gland transcriptome assembly of the mussel *Mytilus galloprovincialis* provide insights on mussel response to paralytic shellfish poisoning

Gerdol Marco[a], De Moro Gianluca[a], Manfrin Chiara[a], Milandri Anna[b], Riccardi Elena[b], Beran Alfred[c], Venier Paola[d], Pallavicini Alberto[a]

[a] Laboratory of Genetics, Department of Life Sciences, University of Trieste, Via Licio Giorgeri 5, 34126 Trieste, Italy

[b] Fondazione Centro Ricerche Marine, viale Amerigo Vespucci 2, 47042 Cesenatico, Forlì-Cesena, Italy

[c] Istituto nazionale di Oceanografia e di Geofisica sperimentale, Dipartimento di Oceanografia Biologica, via Auguste Piccard 54, 34151 Santa Croce, Trieste, Italy

[d] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

Manuscript in preparation

Il cossiddetto Paralytic Shelfish Poisoning (PSP) è una delle forme di avvelenamento di origine marina che rappresenta un serio ed emergente problema per la salute umana, oltre a provocare gravi perdite economiche a causa della chiusura delle molluschicolture in tutto il mondo a seguito della fioritura delle alghe produttrici tossine paralitiche, che vengono accumulate dagli organismi filtratori nei propri tessuti. Nonostante gli effetti del PSP siano stati estensivamente studiati nei molluschi ad un livello fisiologico ed isto-patologico, essi sono stati solo marginalmente investigati ad un livello molecolare. Il mitilo *M. galloprovincialis* è in grado di accumulare tossine paralitiche a livelli particolarmente elevati e dunque potenzialmente rischiosi per il consumo umano. Nonostante la sua importanza commerciale, gli studi molecolari su questo organismo sono ancora agli stadi iniziali principalmente a causa dei limitati dati di sequenza disponibili al momento, problema comune alla maggior parte degli invertebrati, insetti esclusi.
In questo lavoro sono state sfruttate le più recenti tecnologie di sequenziameno di nuova generazione per un'esperimento di RNA-sequencing, al fine di comparare i profili di espressione genica di mitili nutriti con ceppi tossigenici e non-tossigenici dell'alga dinoflagellata *Alexandrium minutum*. Questo studio ha permesso di investigare la risposta trascrizionale del mitilo all'accumulo delle tossine paralitiche, fornendo anche le basi per la costruzione di un avanzato database di sequenze espresse nella ghiandola digestiva di questo organismo.
Il sequenziamento, effettuato su piattaforme Illumina, ha generato 7,4 Gb di dati di sequenza, che sono stati utilizzati assemblare *de novo* 39,289 contigs, con una lunghezza media di 689 paia di basi. Inoltre, gli stessi dati sono stati utilizzati per ottenere dati di espressione, rivelando che l'accumulo delle tossine non ha avuto effetti apprezzabili sui mitili. Infatti, è stato possibile identificare solamente 16 geni come differenzialmente espressi nella ghiandola digestiva degli animali nutriti con il ceppo tossigenico. Cionondimeno, successivi esperimenti di real-time PCR condotti su un più ampio set di tempi sperimentali hanno rivelato che i 16 geni identificati erano probabilmente falsi positivi, visto che i profili di espressione ottenuti non sono risultati essere

compatibili con geni responsivi all'accumulo di tossine. Per questo motivo non è stato possibile identificare dei chiari marker molecolare per il PSP in mitilo.
Il nostro studio è stato il primo ad investigare gli effetti dell'accumulo di tossine paralitiche nei molluschi da un punto di vista molecolare. Sebbene non definitivi, i risultati supportano la classificazione dei mitili come organismi insensibili al PSP, visto che nessun processo molecolare in particolare è stato alterato in modo marcato in risposta all'accumulo delle tossine. Ciò è in pieno accordo con la maggior parte degli studi fisiologici e comportamentali già condotti nelle varie specie di mitilo, che risultano essere tra le più resistenti nei molluschi.
Inoltre l'assemblaggio *de novo* e l'annotazione del trascrittoma di mitilo ha notevolmente implementato i dati pre-esistenti inlusi in Mytibase ed ha fornito uno dei pochi database di sequenze disponibili al momento basati su tecnologie di sequenziamento di nuova generazione. Questo database potrà sicuramente rappresentate una risorsa importantissima come supporto a futuri studi genomici in questa specie, ed in parte ha già dato i suoi frutti anche in altri studi dedicati a specifiche famiglie geniche presentati in questa tesi.

TESTO ORIGINALE

Abstract

Background

Paralytic shellfish poisoning (PSP) represents a serious and emerging issue for human health and causes severe economic losses worldwide, due to the closure of shellfish aquacultures. Although the possible physiological and histopathological effects of PSP on mollusks have been extensively studied, they have only been marginally investigated at a molecular level. *Mytilus galloprovincialis* is able to accumulate high levels of paralytic toxins, thus causing health risks in the case of human consumption. Despite its commercial importance, the molecular research on this organism is still at the early stages, mainly because of the limited amount of *Mytilus* sequence data currently available.
We used deep RNA sequencing to compare gene expression profiles of the digestive gland in mussels fed for 5 days with toxic and non-toxic strains of the dinoflagellate *Alexandrium minutum*, and to extend the transcriptional mussel dataset as a basis for future studies.

Results

RNA-sequencing generated 7,4 Gb data and allowed the assembly of 39,289 unique contigs, with a mean length of 689 bp. The gene expression analysis based on the new sequencing reads indicated that toxin accumulation scarcely affected mussels. In fact, only 16 genes were identified as possibly differentially expressed in the digestive gland of mussels fed with the toxin-producing strain, and real-time PCR performed on a larger set of time points indicated such gene expression changes as probable false positives. Therefore, we could not identify a reliable molecular marker for paralytic shellfish poisoning in mussels.

Conclusion

To our knowledge, this study presents for the first time a large-scale molecular analysis of the effects possibly induced by the accumulation of paralytic shellfish toxins (PSTs) in mollusks. Although not conclusive, the overall lack of transcriptional changes support the classification of mussels as organisms refractory to PSP toxins, experimental evidence consistent with the majority of the previous physiological and behavioral studies. Comprehensive *de novo* assembly of the publicly available mussel ESTs with the new dataset, and bulk re-annotation of mussel transcriptome, yielded one of the few NGS-based molluscan sequence databases available to date, a basic resource for expanding functional genomics investigations in the Mediterranean mussel.

Keywords: *Mytilus galloprovincialis*, *Alexandrium minutum*, paralytic shellfish poisoning

Abbreviations: PSP: paralytic shellfish poisoning; PSTs: paralityc shellfish toxins; NGS: next generation sequencing; STX: saxitoxin; HABs: harmful algal blooms.

Background

PSP (paralytic shellfish poisoning) is a syndrome associated with the consumption of filter-feeding mollusks contaminated with toxins usually produced by various unicellular algae. Although paralytic shellfish toxins (PSTs) can be produced by some cyanobacteria species [1], the organisms most commonly associated with PSP are dinoflagellates, such as *Alexandrium catenella*, *A. tamarense* [2] *A. minutum* [3], *A. cohorticula* [4], *A. fundyense* [5], *A. ostenfeldii* [6], Gymnodinium catenatum [7] and Pyrodinium bahamense [8]. Filter-feeding organisms, including bivalves, can accumulate PSTs at very high concentrations and act as lethal vectors of toxins for organisms at higher trophic levels, including humans.

PSTs are toxins analogue to STX (saxitoxin) and their paralytic effects depend on their high affinity to the neuronal voltage-gated sodium channels [9]. The binding of STX to the channel blocks action potentials, in a similar fashion to tetrodotoxin [10]. The symptoms of intoxication are mainly neurological and include numbness, tingling, weakness, shortness of breath and ataxia in humans [11]. While recovery is generally complete and uncomplicated, in some cases respiratory paralysis and death may occur, especially with the consumption of heavily contaminated mollusks [12].

PSP is a phenomenon widespread worldwide, reflecting the broad distribution of the causative algae. In fact, cases of PSP toxicity have been extensively reported in Japan [13-15], both in the eastern and the western coast of Northern America [16-18] and Southern America [19, 20], Britain [21], and on the Atlantic coasts of the Iberian peninsula [22, 23] and France [24]. Sporadic cases have also been described elsewhere, i.e. in the Mediterranean Sea [25-27]. According to HAEDAT (The Harmful Algae Event Database, http://iodeweb6.vliz.be/haedat), almost 800 blooms of PSP-producing dinoflagellates have been recorded worldwide since 1987.

PSP certainly represents a serious threat for human health [11], but also causes severe economic damage to the aquaculture industry because of the closure of bivalve shellfish cultivation grounds [28, 29]. The current toxicity limits set by both EU (Regulation (EC) No 853/2004 of the European Parliament) and FDA (Compliance Policy Guide Sec. 540.250) for human consumption of shellfish is set at 80 μg STX eq $100g_{-1}$ meat. Considering the possibility of ingestion of large quantities of shellfish meat, the European Food Safety Authority established that the limit concentration in shellfish for human consumption should be considerably lower, about 7.5 μg STX eq $100g^{-1}$, in order not to exceed the acute reference dose (ARfD) of 0.5 μg STX eq kg^{-1} body weight [30].

Likewise human and other vertebrates [31-34], several species of mollusks are affected by the paralytic effect of PSPs. Shell valve closure, siphon retraction and burrowing incapacitation are the most commonly observed effects in susceptible species such as *Mya arenaria* and *Geukensia demissa*, whereas others such as *Spisula solidissima* and *Modiolus modiolus* seem to be completely unaffected [35]. Although different species display different behavioral responses to paralytic HABs, there is a clear negative relationship between the sensitivity to PSTs and the ability to feed and consequently to accumulate toxins [35, 36]. One of the most common behavioral modifications observed in susceptible bivalves is the reduction of filtration rate [37-39], which could be either interpreted as a paralytic effect or as a strategy adopted to avoid contamination [40, 41]. Other mechanisms adopted by susceptible species to reduce the intoxication involve toxin accumulation in specific tissues [42, 43], binding of PSTs to sequestering proteins [44], enzymatic or chemical

transformation and degradation reactions [45-47], though it is not clear whether these processes depend on symbiotic bacteria or the bivalves themselves [48, 49].

Electrophysiological studies have demonstrated that mussel nerves are insensible to the paralytic effects of STX [50, 51]. Such a resistance may reflect adaptive evolution to recurrent toxic algal blooms, since a direct link between the sensitivity to PSTs and the frequency of red tides has been observed in clam populations. This hypothesis is supported by the evidence of sodium channel mutations leading to a decreased affinity to toxins in resistant populations [52, 53]. Based on the substantial lack of physiological and behavioral changes [54] also when fed with PSP-producing dinoflagellates [55, 56] mussels (*M. edulis*, *M. galloprovincialis*) are considered refractory to PSP. On the other hand, some cases of adverse effects have been reported. In fact, increased valve closure, decreased filtration rates, and a reduced byssus production have been occasionally observed and associated with an increased mortality in *M. edulis* individuals fed with toxic *A. tamarense* [57, 58] and extensive histopathological modifications have been described in individuals exposed to *A.* fundyense [59].

While the kinetics of toxin accumulation and decontamination in mussels have been thoroughly investigated, relatively little attention has been paid to molecular aspects of the mussel response to PSP. Detectable changes occurring in response to toxin accumulation could be used as early warning signals of contamination, and reveal which strategy, if any, mussels adopt to cope with significant amounts of bioaccumulated PSTs.

In the present study, we investigated the response of the Mediterranean mussel to in vivo bioaccumulated PSTs by comparing the digestive gland transcription profiles of animals fed with toxigenic or non-toxigenic strains of the dinoflagellate *A. minutum*. Since the analysis was conducted by RNA-seq, the new sequence data also allowed the de novo assembly of *M. galloprovincialis* digestive gland transcriptome. The advent of next generation sequencing is increasingly providing the opportunity to expand the “omics” studies to non-model invertebrates [60]. Based on the 454 [61-67], SOLiD [68] and Illumina [69] technologies, new and deeper investigation of the bivalve transcriptomes is paving the way to phylogenomics, i.e. an improved view of the evolutionary relationships within mollusks based on deep sequencing [70, 71]. The sequencing effort described in the present work has significantly enriched the total number nucleotide sequences of *M. galloprovincialis* [64, 72] and gave us the opportunity to create one of the few Illumina sequencing-based transcript databases available to date in the Mollusca phylum.

Results and discussion

Toxin accumulation

Concentrations of *A. minutum* varying from 1 to 47x 10^6 cells L^{-1} have been reported in toxic blooms [73-77]. We exposed adult *M. galloprovincialis* individuals for five days to 5x10^6 cells L^{-1} of the PSP-producing *A. minutum* AL9T strain, a significant but not extreme concentration selected to simulate mussel PSP contamination at levels comparable to those commonly observed during PSP-producing dinoflagellate blooms. Another group of mussels was exposed to identical concentrations of the non-toxigenic strain AL1T in parallel. Mussels from both groups were sacrificed at selected time points (5 days of intoxication followed by depuration), before the feeding, to evaluate the toxin accumulation and to collect digestive glands for RNA extraction. The experimental design is summarized in Table 1 and detailed in the Methods section.

The toxicity tests revealed the production of toxins by the *A. minutum* strain AL9T, with a concentration of 70,6 +5 fg STXdiHCleq\cell, whereas as expected the strain AL1T did not produce any toxins. The estimate of toxin bioaccumulation was performed on the soft mussel tissues, with the digestive gland being taken apart for RNA extraction. The PSP levels in the remaining tissues resulted to be about 15 μg STX eq per 100 grams of meat at T4 (five days from the start of the experiment). Visceral organs are known to accumulate approximately 95% of PSTs in mussel [55]: considering the removal of the digestive gland, the accumulation of toxins produced by the feeding protocol can be estimated to be around 300 μg STX eq 100 g^{-1} of meat, well above the EU and US limits (set at 80 μg STX eq 100g-1 meat). Although PSTs accumulation was detected also at T1 in the AL9T strain-fed mussels, it was not possible to exactly calculate the PSTs concentration in soft tissues deprived of the digestive gland, as it was below the limit of quantification of the method used. Nevertheless, studies previously conducted on the kinetics of paralytic toxins accumulations revealed that mussels accumulate toxins at very high rates, resulting in high toxicity in the matter of a few hours [78-80].

Although exceptional blooms can produce massive toxification, with PSTs reaching concentrations up to 30,000 STX eq 100 g^{-1} meat [79], the accumulation we achieved is better suited for the detection of possible PSP-specific biomarkers, as during an extreme event of contamination a general stress response could be activated.

De novo assembly of *Mytilus galloprovincialis* digestive gland transcriptome

The Illumina sequencing of the digestive gland samples collected at the two experimental time points T1 and T4 from the mussel groups (see Table 1), generated 74,470,393 trimmed nucleotide reads (129,003 single and 74,341,390 paired-end reads). The average read length was 97.75 bp, overall equivalent to ~7,4 GB of sequence. Table 2 summarizes the trimming statistics and the number of trimmed reads per sample. The raw Illumina sequencing reads have been deposited at the NCBI Short Read Archive (study ID: SRP011280.2). Aiming at the creation of a high quality set of sequences expressed in the digestive gland, the preliminary assembly of the Illumina trimmed reads with the pre-existing Sanger and 454 Life Sciences sequences plus an additional set of Illumina reads, required a careful filtering procedure. This contig processing step was used to overcome the creation of short and low quality or misassembled contigs, (a problem commonly arising from the assembly of next generation sequencing data [81]), and to remove contaminant sequences (mainly

arising from ingested *A. minutum* cells). The contigs processing step contributed to a remarkable improvement of the assembly quality (see Figure 1), producing a shift of contig lengths towards higher ranges and reducing the bias towards short, misassembled contigs.
Overall, the final assembly resulted in a collection of high quality *M. galloprovincialis* transcripts representative of the digestive gland which included 39,289 unique assembled contigs. The average length of the contigs obtained, variable from 250 (the minimum length allowed) to 13,211 bp, was 689 bp. The N50 statistic of the assembly was 814, with the average and median lengths being 689 bp and 510 bp, respectively. A total of 6,545 contigs longer than 1 Kb was generated, the longest one being over 14 Kb. A summary of the assembly statistics is shown in Table 2. In the present work we describe one of the first de novo assemblies of a bivalve mollusk transcriptome based on Illumina sequencing data, which was able to generate a high quality collection of transcripts expressed in the digestive gland of the Mediterranean mussel *M. galloprovincialis*.

RNA-seq expression analysis

After the initial contig filtering step, an additional selection was applied in order to remove transcripts whose expression was too low prior to the RNA-seq expression analysis (only 5,523 contigs with a global average coverage higher than 5 were selected for the analysis). In fact the sequencing depth applied was certainly able to allow the creation of many contigs resulting from transcripts which were not digestive gland-specific and whose low expression levels could therefore be subject to random fluctuations. Furthermore, since the representation of these transcripts is highly dependent on the total number of reads obtained from a sample, it has been demonstrated that sensible random variability can occur, even between technical replicates, when the sequence coverage is particularly low [82]. Overall, the removal of these contigs guaranteed the achievement of a less noisy dataset, less prone to false positives detection.
The statistical analysis performed with DEB [83] revealed a very low number of genes as significantly differentially expressed (FDR <0.01) in the AL9T toxigenic strain-fed mussels in both the time points analyzed (T1 and T4, see Table 1). More in detail, 21 genes were identified by EdgeR [84] and 66 genes by DESeq. A schematic representation of the DEB results is summarized in the Venn diagram in Figure 7. Only the 20 transcripts identified by both EdgeR and DEseq were subject to further analysis. The inspection of sequencing reads mappings permitted to identify six contigs which were created by misassembly, which were therefore disregarded.
Twelve out of the 14 resulting contigs were up-regulated in respect with both the non-toxigenic strain-fed and the standard diet control mussels, while, on the contrary, only two transcripts displaying significant down-regulation were identified. The complete list of the differentially expressed contigs in the toxigenic dinoflagellate-fed mussels and the detected expression levels are presented in Table 3.
The functional classification could not directly link any of the positively regulated sequences to PSP. More in detail, seven contigs were found to have no similarity with other known sequences (and were therefore named “transcript of unknown function”), reflecting the low representation of molluscan sequences in public databases, often resulting in difficult annotation of poorly conserved sequences.
Two contigs were pertaining to the C1qDC family, which is an extremely large class of immunity related lectin-like molecules, possibly including several hundred genes in bivalves [85] and representing the largest protein family in *M. galloprovincialis*. In fact, the abundance of C1qDC

transcripts appears to be close to 1,5%, as 580 contigs were annotated as containing the C1q domain IPR001073 (see Table 4). One sequence encoded an IMAP family GTPase, characterized by the AIG1 InterPro domain IPR006703, one of the most commonly found in mussel, and one contig was annotated as C-type lectin, another widely represented family, with more than 200 contigs identified. The three remaining transcripts encoded a hemicentin-1-like protein, a zona pellucida sperm binding domain containing protein and a DED domain containing protein. Hemicentin-1 (also known as fibulin-6) is a protein rich in immunoglobulin-like domains, with a structural function in the extracellular environment [86]. The zona pellucida binding domain is found in a variety of eukaryotic receptor-like glycoproteins [87], which is also rather common in mussel (82 contigs harboring a IPR001507 were found). Finally, DED-containing proteins are often involved in the regulation of apoptosis [88].
Despite the important role of these molecules in many different aspects of mussel life, none of them could be directly linked to functions related to toxin accumulation, excretion, transport or metabolism.

Real Time PCR

The real time PCR expression analysis, performed on samples from all the experimental time points available (T0 to T7, see Table 1), was used to monitor the expression trends of 4 selected genes (transcript of unknown function I and II, C1q domain-containing protein I and hemicentin 1-like) identified as differentially expressed by the edgeR and DEseq analyses in mussels fed with toxigenic (AL9T) and non-toxigenic (AL1T) strains of *A. minutum* and to validate the expression data gathered with RNA deep sequencing. The results confirmed the data obtained at T1 and T4, showing significantly higher expression values in the AL9T strain-fed mussels in all cases (see Figure 3).
On the other hand the analysis also revealed remarkable fluctuations in the expression levels throughout the experimental time points, apparently independent from toxin accumulation. The complete expression profiles obtained for each of the 4 genes analyzed (see Table 3) are shown in Figure 3. In general, the expression trends observed were substantially not consistent with those expected from transcripts positively regulated in response to PSP accumulation, often showing an expression level lower by several orders of magnitude in the AL9T strain-fed samples in respect with the T0 control. Therefore, although the real-time PCR confirmed the experimental data obtained with RNA-sequencing, it also showed that all the four tested candidate PSP-responsive genes were just subject to random expression fluctuations, which showed trends consistent with PSP responsiveness just by chance in the two time points analyzed by RNA-seq (T1 and T4), but failed to do so when the expression analysis was extended to the entire experimental time-course.

Transcriptome annotation

The 86,94% (34,157 out of the total) contigs had a positive BLASTx hit in the NCBI nr protein database: The proportion of contigs without similarity with published sequences is lower than the one previously observed in Mytibase [72] and reflects the higher overall average transcript length. The BLAST e-value and sequence similarity distributions are shown in Figure 4. Not surprisingly, the most represented species as top BLAST hits are invertebrates whose genome has been fully sequenced and released, namely the cephalocordate *Branchiostoma floridae*, the hemicordate *Saccoglossus kovalevskii*, the echinoderm *Strongylocentrotus purpuratus* and the cnidarian

Nematostella vectensis (see Figure 5). Despite the few *Mytilus* sequences available in the nr database, *M. galloprovincialis* stands as the fifth top hit species. Because of the strong bias towards vertebrate sequences produced by genome sequencing efforts, organisms such as *Danio rerio*, *Oreochromis niloticus*, *Xenopus tropicalis*, *Mus musculus* and *Homo sapiens* are also heavily represented as top hit species, despite the extremely large evolutionary divergence with mussel. Nevertheless, the BLAST results are distributed within a very broad range of organisms, as shown by the very large "others" bar in Figure 5.

Due to the limited sequence resources available for mollusks in public databases, homologies were also investigated in the only fully sequenced molluscan genome available to date, the gastropod snail *Lottia gigantea* [89]. The 54% of assembled contigs found a significant similarity (e-value < 10^{-6}) to the proteins predicted from the snail genome, and in 16% cases the e-value was highly significant (< $10e^{-50}$, see Figure 6 for details).

Transcript fragmentation likely occurred, especially in the case of long and poorly covered sequences, as suggested by the contig length distribution, dominated by rather short contigs (see Figure 1). To address this issue, an ortholog hit ratio analysis was used [90]. Although this measure is strongly influenced by the availability of sequence data from closely related organisms, due to evolutionary divergence resulting in sensible underestimations, it can still be helpful to give a rough idea of the transcripts integrity. In this specific case, to balance the unavailability of fully sequenced bivalve genomes, the analysis was performed on putatively highly conserved sequences from the nr database. Although the correct assessment of transcripts fragmentation is difficult in this specific case, the ortholog hit ratios distribution (Figure 7) shows that approximately 50% contigs were assembled to a length corresponding to >75% of their ortholog.

On the basis of the homologies identified with BLAST, Gene Ontology (GO) terms could be assigned to 17,738 contigs (45%). More in detail 5,634 were mapped to a cellular component, 12,290 to a biological process and 14,625 to a molecular function. The complete summary of GO mappings is shown in Figure 8. Most transcripts resulted to be located within the cell, while only a minority resulted to be located elsewhere (mostly in organelles, macromolecular complexes or in the extracellular environment). The predominant molecular function was, by far "binding", with catalytic activity as the second most abundant GO term, reflecting the high enzymatic activity of the digestive gland. The scenario concerning biological processes was more variegated, with "cellular processes" and "metabolic processes" being the two most abundant GO terms, reflecting once again the high metabolic activity of digestive gland. This found further confirmation while exploring the level 3 GO terms, as "primary metabolic process", "cellular metabolic process" and "macromolecule metabolic process" resulted to be the three most abundant terms (see Table S1).

InterPro domains could be assigned to 17,726 contigs (45% out of the total). The most abundant Interpro domains are shown in Table 4. Consistently with the molecular function GO assignements, several of the most abundant domains (i.e. immunoglobulin-like, ankyrin, C1q, etc.) are characterized by marked binding properties.

Comparison with Mytibase

The *M. galloprovincialis* sequence data generated by RNA-seq in the present study can clearly only provide a limited view of the entire complement of transcripts expressed in the different tissues, different life stages and in response to different environmental variables and biotic\abiotic stimuli in this organism. Although the digestive gland has already been reported as a tissue characterized by

the expression of a broad spectrum of transcripts because of its active involvement in homeostatic mechanisms, the overlap with transcriptomes of other tissues is quite low [64] and therefore a significant proportion of genes is expected to be almost exclusively expressed in other tissues.
Mytibase includes 24,937 ESTs (assembled into 7,112 non-redundant sequences), obtained by the Sanger sequencing of many different tissues (namely digestive gland, haemocytes, gills, foot, anterior and posterior abductor muscle and mantle) [72]. Despite the low sequencing depth applied in respect with the present study, it is likely to include a certain amount of sequences which are expressed at all in the digestive gland and which therefore could not be obtained by RNA-seq.
In fact, not surprisingly, only about 30% of Mytibase sequences were included in the final "high quality set" of transcripts expressed in the digestive gland. A significant number of Mytibase sequences resulted to be expressed in this tissue at rather low levels and were discarded during the contig processing step and therefore were not included in the final transcripts set. Despite the high sequencing depth used, about 2,000 Mytibase transcripts were not found expressed at all in *M. galloprovincialis* digestive gland and may represent genes whose expression is strictly regulated and extremely specific in certain tissues.
The comparison between the functional domain relative abundance within the two sequence datasets revealed a sensible enrichment of many of the most common Interpro signatures in Metazoans (i.e. immunoglobulin-like, zinc-finger C2H2, etc.), most closely approaching the expected frequencies expected from a complete transcriptome (see Table 4). This can be explained by the advantages provided by next generation sequencing technologies over classic Sanger sequencing. Namely, the higher sequencing depth offers the possibility to obtain the sequences of transcripts expressed at very low levels, which would be difficult to observe in libraries comprising a few thousands of ESTs. Furthermore, the unbiased nature of RNA-seq (in respect with the 5'-biased nature of EST sequencing), guarantees the assembly of a higher proportion of full length transcripts (see Figure 1) [91], which can be most effectively and comprehensively annotated.
On the contrary, within the digestive gland transcriptome, Interpro signatures closely associated to immununity-related functions, such as C1q, C-type lectin-like and fibrinogen C-terminal globular domain didn't show a proportional increase in respect with the relative abundance observed in Mytibase. This is consistent with the previous observations of Mytibase being a valuable source of immune- and defense- related transcripts [92], as it includes a high proportion of ESTs produced from the sequencing of libraries constructed from haemolymph of immune-stimulated mussels [72].
The noteworthy difference between the number of total sequences contained in Mytibase and in the newly obtained digestive gland transcriptome clearly indicates that the newly obtained *M. galloprovincialis* digestive gland transcriptome integrates and sensibly enriches the previous version of Mytibase and provides the basis for future extensive studies on mussel on a whole-transcriptome scale [93], including the RNAseq analysis applied in the specific case of the present experiment. Nevertheless, despite the high number of expressed sequences assembled, further RNA-sequencing from additional tissues would be necessary to obtain a complete overview of the entire complement of transcripts expressed by mussel during its life cycle, as highlighted by the incomplete overlap between Mytibase and the new digestive gland transcripts database.

Conclusion

This study provides the first comprehensive analysis of the effects of PSTs accumulation in a molluscan species on a transcriptomic level. Although the effects of PSP on mussels and other mollusks had already been extensively investigated from the physiological, behavioral and histo-pathological point of views, the issue had never been addressed from a molecular perspective before.

The analysis of the transcriptional profiles revealed that PSP did not affect mussels, at least not at the concentrations reached in the experiment (about 300 μg STX eq 100 g^{-1} of meat), which were well above the consented limit for human consumption. Most of the previous studies classified Mytilids as organisms not responsive to PSP [55] or just observed a mild early response followed by an extremely rapid acclimatization [78, 94, 95] (even though this response could be merely related to the adaptation to a different alimentation regime). Nevertheless, there are occasional reports of increased mortality linked to PSP in mussels [57] and of severe histo-pathological modifications [59] following the ingestion of PTS producing dinoflagellates. These reports of mussels being seriously compromised by PSP did not find any confirmation in our experiment. Therefore our study provided the first molecular lines of evidence supporting the classification of mussels as organisms not responsive to PSP.

The identification of molecular markers typical of PSP could provide the basis for straightforward studies aimed at the development of tools for the biomonitoring of PSP contamination. In particular, the identification of alternative methods is a priority for the monitoring authorities, in order to replace the unreliable mouse bioassay and support the HPLC-based methods [30] and for the shellfish aquaculture sector, in order to be able to adopt strategies to minimize the possibility of PSP contamination [96]. Nevertheless, given the virtually null responsiveness of mussels evidenced by our study, we argue that the possibility of identifying PSP molecular markers in this organism is extremely unlikely. Such a task will be probably easier in responsive bivalves, such as oysters and clams, where the remarkable physiological modifications observed are likely matched by evident alteration of gene expression.

In addition to clarifying to which extent mussels are affected by PSP, our study, thanks to the massive NGS data obtained, provided extremely valuable resources which were used for the de novo assembly of the *M. galloprovincialis* digestive gland transcriptome. RNA deep sequencing had already been applied to a few bivalve mollusks species [61-64, 68], but this is the first Illumina technology-based sequencing effort aver reported in the Bivalvia class. The newly annotated sequence set remarkably improved the existing sequence database Mytibase and will certainly provide an important resource for improving the molecular knowledge of this species and have already been used as a basis for studies requiring whole-transcriptome mining approaches [93].

Acknowledgements

This work was supported by Regione Friuli Venezia Giulia, Direzione Centrale Risorse Agricole, Naturali, Forestali e Montagna, L.R. 26/2005 prot. RAF/9/7.15/47174.

Methods

Mussel specimens

The *Mytilus galloprovincialis* (Lamarck, 1819) specimens used in the experiments were obtained from a commercial producer from the Gulf of Trieste. All the mussels were collected from the same location. Individuals of similar size and weight (medium length 55 ±4 mm, mean fresh weight 2,48 ± 0,42 g) were acclimated for one week in running prefiltered seawater and for 3 days in bacteria-free filtered seawater (Millipore Durapore GV 0,22 μm, hydrophile PVDF) in experimental conditions. Two pools of 3 mussels were tested by HPLC before the start of the experiment and were found free of PSP toxins.

Alexandrium minutum cultures

The *A. minutum* strains AL1T (non-toxigenic) and AL9T (toxin producer), previously isolated from the Gulf of Trieste, were cultured in medium B [97] in a sufficient number of aerated 1 L batch cultures. The cultures were maintained at 15°C at a 10:14 h dark:light cycle with an irradiance of 60 μE m-2 s-1. Algae were harvested in the late exponential phase of growth.
Both strains was tested for toxicity during the course of the experiment, using the method described in the toxins analysis paragraph: at each of the two time points sampled for RNA-seq (T1 to T4), 100 ml of culture were filtered on Millipore Durapore GV 0,22 μm filters and immediately frozen at -18°C for HPLC. The toxic profile of the AL9T strain is shown in Figure 9.

Experimental design

The mussels were incubated at 15°C at a 12:12 h dark:light cycle in aquariums containing 0.4L of 0,22 μm filtered seawater per mussel. The water in the aquariums was substituted with fresh bacteria-free filtered water every morning at 9 AM. Six sets of aquariums were prepared for the incubations. Three out of these aquariums were incubated with the AL1T (non-toxigenic) and three with AL9T (toxigenic) *A. minutum* strain. For 5 days $2x10^6$ cells of *A. minutum* per mussel were added every 2 hours, 5 times a day, beginning at 10 AM. The mussels were further maintained in the aquariums with regular changes of the water but without food supply for other six days. At each experimental time point, at 9.00 AM one mussel per aquarium was sacrificed for further analyses. The selected experimental points were T1 (24 hours), T2 (48 hours) T3 (3 days), T4 (5 days), T5 (2 days into the detoxification phase), T6 (4 days into the detoxification phase) and T7 (6 days into the detoxification phase). At T0, before the feeding, one mussel per aquarium was sacrificed in order to provide the T0 control samples for the real-time PCR. A schematic representation of the experiment, summarizing the analysis performed at each time point for each group, is showed in Table 1.
Three additional aquariums were kept as a “standard diet” controls. Mussels were regularly fed once a day, with 36 mg marine invertebrate feed (Brightwell Reef Snow) per animal. Mussels were sacrificed at T0 to provide two biological replicates as the control material for the RNA-seq analysis.

Toxin analysis

The analysis of the PSP overall toxicity was undertaken on *A. minutum* samples and on shellfish during the feeding experiments (specifically at T0 before the first feeding with *Alexandrium* and at T4, where the maximum accumulation of toxins was supposedly achieved). The method, based on AOAC 2005.06 [98], is based on a pre-column oxidation High Performance Liquid Chromatography with Fluorescence Detection (HPLC-FLD).

The algal pellet was suspended in 0.1 mM acetic acid, with the total volume made up to 3 mL. The acidic algal suspension was transferred to a 50 mL centrifuge tube and the contents were subjected to probe sonication (sonicator Ultrasonic® Liquid Processor Model XL2020, Heat Systems Inc.) for 30 min, in order to break the algal cells. Sonicated algal suspensions were centrifuged (10 min 4500 rpm) and aliquots subjected to analysis.

Following the feeding experiments, mussel tissues, previously deprived of the digestive gland (used for RNA extraction), were homogenised and 1,7 g of tissue were analyzed. Through the use of both a peroxide and a periodate oxidation step, the method allows quantitation of individual PSP toxins, with the exception of the epimeric pairs (GTX1\4; GTX2\3, and C1\2) which form identical oxidation products and cannot be separated [99]. Toxins were quantified against linear calibrations of all currently-available PSP toxin certified reference standards and toxicity equivalence factors (TEFs) proposed by the CONTAM Panel [30] were used to calculate STX equivalence concentrations. HPLC-FLD analysis of mussel tissues without digestive glands enabled the quantitation of PSP toxin concentrations.

RNA extraction and analysis

Digestive glands were excised from 1 mussels per aquarium at each of the selected time points (see Table 1) and immediately homogenized in TRIzol® reagent (Life Technologies, Carlsbad, California). Total RNA was individually purified according to the manufacturer's instructions. Following extraction, the RNA quality was assessed by electrophoresis on denaturing agarose gel and its quantity was estimated by UV-spectrophotometry. Complementary DNA from the mussel samples fed with the AL9T (toxic) and with the AL1T (non-toxic) *A. minutum* strains was prepared by retro-transcription with the iScript™cDNA Synthesis Kit (Bio-Rad) to be used for Real Time quantitative PCR.

RNA extracted from the three individual mussels sampled at each experimental time point from the two groups (the AL1T and AL9T *A. minutum* strains-fed mussels) and from the control aquariums at T0 were pooled in equal quantities and used for the RNA-seq analysis.

Sequencing and de novo transcriptome assembly

cDNA libraries were prepared and subjected to massive sequencing analysis at the Biotechnology Center of the University of Illinois, using an Illumina GAII sequencing platform.

The obtained sequencing reads were further processed for adapter removal and trimmed according to base calling quality. The resulting 57,108,418 sequences were assembled with the CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark), assuming a paired reads distance comprised between 100 and 350 base pairs, and setting the penalties for mismatches, insertions and deletions at 3 and the length fraction and similarity to 0,5 and 0,9, respectively. To increase the overall quality of the assembly, the pre-existing Sanger sequences included in Mytibase [72] were included in the assembly (mismatch, insertion and deletion cost set at 3\3\2 and length fraction and similarity

at 0,2\0,9), together with 115,557 454 Life Sciences sequencing reads obtained from different tissues (gap, insertion and mismatch penalties set at 2\2\2 and length fraction\similarity at 0,4\0,8) and 15,058,580 additional Illumina reads obtained from the digestive gland (using the same parameters previously stated for Illumina sequencing reads). The minimum allowed assembled contig length was set at 250 base pairs.

The resulting contigs were filtered to eliminate sequences originated from ingested *A. minutum* cells and contamination from symbiont and parasytic bacteria as follows: all contigs were subject to BLASTx searches against both the Metazoa and Viridiplantae + Bacteria subsets of UniprotKB sequences. Contigs achieving a higher BLAST e-value in the latter selection by a 10^{-5} factor were discarded as probable contaminants. Furthermore, a BLASTn analysis was conducted against an assembly of the *Alexandrium catenella* ESTs obtained by Toulza et al. [100] and contigs showing identity to this assembly were considered as originating from the ingestion of *A. minutum* cells and also discarded. Contigs created by a very low number of sequencing reads (less than 50) were considered as not relevant to the digestive gland and therefore were not included in the final "high quality" set of sequences. Finally, all the transcripts without an open reading frame of at least 50 codons, the possible result of longer transcripts fragmentation or misassembly, were discarded before the annotation set.

Transcripts annotation

The BLASTx algorithm [101] was used to determine the contigs homology to known sequences, with an e-value cut-off of 10^{-6}. The NCBI non-redundant protein database was used for BLAST. The annotation was performed with Blast2GO [102], and Gene Ontologies mapping and InterPro terms [103] annotation were performed using the default settings. The Gene Ontology mapping were used to generate graphs summarizing Biological Process, Molecular Function and Cellular Component annotations at Level 2. BLASTx was also used to determine homologies with the only molluscan species whose genome has been fully sequenced to date, Lottia gigantea, by using the predicted protein models.

Ortholog hit ratios were calculated as suggested by O'Neil et al. [90], based on the BLASTx output, analyzing only the 1,000 contigs with the best e-values, in order to select conserved orthologs, balancing the evolutionary divergence and the low representation of molluscan sequences.

Expression analysis by RNA-seq

The filtered contigs were further processed to generate the reference set for the RNAseq mapping of sequencing reads originated from each of the six analyzed samples (T1 toxic, T1 non-toxic, T4 toxic, T4 non-toxic, control 1 and control 2). Contigs with a global coverage lower than 5 (calculated with the mapping of all the sequencing reads from all the six samples) were discarded prior to the analysis, as they could be subject to random expression fluctuations [82].

Raw count data from the six samples were used in the statistical analysis to identify differentially expressed transcripts with DEB [83], which simultaneously analyzes data with edgeR [84] and DEseq [82]. The analysis was aimed at the identification of differentially expressed genes in response to PSP, independently from the mussel diet. In this case the two "standard diet" groups and the groups T1 non-toxic and T4 non-toxic were considered as controls, whereas the two groups T1 toxic and T4 toxic were considered as treated samples. Differential expression was concluded

with a FDR (False Discovery Rate) lower than 0.01. Only contigs identified as differentially expressed by both edgeR and DEseq were selected for further analysis.

Quantitative PCR expression analysis

Four transcripts among those identified as differentially regulated in response to PSP contamination by both edgeR and DEseq were selected for the expression analysis via real-time quantitative PCR. Namely, "transcript of unknown function" I and II, C1q domain-containing protein I and hemicentin 1-like were chosen. The complete list of primers used for the quantitative PCR analysis is provided in Table 5. Expression levels were monitored at all the available experimental time points (see Table 1) in the digestive gland samples from the AL1T and AL9T strain-fed mussels.

All the PCR assays were performed using a Bio-Rad CFX96 system. The 15 µL reaction mix included 7,5 µL of 2x IQ ™ SYBR Green ® Supermix (Biorad), 0,3 µL of each 10 µM primer and 2 µl of a 1:10 cDNA dilution. The following thermal profile was used: an initial 3' denaturation step at 95°C, followed by 40 cycles at 95° for 20", 60° for 15" and 72° for 20". Amplification products were analyzed with a 65°/95°C melting curve.

The expression levels of the selected transcripts were determined using the comparative Ct method (2-ΔΔ Ct method) [104]. Ct values used for quantification were corrected based on PCR efficiencies using LinRegPCR [105]. The expression values were normalized using the elongation factor EF-1 as housekeeping gene (EF-1 primers are shown in Table 5) and scaled on the expression level observed at T0. Results are given as the mean with standard deviation of three technical replicates.

TABLES

day	experimental time point	feeding with the toxigenic strain AL1T	RNA-seq*	Real Time PCR*	phase
0	T0	x		x	accumulation phase
1	T1	x	x	x	
2	T2	x		x	
3	T3	x		x	
4		x			
5	T4		x	x	
6					detoxification phase
7	T5			x	
8					
9	T6			x	
10					
11	T7			x	

Table 1: Experiment summary. The analysis were performed on two aquariums, differing from the mussel feeding (*A. minutum* toxigenic strain AL1T vs non-toxigenic strain AL9T) *1 mussel per aquarium was sacrificed at each time point before the feeding and the analysis were performed on pools of 3 individuals.

Trimming statistics	
Number of reads before trimming	79,595,897
Number of reads after trimming	74,341,390
Paired reads after trimming	74,341,390
Single reads after trimming	129,003
Sequences discarded during trimming	5,125,504 (6,43%)
Average length before trimming	95,24 bp
Average lenth after trimming	97,75 bp
Number of reads per sample	
T1 non-toxic	6,104,184
T1 toxic	14,574,893
T4 non-toxic	16,423,414
T4 toxic	15,682,924
control 1	12,996,171
control 2	8,688,807
Assembly statistics	
Assembly size	27,094,215 bp
Total number of contigs	39,289
N50	814 bp
N75	494 bp
N90	349 bp
Mean contig length	689 bp
Median contig length	510 bp
Longest contig	14,211 bp
Number of contigs longer than 1 Kb	6,545
GC content	37,42%

Table 2: Trimming statistics, total number of sequencing reads obtained per sample and assembly summary.

Transcript name	edgeR FDR	DEseq FDR	expression level (RPKM)					
			control 1	control 2	T1 non-toxic	T4 non-toxic	T1 toxic	T4 toxic
unknown (I)*	4.72e-06	1.66e-05	0	0	0	0	37,71	25,45
unknown (II)*	2.20e-04	7.41e-05	0	0	0	0	18,36	19,82
unknown (III)	2.10e-04	7.39e-05	4,46	0	0	0	103,18	173,85
unknown (IV)	5.96e-03	6.44e-03	0	0	0	0	0	24,81
unknown (V)	5.03e-09	7.14e-05	0	10,3	0	0	4760,72	41,61
unknown (VI)	9.34e-03	4.58e-06	573,12	322,37	79,13	3,25	1,01	0,65
unknown (VII)	2.63e-04	1.40e-04	0	0	0	0	22,33	24,11
C1q domain containing protein (I)*	3.69e-03	9.79e-03	0	0	2,1	0	33,37	25,26
C1q domain containing protein (II)	2.06e-03	1.68e-04	3,05	1,74	0	0	34,18	100,26
DED domain containing protein	3.00e-04	1.15e-04	2,05	2,34	0	0	79,65	128,92
IMAP family GTPase	3.00e-04	2.10e-04	0	2,65	0	0	82,42	47,86
C-type lectin	2.62e-03	3.44e-03	0	0	0	9,9	27,36	175,63
zona pellucida binding domain containing protein	3.33e-03	4.21e-07	431,05	571,68	0	0	0	0
hemicentin 1-like*	8.01e-06	3.55e-05	0	0	0	0	53,38	13,19

Table 3: List of differentially expressed genes according to the DEB analysis. FDR values calculated by edgeR and DEseq are shown, as well as the expression values calculated with the RNA-seq analysis. *These genes were also selected for validation via real-time quantitative PCR.

Interpro domain	Description	Digestive gland contigs	Mytibase contigs	Rate*
IPR013783	Immunoglobulin-like fold	777	46	3,05
IPR011042	Six-bladed beta-propeller, TolB-like	710	25	5,12
IPR020683	Ankyrin repeat-containing domain	676	43	2,84
IPR008983	Tumour necrosis factor-like	624	145	0,78
IPR001073	Complement C1q protein	580	140	0,75
IPR002110	Ankyrin repeat	558	39	2,58
IPR007110	Immunoglobulin-like	453	33	2,48
IPR000315	Zinc finger, B-box	392	46	1,54
IPR007087	Zinc finger, C2H2	389	17	4,13
IPR015943	WD40/YVTN repeat-like-containing domain	383	44	1,57
IPR013032	EGF-like region, conserved site	363	26	2,52
IPR000742	Epidermal growth factor-like, type 3	335	25	2,42
IPR015880	Zinc finger, C2H2-like	323	11	5,30
IPR013098	Immunoglobulin I-set	292	13	4,05
IPR013087	Zinc finger, C2H2-type/integrase, DNA-binding	288	8	6,50
IPR011009	Protein kinase-like domain	284	28	1,83
IPR006210	Epidermal growth factor-like	275	19	2,61
IPR006703	AIG1	265	23	2,08
IPR003599	Immunoglobulin subtype	261	15	3,14
IPR000719	Protein kinase, catalytic domain	250	24	1,88
IPR002181	Fibrinogen, alpha/beta/gamma chain, C-terminal globular	234	58	0,73
IPR016196	Major facilitator superfamily domain, general substrate transporter	218	7	5,62
IPR003961	Fibronectin, type III	217	10	3,92
IPR003598	Immunoglobulin subtype 2	212	15	2,55
IPR013083	Zinc finger, RING/FYVE/PHD-type	208	38	0,99
IPR014716	Fibrinogen, alpha/beta/gamma chain, C-terminal globular, subdomain 1	207	52	0,72
IPR016187	C-type lectin fold	203	106	0,35
IPR016186	C-type lectin-like	202	106	0,34

Table 4: most abundant IPR domains in the digestive gland assembly according to the Interproscan assignments. *This value represents the rate between the number of contigs observed and the number of contigs expected based on Mytibase abundances. A rate > 1 means an enrichment in the digestive gland transcriptome, whereas a rate <1 means an over-representation of the domain in Mytibase.

transcript name	FOR primer	REV primer
EF-1	cctcccaccatcaagaccta	ggctggagcaaaggtaacaa
transcript of unknown function (I)	tcagcgtagcacctttacca	ccatctggcaaagccttact
transcript of unknown function (II)	acagcttgaaacggaccttc	tattcacgtgccttgtcctc
C1q domain containing protein (I)	gacaactcaaggcgcatgtt	ttccaaaggtagacccgtca
hemicentin 1-like	gagataccccagcacttcca	aaccaatgaggcatctggac

Table 5: Primers designed for the expression analysis via Real-Time quantitative PCR. EF-1 was used as a housekeeping gene for normalization.

GO-id	GO-term	numer of contigs
GO:0044238	primary metabolic process	2561
GO:0044237	cellular metabolic process	1961
GO:0043170	macromolecule metabolic process	1845
GO:0050789	regulation of biological process	1071
GO:0006807	nitrogen compound metabolic process	972
GO:0051234	establishment of localization	947
GO:0051716	cellular response to stimulus	945
GO:0007154	cell communication	908
GO:0009058	biosynthetic process	644
GO:0055114	oxidation-reduction process	567
GO:0044281	small molecule metabolic process	494
GO:0009056	catabolic process	266
GO:0007155	cell adhesion	251
GO:0006950	response to stress	166
GO:0033036	macromolecule localization	161
GO:0008219	cell death	155
GO:0051641	cellular localization	127
GO:0007017	microtubule-based process	104
GO:0006996	organelle organization	95
GO:0007049	cell cycle	82
GO:0006928	cellular component movement	80
GO:0022607	cellular component assembly	61
GO:0065008	regulation of biological quality	60
GO:0042221	response to chemical stimulus	60
GO:0019637	organophosphate metabolic process	56
GO:0065009	regulation of molecular function	54
GO:0006955	immune response	47
GO:0043933	macromolecular complex subunit organization	42
GO:0034621	cellular macromolecular complex subunit organization	39

Table S1: Top 30 most represented Gene Ontology (GO) at the level 3 of Biological Process.

FIGURES

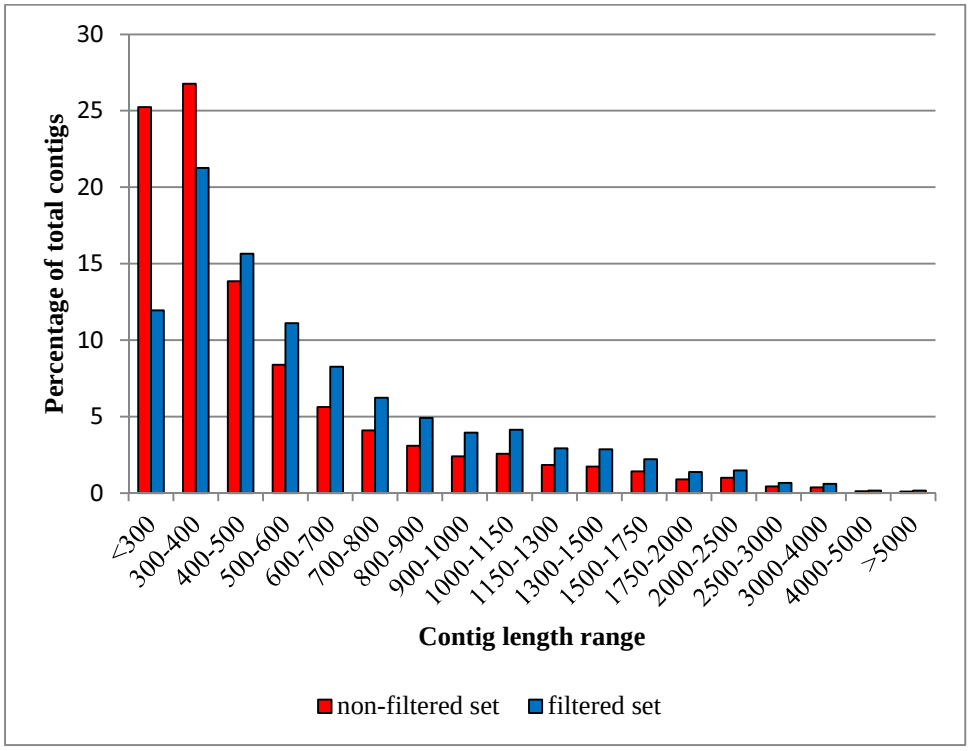

Figure 1: contig length distribution, before and after the filtering procedures. The graph highlights a shift towards higher length ranges in the high quality set, meaning a reduction of the bias towards short, low quality contigs, which are often the result of misassebly or fragmentation.

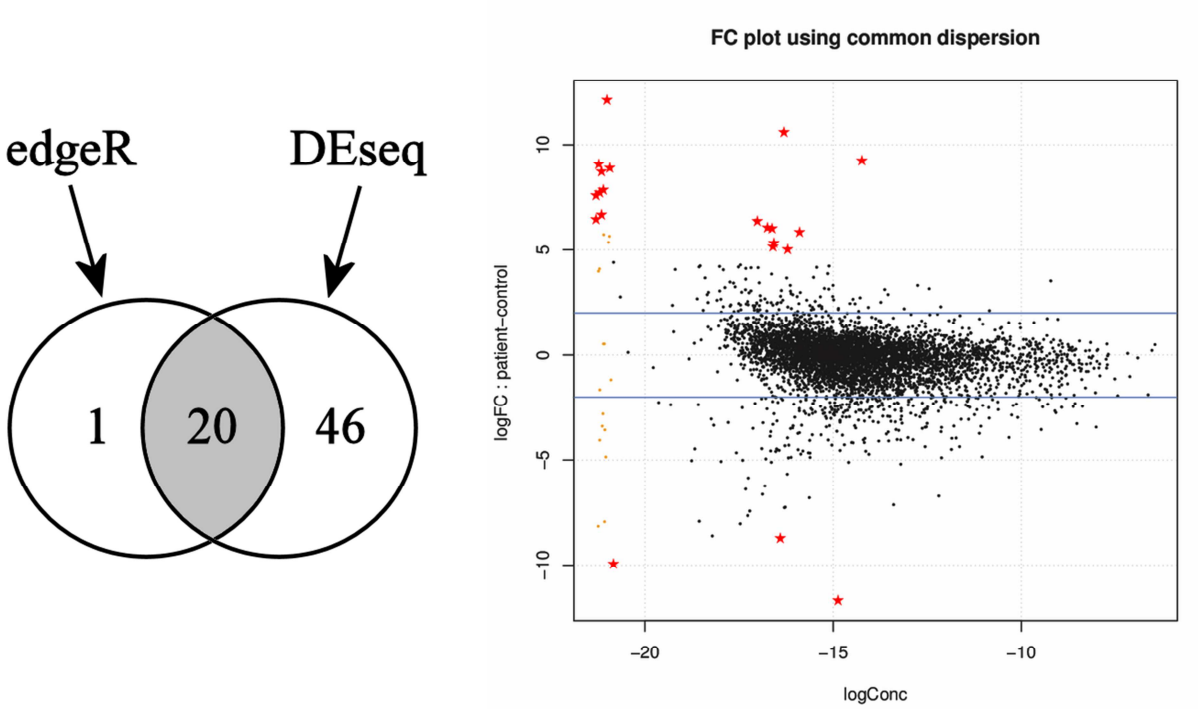

Figure 2: Schematic representation of the DEB analysis. On the left side, a Venn diagram showing the predicted differentially expressed genes overlaps by edgeR and DEseq using a 1% FDR cutoff; only the 20 genes predicted by both edgeR and DEseq were selected for further analysis. On the right, the MA plot generated by edgeR. Differentially expressed genes are indicated by a star.

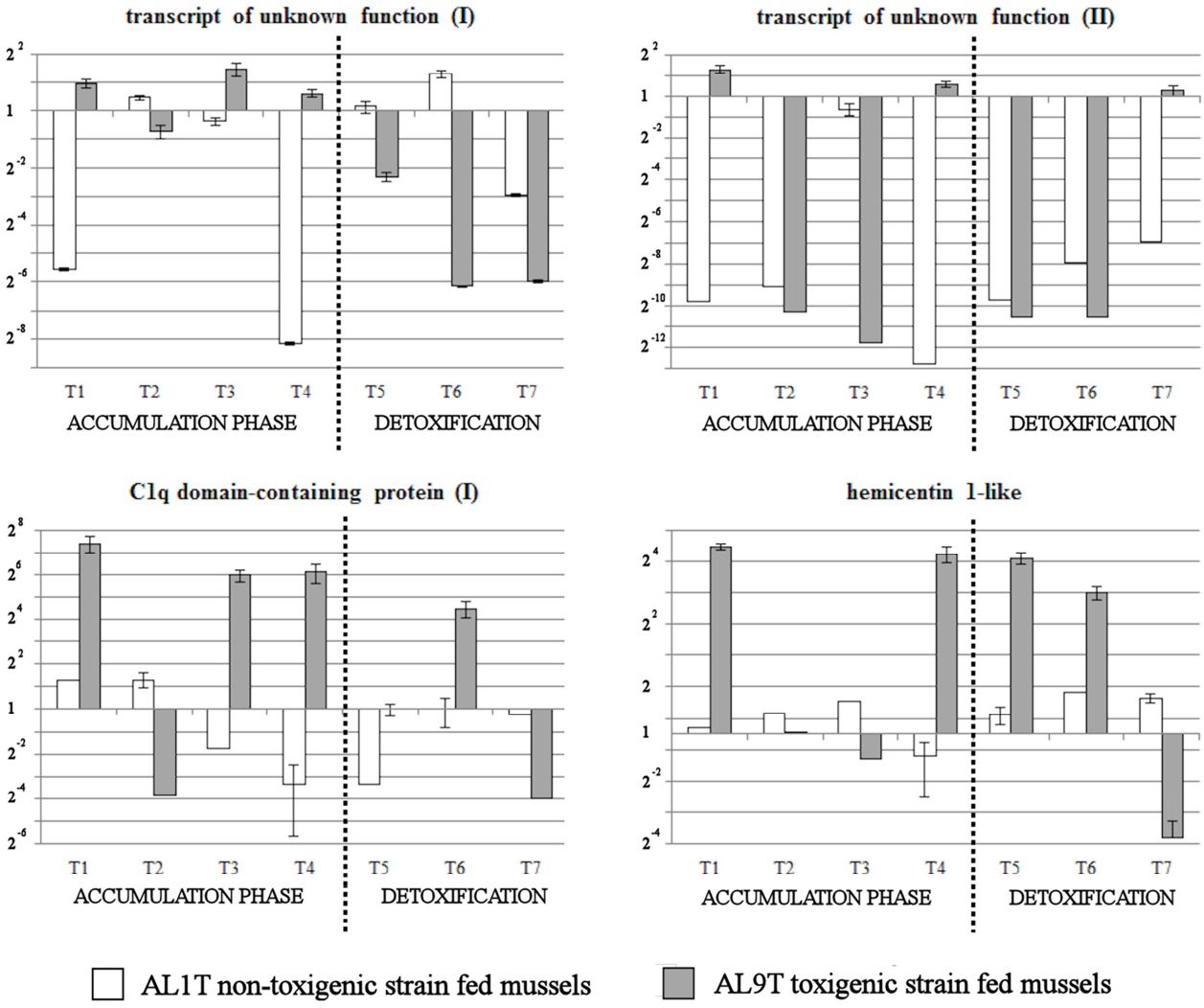

Figure 3: expression profiles obtained during the whole time course of the experiment by real-time quantitative PCR for the 4 selected genes identified as differentially expressed by the DEB statistical analysis. Expression values are scaled on the expression observed at T0.

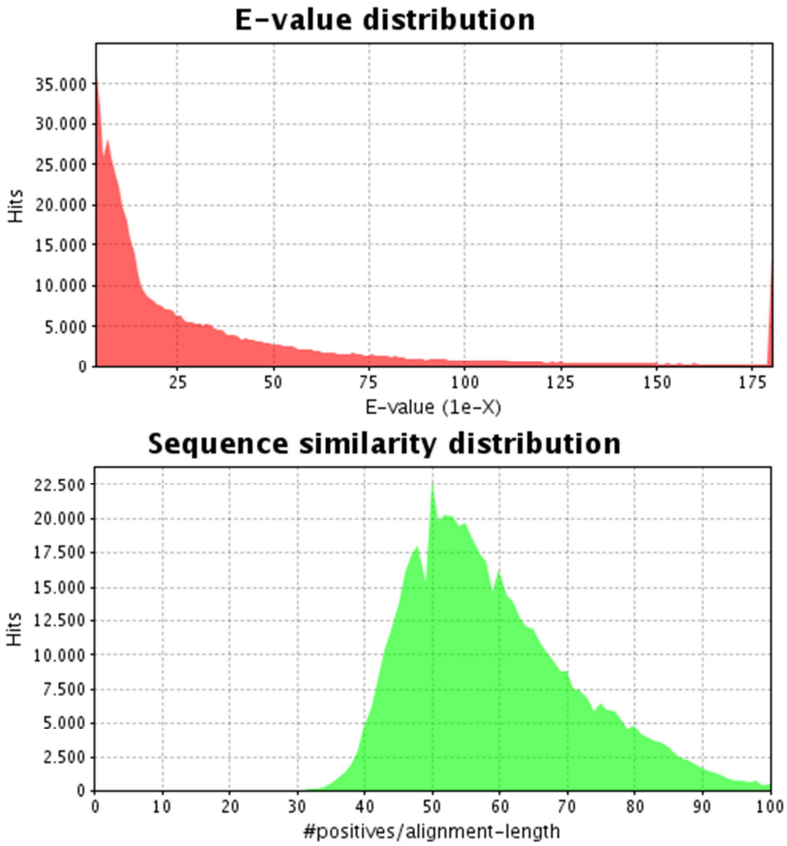

Figure 4: BLASTx e-value and sequence similarity distribution in the NCBI non redundant sequence database.

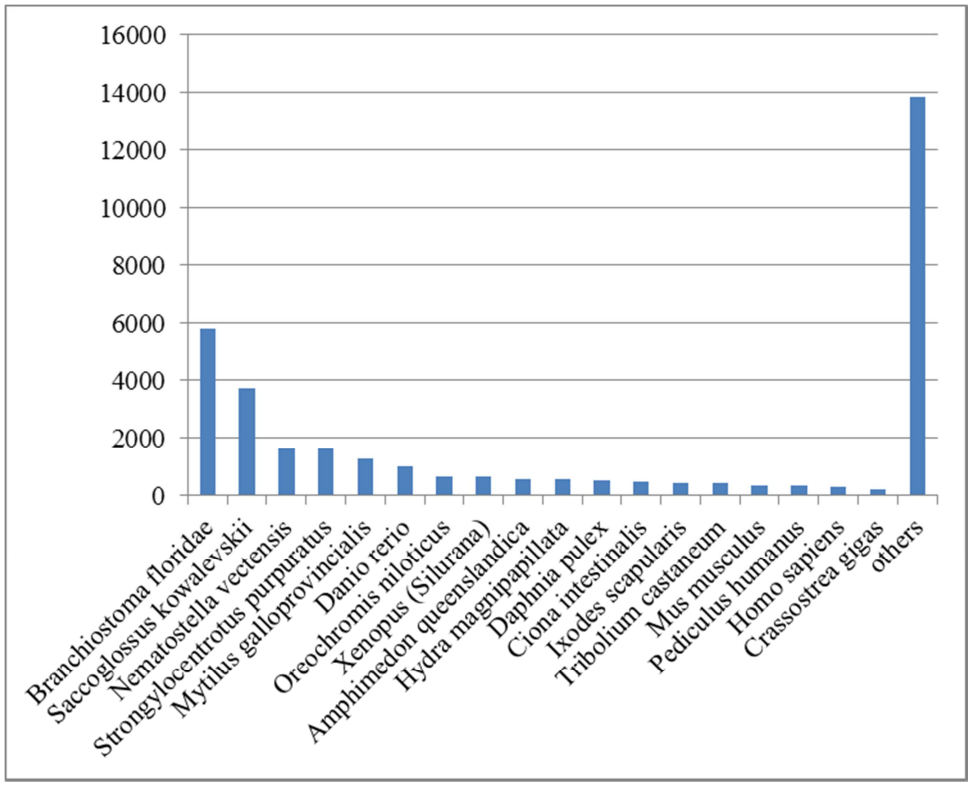

Figure 5: top hit BLAST species in the NCBI non redundant protein database.

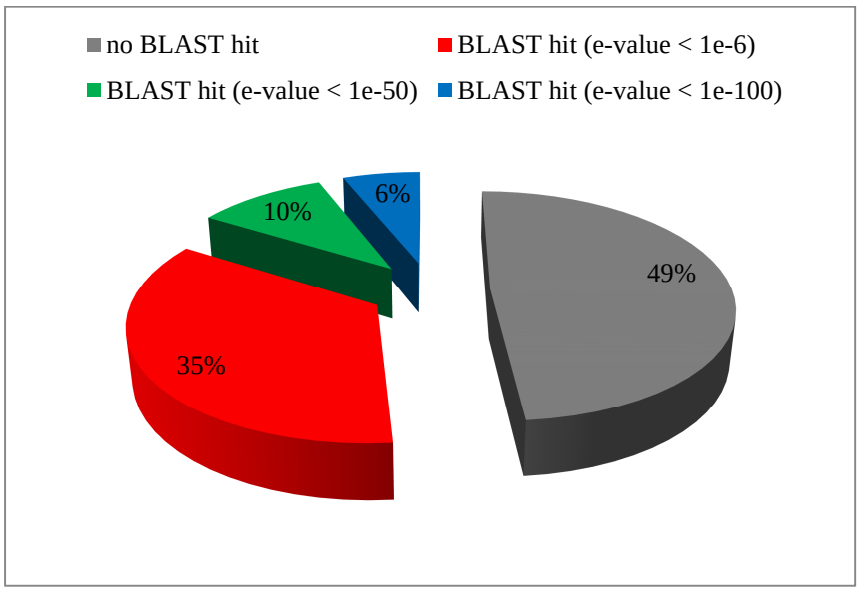

Figure 6: BLAST homologies of the *M. galloprovincialis* filtered contigs set to the *L. gigantea* proteins predicted from the genome assembly.

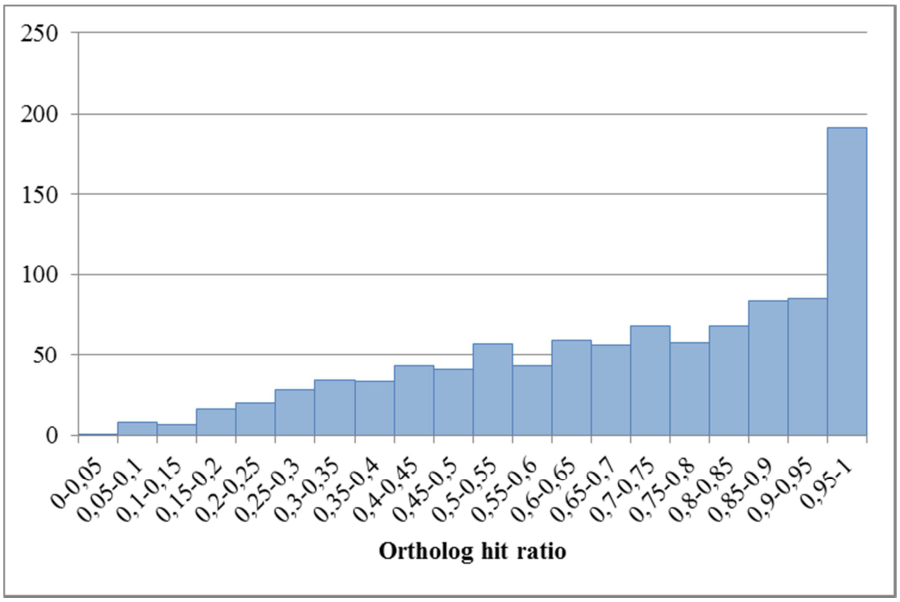

Figure 7: Distribution of ortholog hit ratios obtained from the BLASTx of the *M. galloprovincialis* filtered contig set vs the 1000 top BLAST hits in the nr NCBI protein database. An ortholog hit ratio of 1 means that a transcript has been likely assembled to its full length. Ratios >1 (indicating insertions) were collapsed within the 0,95-1 category.

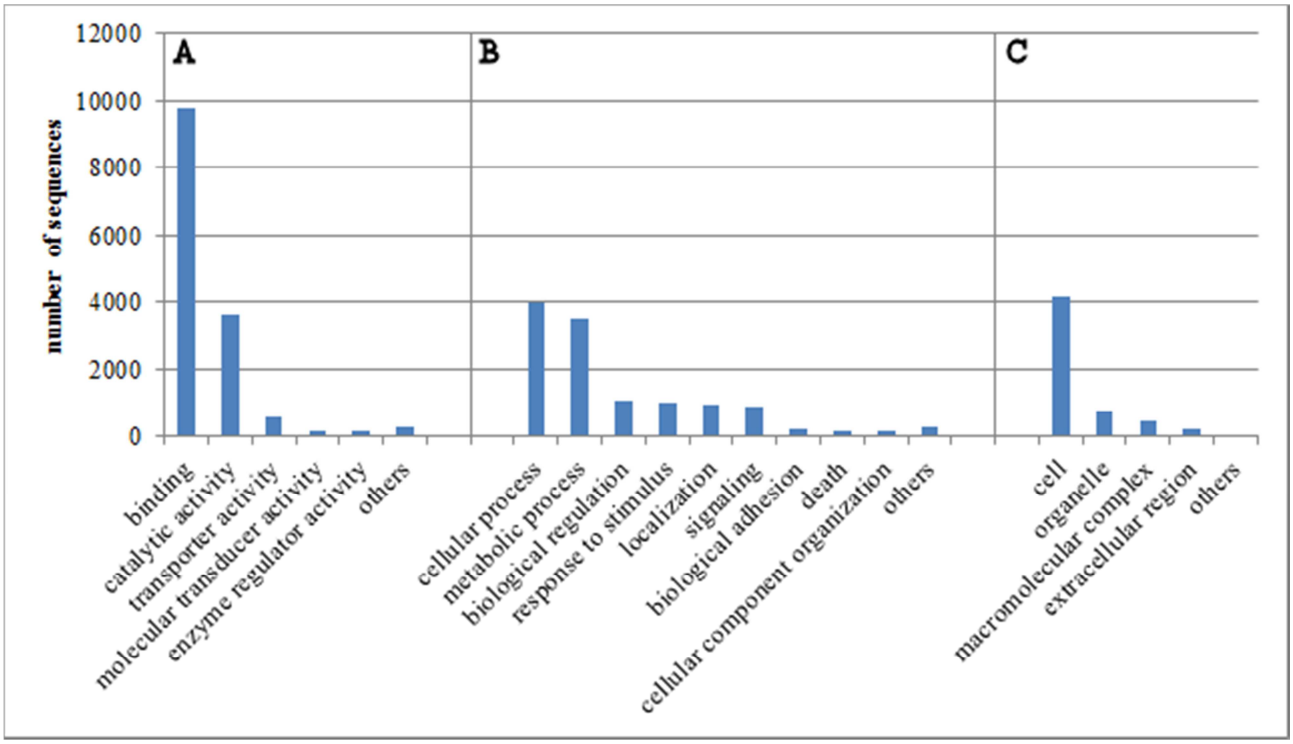

Figure 8: Gene Ontology (GO) terms assignments. A: molecular function. B: biological process. C: cellular component.

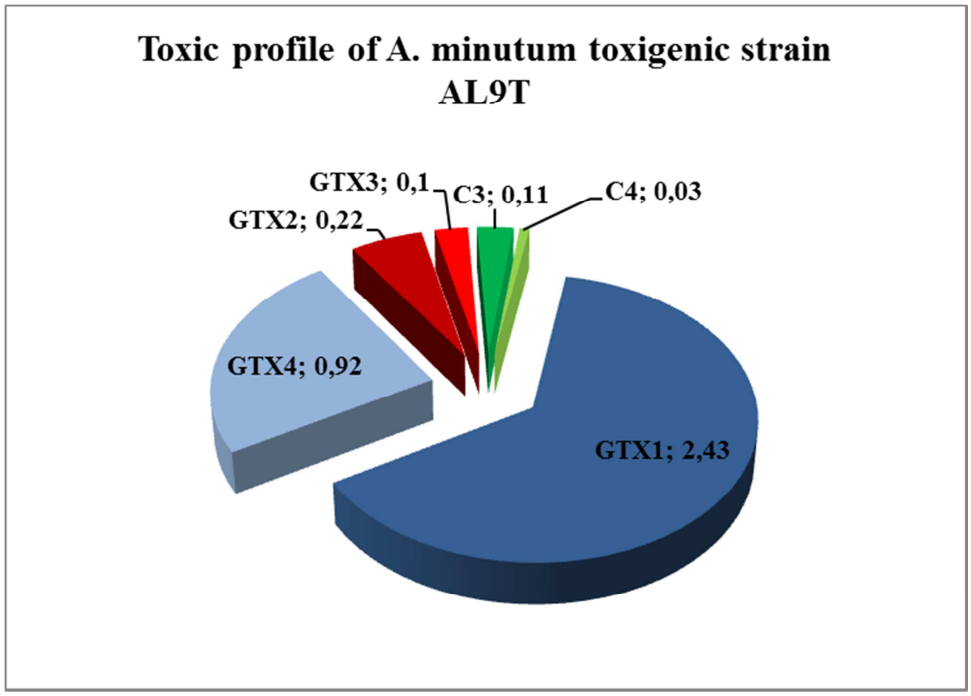

Figure 9: toxic profile of *A. minutum* AL9T strain, as determined by Jaime et al. (personal communication).

REFERENCES

1. Humpage AR, Rositano J, Bretag AH, Brown R, Baker PD, Nicholson BC, Steffensen DA: Paralytic shellfish poisons from Australian cyanobacterial blooms. AUSTJMARFRESHWATER RES 1994, 45(5):761-771.

2. Cembella AD, Sullivan JJ, Boyer GL, Taylor FJR, Andersen RJ: Variation in paralytic shellfish toxin composition within the *Protogonyaulax tamaronsis*/*catenella* species complex; red tide dinoflagellates. Biochemical Systematics and Ecology 1987, 15(2):171-186.

3. Hallegraeff GM, Steffensen DA, Wetherbee R: Three estuarine Australian dinoflagellates that can produce paralytic shellfish toxins. J PLANKTON RES 1988, 10(3):533.

4. Kodama M, Ogata T, Fukuyo Y, Ishimaru T, Wisessang S, Saitanu K, Panichyakarn V, Piyakarnchana T: *Protogonyaulax cohorticula*, a toxic dinoflagellate found in the Gulf of Thailand. Toxicon 1988, 26(8):707-712.

5. Schwinghamer P, Hawryluk M, Powell C, MacKenzie CH: Resuspended hypnozygotes of *Alexandrium fundyense* associated with winter occurrence of PSP in inshore Newfoundland waters. Aquaculture 1994, 122(2-3):171-179.

6. Hansen PJ, Cembella AD, Moestrup O: The marine dinoflagellate *Alexandrium ostenfeldii*: paralytic shellfish toxin concentration, composition, and toxicity to a tintinnid ciliate. Journal of Phycology 1992, 28(5):597-603.

7. Mee LD, Espinosa M, Diaz G: Paralytic shellfish poisoning with a *Gymnodinium catenatum* red tide on the Pacific coast of Mexico. Marine Environmental Research 1986, 19(1):77-92.

8. Gacutan RQ, Tabbu MY, Aujero EJ, Icatlo Jr F: Paralytic shellfish poisoning due to *Pyrodinium bahamense* var. compressa in Mati, Davao oriental, Philippines. Marine Biology 1985, 87(3):223-227.

9. Terlau H, Heinemann SH, Stuhmer W, Pusch M, Conti F, Imoto K, Numa S: Mapping the site of block by tetrodotoxin and saxitoxin of sodium channel II. FEBS Letters 1991, 293(1-2):93-96.

10. Narahashi T, Haas HG, Therrien EF: Saxitoxin and tetrodotoxin: comparison of nerve blocking mechanism. Science 1967, 157(795):1441-1442.

11. James KJ, Carey B, O'Halloran J, Van Pelt FNAM, Škrabáková Z: Shellfish toxicity: Human health implications of marine algal toxins. Epidemiology and Infection 2010, 138(7):927-940.

12. García C, Bravo MDC, Lagos M, Lagos N: Paralytic shellfish poisoning: Post-mortem analysis of tissue and body fluid samples from human victims in the Patagonia fjords. Toxicon 2004, 43(2):149-158.

13. Hashimoto T, Matsuoka S, Yoshimatsu SA, Miki K, Nishibori N, Nishio S, Noguchi T: First paralytic shellfish poison (PSP) infestation of bivalves due to toxic dinoflagellate *Alexandrium*

tamiyavanichii, in the southeast coasts of the Seto Inland Sea, Japan. Journal of the Food Hygienic Society of Japan 2002, 43(1):1-5.

14. Okumura M, Yamada S, Oshima Y, Ishikawa N: Characteristics of paralytic shellfish poisoning toxins derived from short-necked clams (*Tapes japonica*) in Mikawa Bay. Natural Toxins 1994, 2(3):141-143.

15. Takatani T, Morita T, Anami A, Akaeda H, Kamijo Y, Tsutsumi K, Noguchi T: Appearance of *Gymnodinium catenatum* in association with the toxification of bivalves in Kamae, Oita Prefecture, Japan. Journal of the Food Hygienic Society of Japan 1998, 39(4):275-280.

16. Gessner BD, Middaugh JP: Paralytic shellfish poisoning in Alaska: A 20-year retrospective analysis. American Journal of Epidemiology 1995, 141(8):766-770.

17. Jester RJ, Baugh KA, Lefebvre KA: Presence of *Alexandrium catenella* and paralytic shellfish toxins in finfish, shellfish and rock crabs in Monterey Bay, California, USA. Marine Biology 2009, 156(3):493-504.

18. Shumway SE, Sherman SA, Cembella AD, Selvin R: Accumulation of paralytic shellfish toxins by surfclams, *Spisula solidissima* (Dillwyn, 1897) in the Gulf of Maine: Seasonal changes, distribution between tissues, and notes on feeding habits. Natural Toxins 1994, 2(4):236-251.

19. Álvarez G, Uribe E, Vidal A, Ávalos P, González F, Mariño C, Blanco J: Paralytic shellfish toxins in *Argopecten purpuratus* and *Semimytilus algosus* from northern Chile. Aquatic Living Resources 2009, 22(3):341-347.

20. Montebruno D: Paralytic shellfish poisoning in Chile. Medicine, Science and the Law 1993, 33(3):243-246.

21. Ayres PA: Mussel poisoning in Britain with special reference to paralytic shellfish poisoning. A review of cases reported 1814-1968. ENVIRONMHLTH 1975, 83(7):261-265.

22. Anderson DM, Sullivan JJ, Reguera B: Paralytic shellfish poisoning in northwest Spain: The toxicity of the dinoflagellate *Gymnodinium catenatum*. Toxicon 1989, 27(6):665-674.

23. Bravo I, Reyero MI, Cacho E, Franco JM: Paralytic shellfish poisoning in *Haliotis tuberculata* from the Galician coast: Geographical distribution, toxicity by lengths and parts of the mollusc. Aquatic Toxicology 1999, 46(2):79-85.

24. Amzil Z, Quilliam MA, Hu T, Wright JLC: Winter accumulation of paralytic shellfish toxins in digestive glands of mussels from Arcachon and Toulon (France) without detectable toxic plankton species revealed by interference in the mouse bioassay for lipophilic toxins. Natural Toxins 1999, 7(6):271-277.

25. Lilly EL, Kulis DM, Gentien P, Anderson DM: Paralytic shellfish poisoning toxins in France linked to a human-introduced strain of *Alexandrium catenella* from the western Pacific: Evidence from DNA and toxin analysis. Journal of Plankton Research 2002, 24(5):443-452.

26. Honsell G, Poletti R, Pompei M, Sidari L, Milandri A, Casadei C, Viviani R: *Alexandrium minutum* Halim and PSP contamination in the northern Adriatic Sea (Mediterranean Sea). In: Harmful and Toxic Algal Blooms. Edited by Yasumoto T, Oshima Y, Fukuyo Y. Paris: Intergovernmental Oceanographic Commission of UNESCO; 1996: 77-80.

27. Ujević I, Roje R, Ninčević-Gladan T, Marasović I: First report of Paralytic Shellfish Poisoning (PSP) in mussels (*Mytilus galloprovincialis*) from eastern Adriatic Sea (Croatia). Food Control, 25(1):285-291.

28. Anderson MD, Hoagland P, Kaoru Y, White A: Estimated Annual Economic Impacts from Harmful Algal Blooms (HABs) in the United States. In.: Woods Hole Oceanographic Institution; 2000.

29. Conte FS: Economic impact of paralytic shellfish poison on the oyster industry in the Pacific United States. Aquaculture 1984, 39(1-4):331-343.

30. EFSA: Scientific Opinion of The Panel on Contaminants in The Food Chain on a request from the European Commission on Marine Biotoxins in Shellfish – Saxitoxin Group. The EFSA Journal 2009, 1019:1-76.

31. Coulson JC, Potts GR, Deans IR, Fraser SM: Dinoflagellate crop in the North Sea: Mortality of shags and other Sea Birds caused by paralytic shellfish poison. Nature 1968, 220(5162):23-24.

32. Geraci JR: Humpback whales (Megaptera novaeangliae) fatally poisoned by dinoflagellate toxin. Canadian Journal of Fisheries and Aquatic Sciences 1989, 46(11):1895-1898.

33. Kvitek RG, Degange AR, Beitler MK: Paralytic shellfish poisoning toxins mediate feeding behavior of sea otters. Limnology & Oceanography 1991, 36(2):393-404.

34. Cembella AD, Quilliam MA, Lewis NI, Bauder AG, Dell'Aversano C, Thomas K, Jellett J, Cusack RR: The toxigenic marine dinoflagellate *Alexandrium tamarense* as the probable cause of mortality of caged salmon in Nova Scotia. Harmful Algae 2002, 1(3):313-325.

35. Bricelj VM, Lee JH, Cembella AD: Influence of dinoflagellate cell toxicity on uptake and loss of paralytic shellfish toxins in the northern quahog *Mercenaria mercenaria*. Marine Ecology Progress Series 1991, 74(1):33-46.

36. MacQuarrie SP, Bricelj VM: Behavioral and physiological responses to PSP toxins in *Mya arenaria* populations in relation to previous exposure to red tides. Marine Ecology Progress Series 2008, 366:59-74.

37. Gainey Jr LF, Shumway SE: Physiological effects of *Protogonyaulax tamarensis* on cardiac activity in bivalve molluscs. Comparative Biochemistry and Physiology Part C, Comparative 1988, 91(1):159-164.

38. Basti L, Nagai K, Shimasaki Y, Oshima Y, Honjo T, Segawa S: Effects of the toxic dinoflagellate *Heterocapsa circularisquama* on the valve movement behaviour of the Manila clam *Ruditapes philippinarum*. Aquaculture 2009, 291(1-2):41-47.

39. Nagai K, Honjo T, Go J, Yamashita H, Seok Jin O: Detecting the shellfish killer *Heterocapsa circularisquama* (Dinophyceae) by measuring bivalve valve activity with a Hall element sensor. Aquaculture 2006, 255(1-4):395-401.

40. Tran D, Haberkorn H, Soudant P, Ciret P, Massabuau JC: Behavioral responses of *Crassostrea gigas* exposed to the harmful algae *Alexandrium minutum*. Aquaculture 2010, 298(3-4):338-345.

41. Haberkorn H, Tran D, Massabuau JC, Ciret P, Savar V, Soudant P: Relationship between valve activity, microalgae concentration in the water and toxin accumulation in the digestive gland of the Pacific oyster *Crassostrea gigas* exposed to *Alexandrium minutum*. Marine Pollution Bulletin 2011, 62(6):1191-1197.

42. Sagou R, Amanhir R, Taleb H, Vale P, Blaghen M, Loutfi M: Comparative study on differential accumulation of PSP toxins between cockle (*Acanthocardia tuberculatum*) and sweet clam (*Callista chione*). Toxicon 2005, 46(6):612-618.

43. Kitts DD, Smith DS, Beitler MK, Liston J: Presence of paralytic shellfish poisoning toxins and soluble proteins in toxic butter clams (*Saxidomus giganteus*). Biochemical and Biophysical Research Communications 1992, 184(1):511-517.

44. Takati N, Mountassif D, Taleb H, Lee K, Blaghen M: Purification and partial characterization of paralytic shellfish poison-binding protein from *Acanthocardia tuberculatum*. Toxicon 2007, 50(3):311-321.

45. Tian H, Gao C, Wang Z, Sun P, Fan S, Zhu M: Comparative study on in vitro transformation of paralytic shellfish poisoning (PSP) toxins in different shellfish tissues. Acta Oceanologica Sinica 2010, 29(1):120-126.

46. Oshima Y: Chemical and enzymatic transformation of paralytic shellfish toxins in marine organisms. In: Harmful Marine Algal Blooms. Edited by Lassus P, Arzul G, E. E. Paris: Lavoisier; 1995: 475–480.

47. Sullivan JJ, Iwaoka WT, Liston J: Enzymatic transformation of PSP toxins in the littleneck clam (*Protothaca staminea*). Biochem Biophys Res Commun 1983, 114(2):465-472.

48. Smith EA, Grant F, Ferguson CM, Gallacher S: Biotransformations of paralytic shellfish toxins by bacteria isolated from bivalve molluscs. Appl Environ Microbiol 2001, 67(5):2345-2353.

49. Donovan CJ, Ku JC, Quilliam MA, Gill TA: Bacterial degradation of paralytic shellfish toxins. Toxicon 2008, 52(1):91-100.

50. Twarog BM, Hidaka T, Yamaguchi H: Resistance to tetrodotoxin and saxitoxin in nerves of bivalve molluscs. A possible correlation with paralytic shellfish poisoning. Toxicon 1972, 10(3):273-278.

51. Twarog BM: Immunity to paralytic shellfish toxin in bivalve molluscs. In: Proc Int Symp Coral Reefs, 2nd: 1974; Brisbane, Australia: Great Barrier Reef Comm.; 1974: 505-512.

52. Bricelj VM, Connell L, Konoki K, MacQuarrie SP, Scheuer T, Catterall WA, Trainer VL: Sodium channel mutation leading to saxitoxin resistance in clams increases risk of PSP. Nature 2005, 434(7034):763-767.

53. Connell LB, MacQuarrie SP, Twarog BM, Iszard M, Bricelj VM: Population differences in nerve resistance to paralytic shellfish toxins in softshell clam, Mya arenaria, associated with sodium channel mutations. Marine Biology 2007, 150(6):1227-1236.

54. Bricelj VM, Shumway SE: Paralytic shellfish toxins in bivalve molluscs: Occurrence, transfer kinetics, and biotransformation. Reviews in Fisheries Science 1998, 6(4):315-383.

55. Bricelj VM, Lee JH, Cembella AD, Anderson DM: Uptake kinetics of paralytic shellfish toxins from the dinoflagellate *Alexandrium fundyense* in the mussel *Mytilus edulis*. Mar Ecol Prog Ser 1990, 63:117–188.

56. Marsden ID, Shumway SE: The effect of a toxic dinoflagellate (*Alexandrium tamarense*) on the oxygen uptake of juvenile filter-feeding bivalve molluscs. Comparative Biochemistry and Physiology Part A: Physiology 1993, 106(4):769-773.

57. Shumway SE, Cucci TL: The effects of the toxic dinoflagellate *Protogonyaulax tamarensis* on the feeding and behaviour of bivalve molluscs. Aquatic Toxicology 1987, 10(1):9-27.

58. Shumway SE, Pierce FC, Knowlton K: The effect of *Protogonyaulax Tamarensis* on byssus production in *Mytilus edulis* l., *Modiolus modiolus* linnaeus, 1758 and Geukensia demissa dillwyn. Comparative Biochemistry and Physiology Part A: Physiology 1987, 87(4):1021-1023.

59. Galimany E, Sunila I, Hégaret H, Ramón M, Wikfors GH: Experimental exposure of the blue mussel (*Mytilus edulis*, L.) to the toxic dinoflagellate *Alexandrium fundyense*: Histopathology, immune responses, and recovery. Harmful Algae 2008, 7(5):702-711.

60. Pérez-Enciso M, Ferretti L: Massive parallel sequencing in animal genetics: Wherefroms and wheretos. Animal genetics 2010, 41(6):561-569.

61. Milan M, Coppe A, Reinhardt R, Cancela L, Leite R, Saavedra C, Ciofi C, Chelazzi G, Patarnello T, Bortoluzzi S et al: Transcriptome sequencing and microarray development for the Manila clam, *Ruditapes philippinarum*: genomic tools for environmental monitoring. BMC Genomics 2011, 12(1):234.

62. Hou R, Bao Z, Wang S, Su H, Li Y, Du H, Hu J, Wang S, Hu X: Transcriptome Sequencing and De NovoAnalysis for Yesso Scallop (*Patinopecten yessoensis*) Using 454 GS FLX. PLoS ONE 2011, 6(6):e21560.

63. Clark M, Thorne M, Vieira F, Cardoso J, Power D, Peck L: Insights into shell deposition in the Antarctic bivalve *Laternula elliptica*: gene discovery in the mantle transcriptome using 454 pyrosequencing. BMC Genomics 2010, 11(1):362.

64. Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B, Hutchinson TH, Chipman JK: Pyrosequencing of *Mytilus galloprovincialis* cDNAs: Tissue-Specific Expression Patterns. PLoS ONE 2010, 5(1):e8875.

65. Bettencourt R, Pinheiro M, Egas C, Gomes P, Afonso M, Shank T, Santos R: High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*. BMC Genomics 2010, 11(1):559.

66. Joubert C, Piquemal D, Marie B, Manchon L, Pierrat F, Zanella-Cleon I, Cochennec-Laureau N, Gueguen Y, Montagnani C: Transcriptome and proteome analysis of Pinctada margaritifera calcifying mantle and shell: focus on biomineralization. BMC Genomics 2010, 11(1):613.

67. Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B, Hutchinson TH, Chipman JK: Pyrosequencing of *Mytilus galloprovincialis* cDNAs: Tissue-Specific Expression Patterns. PLoS ONE 2010, 5(1):e8875.

68. Gavery MR, Roberts SB: Characterizing short read sequencing for gene discovery and RNA-Seq analysis in *Crassostrea gigas*. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics (0).

69. Ghiselli F, Milani L, Chang PL, Hedgecock D, Davis JP, Nuzhdin SV, Passamonti M: De novo assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. Molecular Biology and Evolution 2012, 29(2):771-786.

70. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW: Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 2011, 480(7377):364-367.

71. Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B et al: Phylogenomics reveals deep molluscan relationships. Nature 2011, 477(7365):452-456.

72. Venier P, De Pitta C, Bernante F, Varotto L, De Nardi B, Bovo G, Roch P, Novoa B, Figueras A, Pallavicini A et al: MytiBase: a knowledgebase of mussel (*M. galloprovincialis*) transcribed sequences. BMC Genomics 2009, 10(1):72.

73. Delgado M, Estrada M, Camp J, Fernández JV, Santmartí M, Lletí C: Development of a toxic *Alexandrium minutum* Halim (Dinophyceae) bloom in the harbour of Sant Carles de la Ràpita (Ebro Delta, northwestern Mediterranean). Scientia Marina 1990, 54(1):1-7.

74. Maguer JF, Wafar M, Madec C, Morin P, Denn EEL: Nitrogen and phosphorus requirements of an *Alexandrium minutum* bloom in the Penzé Estuary, France. Limnology and Oceanography 2004, 49(4 I):1108-1114.

75. Garcés E, Bravo I, Vila M, Figueroa RI, Masó M, Sampedro N: Relationship between vegetative cells and cyst production during *Alexandrium minutum* bloom in Arenys de Mar harbour (NW Mediterranean). Journal of Plankton Research 2004, 26(6):637-645.

76. Galluzzi L, Penna A, Bertozzini E, Vila M, Garcés E, Magnani M: Development of a Real-Time PCR Assay for Rapid Detection and Quantification of *Alexandrium minutum* (a Dinoflagellate). Applied and Environmental Microbiology 2004, 70(2):1199-1206.

77. Van Lenning K, Vila M, Masó M, Garcés E, Anglès S, Sampedro N, Morales-Blake A, Camp J: Short-term variations in development of a recurrent toxic *Alexandrium minutum*-dominated dinoflagellate bloom induced by meteorological conditions. Journal of Phycology 2007, 43(5):892-907.

78. Blanco J, Reyero MI, Franco J: Kinetics of accumulation and transformation of paralytic shellfish toxins in the blue mussel *Mytilus galloprovincialis*. Toxicon 2003, 42(7):777-784.

79. Bricelj VM, Shumway SE: Paralytic Shellfish Toxins in Bivalve Molluscs: Occurrence, Transfer Kinetics, and Biotransformation. Reviews in Fisheries Science 1998, 6:315-383.

80. Navarro J, Contreras A: An integrative response by *Mytilus chilensis*; to the toxic dinoflagellate *Alexandrium catenella*. Marine Biology 2010, 157(9):1967-1974.

81. Feldmeyer B, Wheat C, Krezdorn N, Rotter B, Pfenninger M: Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 2011, 12(1):317.

82. McIntyre L, Lopiano K, Morse A, Amin V, Oberg A, Young L, Nuzhdin S: RNA-seq: technical variability and sampling. BMC Genomics 2011, 12(1):293.

83. Yao J, Yu F: DEB: A web interface for RNA-seq digital gene expression analysis. Bioinformation 2011, 7(1):44-45.

84. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26(1):139-140.

85. Gerdol M, Manfrin C, De Moro G, Figueras A, Novoa B, Venier P, Pallavicini A: The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A widespread and diverse family of immune-related molecules. Developmental & Comparative Immunology 2011, 35(6):635-643.

86. Vogél BE, Hedgecock EM: Hemicentin, a conserved extracellular member of the immunoglobulin superfamily, organizes epithelial and other cell attachments into oriented line-shaped junctions. Development 2001, 128(6):883-894.

87. Bork P, Sander C: A large domain common to sperm receptors (Zp2 and Zp3) and TGF-β type III receptor. FEBS Letters 1992, 300(3):237-240.

88. Thomas LR, Henson A, Reed JC, Salsbury FR, Thorburn A: Direct Binding of Fas-associated Death Domain (FADD) to the Tumor Necrosis Factor-related Apoptosis-inducing Ligand Receptor DR5 Is Regulated by the Death Effector Domain of FADD. Journal of Biological Chemistry 2004, 279(31):32780-32785.

89. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA et al: The Genome Portal of the Department of Energy Joint Genome Institute. Nucleic Acids Research 2012, 40(D1):D26-D32.

90. O'Neil S, Dzurisin J, Carmichael R, Lobo N, Emrich S, Hellmann J: Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon. BMC Genomics 2010, 11(1):310.

91. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 2008, 5(7):621-628.

92. Venier P, Varotto L, Rosani U, Millino C, Celegato B, Bernante F, Lanfranchi G, Novoa B, Roch P, Figueras A et al: Insights into the innate immunity of the Mediterranean mussel *Mytilus galloprovincialis.* BMC Genomics 2011, 12(1):69.

93. Gerdol M, De Moro G, Manfrin C, Venier P, Pallavicini A: Big defensins and mytimacins, new AMP families of the Mediterranean mussel *Mytilus galloprovincialis*. Developmental and Comparative Immunology 2012, 36(2):390-399.

94. Blanco J, Morono A, Franco J, Reyero MI: PSP detoxification kinetics in the mussel *Mytilus galloprovincialis*. One- and two-compartment models and the effect of some environmental variables. Marine Ecology Progress Series 1997, 158(1):165-175.

95. Fernández-Reiriz MJ, Navarro JM, Contreras AM, Labarta U: Trophic interactions between the toxic dinoflagellate *Alexandrium catenella* and *Mytilus chilensis*: Feeding and digestive behaviour to long-term exposure. Aquatic Toxicology 2008, 87(4):245-251.

96. Desbiens M, Cembella AD: Minimization of PSP toxin accumulation in cultured blue mussels (*Mytilus edulis*) by vertical displacement in the water column. In: Toxic phytoplankton blooms in the sea. Edited by Smayda TJ, Shimizu YT. Amsterdam: Elsevier; 1993: 395-400.

97. Agatha S, StrÜDer-Kypke MC, Beran A: Morphologic and Genetic Variability in the Marine Planktonic Ciliate *Laboea strobila* Lohmann, 1908 (Ciliophora, Oligotrichia), with Notes on its Ontogenesis. Journal of Eukaryotic Microbiology 2004, 51(3):267-281.

98. Lawrence JF, Niedzwiadek B, Menard C: Quantitative determination of paralytic shellfish poisoning toxins in shellfish using prechromatographic oxidation and liquid chromatography with fluorescence detection: interlaboratory study. J AOAC Int 2004, 87(1):83-100.

99. Quilliam MA, Janeček M, Lawrence JF: Characterization of the oxidation products of paralytic shellfish poisoning toxins by liquid chromatography/mass spectrometry. Rapid Communications in Mass Spectrometry 1993, 7(6):482-487.

100. Toulza E, Shin M-S, Blanc G, Audic S, Laabir M, Collos Y, Claverie J-M, Grzebyk D: Gene Expression in Proliferating Cells of the Dinoflagellate *Alexandrium catenella* (Dinophyceae). Appl Environ Microbiol 2010, 76(13):4521-4529.

101. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997, 25(17):3389-3402.

102. Conesa A, Götz S: Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics 2008, 2008.

103. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L et al: InterPro: the integrative protein signature database. Nucleic Acids Research 2009, 37(suppl 1):D211-D215.

104. Livak KJ, Schmittgen TD: Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. Methods 2001, 25(4):402-408.

105. Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM: Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neuroscience Letters 2003, 339(1):62-66.

Developmental and Comparative Immunology 35 (2011) 635–643

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

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

The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A widespread and diverse family of immune-related molecules

Marco Gerdol[a], Chiara Manfrin[a], Gianluca De Moro[a], Antonio Figueras[b], Beatriz Novoa[b], Paola Venier[c], Alberto Pallavicini[a,*]

[a] Department of Life Sciences, University of Trieste, Trieste, Italy
[b] Instituto de Investigaciones Marinas (CSIC), Eduardo Cabello 6, 36208 Vigo, Spain
[c] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A widespread and diverse family of immune-related molecules

Marco Gerdol[a], Chiara Manfrin[a], Gianluca De Moro[a], Antonio Figueras[b], Beatriz Novoa[b], Paola Venier[c], Alberto Pallavicini[a]

[a] Department of Life Sciences, University of Trieste, Trieste, Italy

[b] Instituto de Investigaciones Marinas (CSIC), Eduardo Cabello 6, 36208 Vigo, Spain

[c] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

Developmental & Comparative Immunology **35**(6): 635-643

Il C1q è la component chiave della via classica del complemento ed è ritenuto essere il collegamento principale tra l'immunità innata ed acquisita a causa dell'estrema plasticità del suo dominio trimerico globulare gC1q, che permette il riconoscimento ed il legame ad una vastissima gamma di bersagli. Il dominio gC1q caratterizza anche molte proteine non coinvolte nella via del complemento, implicate in una serie di altri processi biologici, inclusa l'apoptosi, l'infiammazione, l'adesione cellulare e la differenziazione cellulare. Nei molluschi, così come negli altri invertebrati, in cui l'immunità adattativa manca del tutto, le proteine contenenti domini C1q (indicate con l'acronimo C1qDC) sono piuttosto abbondanti e con ogni probabilità sono emerse come lectine, per poi evolversi in un'ampia classe di PRP (pattern recognition molecules) altamente specializzati, proprio grazie alle spiccate proprietà di interazione del dominio gC1q.
Sin dalle prime analisi delle risorse trascrittomiche contenute in Mytibase, era stata subito notata una notevole rappresentazione di trascritti C1qDC in *M. galloprovincialis*, dal momento che l'annotazione dei domini funzionali aveva evidenziato l'impronta molecolare C1q come la più abbondantei in assoluto dell'intero dataset. In questo studio sono state descritti ed analizzati i 168 trascritti full length codificanti per proteine C1qDC di mitilo, le cui sequenze sono state ricavate da Mytibase. Questo numero, ancorchè piuttosto elevato, è sicuramente lontano dal numero effettivo di

geni C1qDC presenti in questo organismo, dal momento che la profondità di sequenziamento utilizzata per la generazione di EST con metodologie Sanger non era particolarmente elevata e che molti di questi trascritti potrebbero essere espressi solamente in alcuni specifici tessuti in risposta a determinati stimoli. Una stima più recente, basata sull'RNA-sequencing condotto da ghiandola digestiva, suggerisce che il numero di geni C1q presenti nel genoma di *M. galloprovincialis* possa avvicinarsi addirittura al migliaio.

La grande abbondanza di proteine C1qDC in mitilo certamente suggerisce una strategia evolutiva di duplicazioni geniche, diversificazione funzionale e selezione di molti domini gC1q con proprietà di riconoscimento estremamente specifiche. Dati pre-esistenti avevano identificato gli emociti come tessuto di espressione principale del primo trascritto C1qDC ad essere descritto, MgC1q. Nel corso di questo lavoro sono stati analizzati i profili di espressione di 8 trascritti (MgC1q1-8) in diversi tessuti, mettendo in luce come non tutti siano espressi in modo specifico in queste cellule circolanti di grandissima importanza nei meccanismi legati all'immunità innata. Tuttavia tutti i 5 trascritti espressi preferenzialmente negli emociti testati, hanno dimostrato un forte incremento neli loro livelli di espressione, valutati con real-time PCR, in seguito a challenge batteriche effettuate sia con batteri Gram- (*V. anguillarum*) che con batteria Gram+ (*M. lysodeikticus*). Ciò suggerisce l'effettivo coinvolgimento delle proteine C1qDC di mitilo nell'immunità innata e più specificamente nel riconoscimento di PAMP, anche se non è affatto esclusa la possibilità di un loro diverso utilizzo in altri processi, in modo simile a quanto osservato nei vertebrati.

Un'analisi più comprensiva delle sequenze espresse per quanto riguarda l'ampio gruppo dei Protostomi ha inoltre rivelato che un numero di geni C1qDC così elevato non è affatto comune negli invertebrati. Risulta infatti evidente che questa famiglia genica si è incredibilmente espansa in modo selettivo soltanto nei Molluschi Bivalvi, mentre ad esempio nei Molluschi Gastropodi essa comprende solamente una manciata di geni. Curiosamente, una simile espansione è stata rilevata anche in *Daphnia pulex* (ma non in altri Crostacei), suggerendo che questi eventi potrebbero essere avvenuti anche in altri specifici *taxa*, in modo del tutto indipendente dagli eventi che hanno portato allo stabilimento di una larga famiglia di geni C1qDC nei Cordati.

TESTO ORIGINALE

Abstract

The key component of the classical complement pathway C1q is regarded as a major connecting link between innate and acquired immunity due to the highly adaptive binding properties of its trimeric globular domain gC1q. The gC1q domain also characterizes many non-complement proteins involved in a broad range of biological processes including apoptosis, inflammation, cell adhesion and cell differentiation. In molluscs and many other invertebrates lacking of adaptive immunity, C1q domain containing (C1qDC) proteins are abundant, they most probably emerged as lectins and subsequently evolved in a specialized class of pattern recognition molecules through the expanding interaction properties of gC1q.
Here we report the identification of 168 C1qDC transcript sequences of *Mytilus galloprovincialis*. The remarkable abundance of C1qDC transcripts in the Mediterranean mussel suggests an evolutionary strategy of gene duplication, functional diversification and selection of many specific C1qDC variants.
A comprehensive transcript sequence survey in Protostomia also revealed that the C1qDC family expansion observed in mussel could have occurred in some specific taxa independently from the events leading to the establishment of a large complement of C1qDC genes in the Chordates lineage.

Keywords: *M. galloprovincialis*; MgC1q; Immune gene; Bacterial infection; Expression level

Abbreviations: C1qDC, C1q domain containing; PRPs, pattern recognition proteins; PAMPs, pathogen-associated molecular patterns; ghC1q, globular head C1q; sghC1q, secreted globular head C1q.

1. Introduction

The gC1q is a globular domain which was first identified in the A, B and C chains of the C1q complement C1 complex subcomponent (Kishore and Reid, 2000). In addition to its fundamental role in the classical complement pathway, C1q provides a major link between innate and adaptive immunity, being involved in a wide range of immunological processes such as apoptotic cells clearance, bacteria and retrovirus recognition, cell adhesion and cell growth modulation (Kishore et al., 2004). Such an extreme versatility is granted by the ligand binding properties of the gC1q domain (Gaboriaud et al., 2003, Kishore and Reid, 1999).
The remarkable similarity of the gC1q and tumor necrosis factor (TNF) domains supports a common evolutionary origin for these two gene families (Shapiro and Scherer, 1998). Decisive amino acid changes and association to other functional domains can explain the wide variety of non-complement proteins: globally referred to C1qDC proteins they consist of an optional leading signal peptide, a central collagen-like region of variable length, acting as oligomerization domain and sometimes missing, and a C-terminal C1q domain (Ghai et al., 2007). Depending on the presence or the absence of the collagen-like region, C1qDC proteins are classified as C1q-like proteins or ghC1q proteins, respectively (Carland and Gerwick, 2010).

C1qDC proteins are probably essential in the innate immune system of early animals, as in the agnathan lamprey, having a still primitive adaptive immunity, C1q was shown to act as a lectin. Actually, lectin-like C1q proteins emerged before the immunoglobulins and expanded through the great flexibility and modulability of the gC1q domain in ligand binding (Fujita et al., 2004; Matsushita et al., 2004). Many C1qDC proteins can be regarded as specialized pattern recognition proteins (PRPs), able to bind pathogens directly through pathogen-associated molecular patterns (PAMPs) and to trigger phagocytosis (Bohlson et al., 2007; Medzhitov and Janeway, 2002).

Despite widespread in animal species, both retention or loss of C1q genes have apparently occurred in the evolution of Metazoa. Seven C1q gene models have been identified in the sea urchin *Strongylocentrotus purpuratus* (Hibino et al., 2006), only two in the Ascidian *Ciona intestinalis* (Azumi et al., 2003) and their number starts growing in ancestral Chordates: 50 C1q gene models in the Cephalochordate *Branchiostoma floridae* which is considered as the most primitive extant of the chordate lineage (Huang et al., 2008; Yu et al., 2008), 52 gene models in zebrafish (Mei and Gui, 2008) and 29 in humans (Tom Tang et al., 2005). On the contrary, C1qDC genes seem to be completely absent in Fungi and Plantae (Yuzaki, 2008).

Some C1qDC proteins with specific ligand recognition properties have been described and characterized also in molluscs. In particular, a sialic acid-binding lectin has been identified in the snail *Cepaea hortensis* (Gerlach et al., 2004) and an LPS-binding protein has been described in the scallop *Chlamys farreri* (Zhang et al., 2008). Other two C1qDC proteins, the major extrapallial fluid protein of *Mytilus edulis* (Hattan et al., 2001; Yin et al., 2005) and a protein highly expressed in the mantle tissue of *Pinctada fucata* (Liu et al., 2007) may be somehow involved also in the process of nacre biomineralization. The role of C1qDC proteins in specific pathogen recognition has been investigated in molluscs only recently: up-regulation of C1qDC proteins has been linked to infections with bacterial and metazoan parasites in molluscs such as *Ruditapes decussatus* (Prado-Alvarez et al., 2009), *Biomphalaria glabrata* (Adema et al., 2010), Crassostrea gigas (Taris et al., 2009) and *Mercenaria mercenaria* (Perrigault et al., 2009). AiC1qDC-1, a novel C1q domain containing protein recently characterized in the scallop *Argopecten irradians*, displays a fungi-agglutinating activity, and highlights, once again, the surprising ability of the gC1q domain to interact with many different PAMPs (Kong et al., 2010). In *Mytilus galloprovincialis*, the expression of MgC1q has been thoroughly examined in different tissues and larval stages (Gestal et al., 2010): MgC1q RNAs are abundant in hemocytes and increase rapidly and strongly in response to the injection of Gram+ and Gram- bacteria. Despite these facts point to an involvement of molluscan C1qDC proteins in pathogen recognition and innate immune response, to date the available data do not clarify the expansion and multifaceted functions of C1qDC proteins in this phylum and, more in general, in the Protostomia.

Recently, Carland and Gerwick (2010) reviewed the distribution of C1qDC proteins in animals, revealing the ancient origin of the gC1q domain and concluding that ghC1q genes became prevalent starting with Protostomia and radiated in the vertebrate animals. Here, we report and discuss for the first time the presence of a large family of C1qDC sequences almost exclusively coding for ghC1q proteins in the transcriptome of a non-Chordate organism, *M. galloprovincialis*. Despite the lack of genomic sequences, such an abundance and diversity of transcripts is suggestive of a similar over-representation of the C1qDC genes in the nuclear DNA. Mining publicly available transcriptomic and genomic data we also show that this astounding gene family expansion is restricted to Bivalvia and possibly to a few other unrelated Protostomia classes, and we raise the hypothesis that multiple

events of C1qDC gene family expansion can have occurred in few taxonomic groups independently from the events leading to the acquisition of a large complement of C1qDC genes in the Chordates lineage.

2. Materials and methods

2.1. Sequence analysis

We used Interproscan (Zdobnov and Apweiler, 2001) to identify the Interpro signature IPR001073 for the C1q domain in the 7112 independent sequences of Mytibase, the annotated EST database of *M. galloprovincialis* (Venier et al., 2009). We selected consensus sequences having a significant score for at least one of the four PRINTS, PROFILE, SMART and PFAM signatures for complement C1q, and the related clustered ESTs were individually checked for possible sequencing errors. To provide a conservative estimate of the C1q gene models present in Mytibase, an ESTs collection derived from many mussels, we collapsed in a single consensus both highly similar clusters, possibly originated by noisy chromatograms or sequencing errors, and clusters coding for peptides with an identity percentage greater than 75%, assuming they could refer to the same gene. All the resulting clusters were translated in putative proteins using the Expasy Translate tool (http://www.expasy.ch/tools/dna.html) and only the full length sequences were retained for subsequent analysis. The combined tools for transmembrane topology and signal peptide prediction Phobius (Kall et al., 2004) and SPOCTOPUS (Viklund et al., 2008) were both used to avoid misclassification of these two classes of hydrophobic regions. Coiled coiled domains were predicted with COILS (Lupas et al., 1991) considering true only the cases predicted with a probability higher than 0.7 in at least two out of the three given window sizes. The coiled coil domain containing sequences were then scanned for the presence of leucine zipper motifs using 2ZIP (Bornberg-Bauer et al., 1998). Interproscan supported the identification of additional domains other than C1q in the same proteins.

The full length mRNAs described in this manuscript have been submitted to the EMBL database under the accession numbers from FR715581 to FR715677.

2.2. Mussel samples

Mussels of 6.5-7 cm shell length were collected from a farming site of the Venice lagoon, Italy. To evaluate the tissue-specific expression of different C1qDC transcripts, total RNA was individually purified from hemolymph and from digestive gland, gill, gonads and posterior abductor muscle, previously homogenized in Tri reagent® (Sigma-Aldrich, St. Louis, MO).

Bacterial challenges were performed on adult mussels from Riá de Vigo, Spain, kept in tanks under controlled conditions (filtered seawater at 15ºC with aeration) and fed daily with *Isochrysis galbana*, Tetraselmis suecica and Skeletonema costatum. After an acclimatization time of 10 days, three groups of 60 mussels were challenged by injection into the adductor muscle with *Vibrio anguillarum* or *Micrococcus lysodeikticus* (100 µl of 107 live bacteria in filtered sea water). Controls were injected with 100 µl of filtered sea water. *M. lysodeikticus* was grown in LB medium at 37ºC and *V. anguillarum* in TSA supplemented with NaCl 1% at 20ºC. All individuals were maintained out of water for 20-30 min before and after the injection. At three, six and 24 hours

post-injection, the hemolymph was collected and pooled from 20 mussels per sampling time and treatment.
Following extraction, the RNA quality was assessed by electrophoresis on denaturing agarose gel and its quantity was estimated using a spectrophotometer. Complementary DNA was prepared by retro-transcription with the iScript™cDNA Synthesis Kit (Bio-Rad) from the pooled RNA samples representing five or 20 individuals.

2.3. Quantitative PCR expression analysis

The expression levels of eight selected C1q transcripts, namely MGC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7 and MgC1q8, were assessed in samples representing the hemocytes, digestive gland, gills, gonads and posterior abductor muscle of five adult mussels. Primer pairs were designed (Table 1) and used to obtain specific PCR amplicons, finally checking the reaction specificity by Sanger sequencing (ABI3130 Genetic Analyzer).
The expression of transcripts classified as hemocyte-specific, according to their relative abundance in the selected tissues, was also analyzed in the hemolymph sampled at 3, 6 and 24 hours post-challenge from mussels injected with Gram+ (*Micrococcus lysodeikticus*) or Gram- (*Vibrio anguillarum*) bacterial cells.
All the PCR assays were performed using a Bio-Rad CFX96 system. The 15 µL reaction mix included 0.75 µl of 20X EvaGreen™ (Biotium), 0.6 µl of 10 µM primer pairs and 5 µL of a 1:20 cDNA dilution. The following thermal profile was used: an initial 3' denaturation step at 95°C, followed by 35 cycles at 95° for 20", 56° for 15" and 72° for 20". Amplification products were analyzed with a 65°/95°C melting curve.
The expression levels of the selected transcripts were determined using the comparative Ct method (2-ΔΔCt method) (Livak and Schmittgen, 2001). Ct values used for quantification were corrected based on PCR efficiencies using LinRegPCR (Ramakers et al., 2003). The MgC1q expression values were normalized using the elongation factor EF-1 as housekeeping gene (EF-1 primers are shown in Table 1). Results are given as the mean with standard deviation of three technical replicates. The results were subjected to One-way Analysis of Variance (ANOVA) to determine significant differences in the mean values between the control and the the challenged groups. Significance was concluded at $P<0.01$.

2.4. Transcriptomic and genomic data mining

Transcriptomic data available for molluscan species with at least 10000 EST sequences were retrieved from the GenBank EST database (http://www.ncbi.nlm.nih.gov/nucest/) and from the SRA archive (http://www.ncbi.nlm.nih.gov/sra). Similarly, data were collected from the EST database for all the suitable Protostomia species, selecting up to three representative species per class.
Globally, 15 Mollusca, five Annelida, 16 Arthropoda, four Nematoda and Platyhelminthes and a single Onychophora and Rothifera species were included in the survey. The full list of the species selected is summarized in Table 2.
Sequence data were assembled with the CLC Genomic Workbench 4.02 (CLC Bio, Katrinebjerg, Denmark) to obtain a raw estimate of the total transcript number for the selected organisms. The longest ORF obtained for each sequence was then translated into the corresponding predicted protein and the resulting sequences were scanned for the presence of C1q profile with HMMER

(http://hmmer.wustl.edu/) in order to estimate the number of C1qDC transcripts and to calculate their relative abundance in the available transcriptome of each species analyzed.
The C1q profile used for the HMMER scanning was created by alignment of the C1q domains defining the PROSITE C1q profile PS50871 together with the publicly available C1q domains of invertebrate C1qDC proteins.
Selected Protostomia genomes available at the DOE Joint Genome Institute (http://www.jgi.doe.gov/), at Vectorbase (http://www.vectorbase.org/), at Wormbase (http://ws210.wormbase.org/) and at the Sanger Institute (http://www.sanger.ac.uk/) were also analyzed for the presence of C1qDC genes, scanning the predicted proteins with the same profile described above to verify the reliability of our transcriptomic approach on a genomic level.
More in detail, we downloaded and screened the whole genome protein models of *Helobdella robusta* (http://genome.jgi-psf.org/Helro1/Helro1.home.html), *Capitella teleta* (http://genome.jgi-psf.org/Capca1/Capca1.home.html), *Lottia gigantea* (http://genome.jgi-psf.org/Lotgi1/), *Daphnia pulex* (http://genome.jgi-psf.org/Dappu1/Dappu1.home.html), *Caenorhabditis elegans* (http://ws210.wormbase.org/), *Aedes aegiptyi* (http://aaegypti.vectorbase.org/), *Culex quinquefasciatus* (http://cquinquefasciatus.vectorbase.org/) and *Schistosoma mansoni* (http://www.genedb.org/Homepage/Smansoni).

3. Results

3.1. Sequence analysis

The Interproscan analysis identified in Mytibase a total of 168 C1qDC sequences, 96 of them coding for a full length protein. Two additional partial transcripts (MgC1q97 and MgC1q98) were elongated to the full length by Rapid Amplification of cDNA Ends (RACE). After virtual translation and conservative clustering of all the full-length sequences, we named them in a sequential order, starting from the first *Mytilus galloprovincialis* C1q transcript described in literature, MgC1q (Gestal et al., 2010). Remarkably, the multiple alignment of the 98 virtually translated full-length C1qDC proteins made evident the high sequence variability of the mussel C1q domains which display just a few conserved residues (Figure 1).
A signal peptide of 17-41 amino acid residues was identified in almost the totality of the *M. galloprovincialis* C1qDC proteins: more specifically a signal peptide was unambiguously predicted by Phobius in 91 out of 98 cases and, in two additional sequences the predicted signal was confirmed with SCOPTOPUS. Four out of the five remaining cases could be reasonably included within the false prediction rate reported to be 3,9% for Phobius (1,7% for SPOCTOPUS) and a trans-membrane domain was unambiguously predicted in a position incompatible with a signal peptide in the only sequence MgC1q98.
According to COILS analysis, 32 C1qDC proteins (32% of the total) also present coiled-coil regions in the N-terminal region. A leucine-zipper motif associated with the coiled-coil domain was identified by 2ZIP analysis in nine of these proteins (9% of the total). No other domain was found associated with C1q with a significant score by Interproscan analysis, with the exception of a collagen-like domain detected in MgC1q98 which is, curiously, also the only protein where a trans-membrane domain was unambiguously detected.

3.2. Tissue-specific expression

According to their abundance in Mytibase, structural diversity and homology to other previously described proteins, we chose eight among the 98 full length C1qDC transcripts to ascertain their expression levels in the main mussel tissues. Bidirectional sequencing of the PCR products obtained with specifically designed primer pairs confirmed the selective amplification of the 8 target sequences (non-specific amplification of similar mussel C1qDC transcripts was not observed).

The expression data are summarized in Figure 2. Though at different levels, constitutive expression of MgC1q1, MgC1q2, MgC1q3, MgC1q4 and MgC1q5 occurred mainly in the mussel hemocytes whereas other MgC1q transcripts resulted more expressed in other tissues (MGg1q7 in the digestive gland, MgC1q6 in the posterior abductor muscle, MgC1q8 in the gills). Melting curve analysis of the real time PCRs was systematically performed to exclude the formation of primer dimers and secondary products: samples where no amplification was observed or whose melting peaks resulted to be given by primer dimers were considered as tissues where the expression of a given transcript was so low to be undetectable, hence marked by "ND" in Figure 2.

3.3. C1q expression changes following microbial challenge

The expression level of the 'hemocyte-specific' transcripts MgC1q1, MgC1q2, MgC1q3, MgC1q4 and MgC1q5 was monitored also in mussels injected with a standard dose of live *Vibrio anguillarum* or *Micrococcus lysodeikticus* cells. Results are summarized in Figure 3.

Not surprisingly, all the selected transcripts showed a time-dependent expression pattern similar to that described by Gestal et al., 2010) for MgC1q1: a rapid expression increase in the very first hours after the challenge, a progressive decrease in the following hours and a return to physiological levels within 24 hours. The expression levels observed in response to both Gram+ and Gram- bacteria after three hours, ranged from 2.5 up to 5.5 times depending on the transcript examined and the type of challenge (Figure 3). Transcriptional down-regulation was evident for all transcripts already at 6 hours post-injection, with the only exception of MgC1q4 whose expression levels remained stable (and even showed an additional increase in the hemocytes of mussels injected with *V. anguillarum*).

At 24 hours post-injection, the MgC1q transcripts returned at levels similar or just slightly higher than those of the control group, except MgC1q3 whose expression significantly increased in both the groups of injected mussels. In general, similar expression trends characterized the response to Gram+ and Gram- bacteria, without evidence of specific induction for any of the tested transcripts.

3.4. Transcriptomic and genomic data mining

The analysis of EST data available for Protostomia evidenced that C1qDC transcripts are usually limited in number, accounting for 0 to about 0,1% of the total number of predicted transcripts, in many classes of large taxonomic groups such as Nematodes, Platyhelminthes and most Arthropods (Table 2).

Exceptions to such a rule of thumb can be found in Bivalvia, with proportions of C1qDC ESTs ranging from 2.36%, 0.87% and 0.543% and 0.34% in *M. galloprovincialis*, *C. virginica* and *M. californianus*, respectively, to values <0.1% in all the species considered for the two other major molluscan classes, Gastropoda and Cephalopoda. The only other case displaying a significant number of C1qDC transcripts resulted to be the crustacean genus Daphnia (0.17 to 0.36%).

These proportions can be influenced by several factors, including tissue of origin, developmental stage and possible immunostimulation. Therefore it has to be taken into account that our estimates of the C1qDC transcripts could be not exactly representative of the whole-organism transcriptome in physiological conditions. Nevertheless, we have no reason to assume that a certain bias towards immune transcripts is present in any of the transcriptomes we analyzed.
The search of predicted C1qDC gene models additionally performed in eight Protostomia genomes available at the DOE Joint Genome Institute (http://www.jgi.doe.gov/), at Vectorbase (http://www.vectorbase.org/), at Wormbase (http://ws210.wormbase.org/) and at the Sanger Institute (http://www.sanger.ac.uk/) confirmed the reliability of our transcriptomic approach and revealed a significant overall correlation ($p<0.01$) between the proportions of transcriptomic and genomic C1qDC sequences of the selected species (see *L. gigantea*, *C. teleta*, *H. robusta*, *A. aegiptyi*, *C. quinquefasciatus*, *D. pulex*, *S. mansoni* and *C. elegans* in Table 2).

4. Discussion

4.1. Evolutionary overview

The 168 C1qDC sequences identified in Mytibase indicate the abundance and molecular diversity of a specific class of molecules expressed in *M. galloprovincialis*. Since the Mytibase ESTs originated from mussels sampled in different locations and time periods, they cannot reveal the exact number of C1q genes present in the mussel genome. Furthermore, possible C1qDC genes expressed at low level are not likely to be present in Mytibase (yet to be discovered). Despite the lack of genomic data from *M. galloprovincialis*, the remarkable multiplicity of the mussel C1qDC sequences (see alignment of the C1q domains in Figure 1) and the specific amplification and sequencing of 8 exemplary C1q sequences also from the genomic DNA of a single mussel) suggest that the majority of Mytibase C1qDC transcripts are the product of different genes. The gene redundancy hyphotesized in mussel is striking, especially considering the number of C1qDC genes found in Chordates, the evolutionary lineage where gC1q apparently became prominent. Actually, the 52, 50 and 29 C1q gene models identified in zebrafish, amphioxus and humans, respectively, would be less than a half of the number of C1qDC genes conservatively estimated in *M. galloprovincialis*.
According to recent reviews, the evolution of C1q is still somehow obscure with unexplained "missing spots": despite being broadly represented in the animal kingdom, C1qDC proteins seem to be completely missing in several major phyla whereas the presence of C1q in some Bacillus species is still not completely understood (Ghai et al., 2007). Nevertheless, we can now report the existence of single C1qDC gene in the recently sequenced genome of the marine choanoflagellate *Monosiga brevicollis* (Joint Genome Institute *Monosiga brevicollis* v1.0. I genome release v1.0, protein ID 22872, King et al., 2008). As Choanozoa are the closest unicellular relatives of animals and fungi, this fact additionally supports the ancient origin of the C1q domain (Carland and Gerwick, 2010).
Despite limited to the available ESTs of selected Protostomia species, our transcriptomic survey provided a comparative overview of the C1q domain abundance in the main classes of invertebrates and, in our opinion, shed some light on the evolution of C1q in Protostomia.
C1qDC transcripts resulted to be infrequent (<0,1%) in many Invertebrate taxa including flat worms, Annelids, Insects, Arachnids and most crustaceans, and apparently completely absent in

Nematoda, Rothifera and Onychophora (see Table 2). Owing to the incompleteness of transcriptome data and the low number of selected species, we cannot exclude the existence of C1qDC genes in the genomes of such species and organism classes.

One of the few exceptions to the low representation of invertebrate C1qDC transcripts is represented by the class of Bivalves: besides *M. galloprovincialis* (2.36%) also other three species display a not negligible proportion of C1qDC transcripts (*Mytilus californianus*, 0.38%; *Crassostrea gigas*, 0.38%, *Crassostrea viginica*, 0.87%) whereas only 0.06% could be reported in the Antarctic clam Laternula elliptica, and the presence of C1qDC transcripts estimated in Gastropoda and Cephalopoda species was very scarce. The amazing multiplicity of C1qDC transcripts in most Bivalvia, compared to the other Mollusca classes, is suggestive of an expansion event of the C1qDC gene family restricted to this class.

A similar event may have occurred also in the Crustacean genus *Daphnia* (C1qDC transcript estimated to be 0,36% in *Daphnia pulex* and 0,17% in *Daphnia magna*) in contrast to all the other selected crustacean species characterized by a negligible expression of C1qDC molecules. The driving forces leading to such an extremely specific and likely unrelated expansion of the C1qDC gene family in two distant groups as the Bivalvia class and the *Daphnia* genus are unknown. Independent expansion of C1qDC genes may have also occurred in other Protostomia classes which have not been analyzed in our transcriptomic survey.

The trends inferred from the transcriptomic survey find a strong support in some ongoing genome sequencing programs. To correlate the C1qDC representation in transcriptomes and the available corresponding genomes, we performed a statistical analysis to calculate the canonical correlation between the two independent variables; the finding of a strong linear combination of the two ratios ($P<0.01$) supports our experimental data also in organisms whose genome is not available yet. In fact, no C1qDC genes could be identified in the insects *Aedes aegiptyi* (http://aaegypti.vectorbase.org/) and *Culex quinquefasciatus* (http://cquinquefasciatus.vectorbase.org/) and in the nematode *Caenorhabditis elegans* (http://ws210.wormbase.org/), but we could identify two C1q gene models in the flatworm *Schistosoma mansoni* (http://www.genedb.org/Homepage/Smansoni), eight and 24 gene models in the Annelida *Helobdella robusta* (http://genome.jgi-psf.org/Helro1/Helro1.home.html) and *Capitella teleta* (http://genome.jgi-psf.org/Capca1/Capca1.home.html) respectively and only 6 gene models in the limpet *Lottia gigantea* (http://genome.jgi-psf.org/Lotgi1/). In other words, the scarce evidence of independent C1qDC transcripts in these seven species is confirmed by an equally small number of predicted genes (the relative abundance of C1qDC transcript and gene models are shown in Table 2). Similarly, in the crustacean *Daphnia pulex* (http://genome.jgi-psf.org/Dappu1/Dappu1.home.html) a total of 70 C1qDC transcripts and 144 gene models (accounting for 0,47% of the total and often organized in dense gene clusters) strongly support the multiplicity of these molecules.

Overall, the relative abundance of C1qDC ESTs in the eight mentioned species reflects the actual abundance of C1qDC genes in their genomes, as supported by the strong canonical correlation observed between the two ratios. Accordingly, the great number of C1qDC transcripts in *M. galloprovincialis* suggests a similar remarkable abundance of C1qDC genes in its genome.

4.2. Structural features of the *M. galloprovincialis* C1qDC proteins

Almost all the 168 C1qDC proteins virtually identified in *M. galloprovincialis* show a N-terminal signal peptide, with the few exceptions likely being the result of mispredictions or sequencing errors and a single case unambiguously predicted as trans-membrane protein (MgC1q98). Hence, almost the entire complement of mussel C1qDC proteins seems to be destined to the secretory pathway.

The usual structure of C1qDC proteins also includes a C-terminal C1q domain, currently regarded as the most widespread although not exclusive feature of this family of proteins, and a central collagen-like region which may or may be not present. With no exception, the C1qDC proteins of *M. galloprovincialis* show a C-terminal C1q domain. On the contrary, the presence of a collagen-like glycine rich region is absolutely uncommon, as it was identified just in MgC1q98 which is also the only non-secreted C1qDC protein detected in mussel. Taken together, the absence of a collagen-like region and the presence of signal peptide classify the vast majority of mussel C1qDC proteins as sghC1q (secretory globular head C1q) proteins with the only exception of the C1q-like MgC1q98 (Carland and Gerwick, 2010).

As the collagen-like region has a stabilizing role on the heterotrimeric structure of C1q and supports the assembly of higher-order complexes, most mussel C1qDC proteins should merely rely on the interactions mediated by the C1q domains or other N-terminal structures. Interestingly, almost one third of mussel C1qDC proteins are characterized by the presence of coiled coil region N-terminal to C1q, occasionally embedding a leucine-zipper motif. Both coiled-coils and leucine zippers are known to act as multimerization domains (Lupas, 1996; Tadokoro et al., 1999) and such a role has been suggested for them in vertebrate C1qDC proteins such as emilins and multimerin (Doliana et al., 1999; Hayward et al., 1995). Given the absence of collagen-like regions, the association of C1q with coiled-coil and leucine-zipper domains in Mollusca, and possibly in other Protostomia, could reasonably represent an alternative strategy for the association of C1qDC proteins in multimeric complexes.

Our transcriptomic survey also revealed several C1qDC proteins of Annelida and Crustacea associated to other N-terminal protein domains, especially chitin-binding domains in Annelida and fibrinogen or COLFI domains in Crustacea. Different from those usually found in vertebrates, the protein domains associated to C1q indicate the need of specific studies based on invertebrate models. Since we could not identify any unconventional N-terminal domain associated to C1q in *M. galloprovincialis*, such a feature could be completely missing in Mollusca or, more simply, in the C1qDC sequences of Mytibase.

Table 3 and Figure 4 illustrate the main structural features of the C1qDC proteins deduced from the eight exemplary MgC1q sequences used in the tissue-specific expression analysis: MgC1q, M1gC1q2, MgC1q4, MgC1q5 and MgC1q8 show show N-terminal signal peptide and C-terminal C1q, hence typical sghC1q proteins, whereas MgC1q3, MgC1q6 and MgC1q7 also have a coiled-coil domain, which contains a leucine-zipper motif in MgC1q7.

Furthermore, we report few cases of C1qDCs with multiple C1q domains, namely the Mytibase clusters MGC06942, MGC07609 and MGC07852. The complete mRNA of MGC07852 was achieved by RACE analysis and named MgC1q97; the deduced protein resulted to include a signal peptide and 3 consecutive C1q domains, with a total length of 441 amino acids. C1qDC proteins with multiple C1q domains have been described in vertebrates (Tom Tang et al., 2005). According to the analysis of the C1qDC genes of *C. teleta* (Joint Genome Institute Capitella teleta v1.0. I

genome release, protein ID: 215797), we can now report the presence of C1qDC proteins with at least 4 C1q domains. As revealed by the transcriptome survey, proteins with multiple C1q domains can be also identified in the oysters *C. virginica* (EST accession numbers: CV089299, CV89256, CV89284, CV133085, CV0874141, CV132342 and CV132710) and *C. gigas* (EST accession numbers: CU987496, CU993590, CU682562, CU993633.1, CU683542, CU996256 and FP003470).

4.3. Expression and response to bacterial challenges

The Real-Time qPCR analyses revealed that 5 of the 8 selected MgC1q transcripts (MgC1q 1-5) are constitutively espressed at variable, often negligible, levels in the main mussel tissues except in hemocytes where their expression increases at significant levels, as previously reported for MgC1q (Gestal et al., 2010). The cells circulating in the hemolymph, and infiltrating tissues when alerted by specific signals, are currently regarded as the major players of the innate immunity system of mussels and, in general, invertebrate organisms. The present data confirm significant constitutive levels of C1qDC transcripts in the hemocytes of adult mussels, as expected from a specialized transcriptome rich of immune-related molecules (Gestal et al., 2010; Pallavicini et al., 2008; Venier et al., 2009).

The expression of MgC1q, C1qDC transcript uniquely clustering 112 Mytibase ESTs, was confirmed about 6 fold higher than the elongation factor 1 in hemocytes. The expression of the other 'hemocyte-specific' C1qDC transcripts ranged from about 0,3 fold (MgC1q5) to about 3 fold (MgC1q3) the level of the elongation factor-1. On the other hand, the expression of MgC1q7 and MgC1q8 was specific to the digestive gland and gills, respectively, whereas MgC1q6 the homologue of the *Mytilus edulis* major extrapallial fluid protein was specifically expressed, about 2.5 fold compared to EF-1, in the posterior abductor muscle. Taken together, these expression data suggest that the diversification which occurred within the C1q family may have led some of its members to carry out specialized functions, other than those of the innate immunity, in different tissues.

As a matter of fact, increased versatility of the gC1q binding and association with different N-terminal domains have likely expanded the functional roles currently recognized in the C1qDC proteins of Chordates. Except for coiled/coil and leucin-zipper multimerization domains no other N-terminal domain has been found associated with C1q molecules in *M. galloprovincialis*, and the search of additional functions is not feasible at the moment.

The remarkable up-regulation observed for all the hemocyte-specific transcripts in response to the injection of both Gram+ and Gram- cells suggests once again the involvement of mussel C1qDC proteins in the innate immune responses and confirms the plasticity of the gC1q domain as potential PAMPs recognition receptor. The significant increase of expression observed at three hours post-injection, already detected for MgC1q1 at one hour post-injection and with higher levels, as shown by Gestal et al. (2010), reinforces the idea of C1qDC proteins as PRPs involved in the early phases of defense and able to trigger later complex modulations of the hemocyte behavior. Overall, the multiplicity of the C1qDC transcripts identified in *M. galloprovincialis* suggests an evolutive strategy of gene duplication and diversification/specialization in response to potential pathogens and, possibly, to other signals; however, the expression levels of 8 exemplary MgC1q in the hemocytes of mussels injected with living bacteria did not reveal a specific pattern of response towards the Gram+ and Gram- cells. Common regulatory mechanisms leading to the up-regulation

of similar gene sets in response to pathogens could explain these findings and do not exclude specific interaction under different experimental conditions.

5. Conclusions

To date, the C1qDC proteins have always been considered to be a family only sporadically represented in animals before the onset of the Chordates lineage. Here we report for the first time the existence of a large C1qDC protein family in a Prostostome. In fact, more than one hundred C1q domain containing proteins are likely to be encoded by the *Mytilus galloprovincialis* genome, mostly pertaining to the sghC1q group. Our experimental data support the possible involvement of many invertebrate C1qDC proteins as ancient innate immune response proteins, but the role of specific members of this highly diverse family in many different processes other than pathogen recognition needs additional study. A comparative transcriptomic survey performed on the C1qDC proteins in many different Protostomia phyla, suggested that expansion of the C1q genes family may have sporadically and independently occurred in a few specific classes, including Bivalvia, separately from the the emergence of a large consolidated C1qDC genes repertoire in the Chordate lineage.

Acknowledgements
This work was supported by the European Integrated Project FOOD-CT-2005-007103 (http://imaquanim.dfvf.dk/info) and by Regione Friuli Venezia Giulia, Direzione Centrale Risorse Agricole, Naturali, Forestali e Montagna, L.R. 26/2005 prot. RAF/9/7.15/47174.

TABLES

primer name	primer sequence
MGC1Q2_FOR	gcaagacaaagtcggagtgga
MGC1Q2_REV	agcaccaacaatgccagacg
MgC1q1_FOR	cagggtcagattacagcgtcttca
MgC1q1_REV	cgatttttgtgctgcccatc
MGC1Q3_FOR	tgtgcctcaggaaaatcctcttgc
MGC1Q3_REV	ccgtctggtatctcggaatcg
MGC1Q4_FOR	aagcagcaagcattcccgta
MGC1Q4_REV	ccatcgctaggtgctgtgaa
MGC1Q5_FOR	taaagccggactgtacttggtgtc
MGC1Q5_REV	atctccctctgctgcctgta
MGC1Q6_FOR	ctggtgctgttttgcgttgtcag
MGC1Q6_REV	ttttcgatttcgtggtggat
MGC1Q7_FOR	aggtggcgttttatgctgcgttga
MGC1Q7_REV	ggagcagtaaacatgccatttaca
MGC1Q8_FOR	ccaattcgcagtgagttttgt
MgC1q8_REV	gtgtggcttgtaaagatcctgctg
EF-1_FOR	cctcccaccatcaagaccta
EF-1_REV	ggctggagcaaaggtaacaa

Table 1: primers designed for assessing tissue-specific expression and up-regulation of the transcripts MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7 and MgC1q8 in response to bacterial challenges.

Species	Class	Sequence type	Number of sequences	Number of assembled contigs	Predicted C1q-DC transcripts	Relative rapresentation of C1q transcripts in the transcriptome
MOLLUSCA						
Mytilus Galloprovincialis	**Bivalvia**	**Sanger**	**18788**	**7112**	**168**	**2,362**
Crassostrea gigas	Bivalvia	Sanger	57279	10031	38	0,379
Mytilus californianus	Bivalvia	Sanger	42354	9570	52	0,543
Crassostrea virginica	Bivalvia	Sanger	14560	1734	15	0,865
Laternula elliptica	Bivalvia	454	123135	6619	4	0,060
Aplysia californica	Gastropoda	Sanger	255605	376698	8	0,002
Aplysia californica	Gastropoda	Illumina	58073706			
Lottia gigantea	Gastropoda	Sanger	252091	19996	3	0,015
Biomphalaria glabrata	Gastropoda	Sanger	54309	35687	13	0,036
Biomphalaria glabrata	Gastropoda	454	704022			
Lymnaea stagnalis	Gastropoda	Sanger	11697	2291	1	0,044
Aplysia kurodai	Gastropoda	Sanger	11445	1290	0	0,000
Ilyanassa obsoleta	Gastropoda	454	1387166	127783	19	0,015
Littorina saxatilis	Gastropoda	454	298623	25832	9	0,035
Crepidula fornicata	Gastropoda	454	1297588	62835	12	0,019
Strombus gigas	Gastropoda	454	286933	26369	12	0,046
Euprymna scolopes	Cephalopoda	Sanger	35420	7361	0	0,000
ANNELIDA						
Alvinella pompejana	Polychaeta	Sanger	218454	20333	2	0,010
Capitella teleta	Polychaeta	Sanger	138404	13694	9	0,066
Helobdella robusta	Citellata	Sanger	101359	11754	8	0,068
Hirudo medicinalis	Citellata	Sanger	26833	6426	4	0,062
Lumbricus rubellus	Citellata	Sanger	20239	2567	2	0,078
ARTHROPODA						
Aedes aegypti	Insecta	Sanger	301596	21424	0	0,000
Culex quinquefasciatus	Insecta	Sanger	205275	7036	0	0,000
Dendroctonus ponderosae	Insecta	Sanger	152724	10578	0	0,000
Onychiurus arcticus	Entognatha	Sanger	16379	3106	0	0,000
Litopenaeus vannamei	Malacostraca	Sanger	161091	13332	0	0,000
Petrolisthes cinctipes	Malacostraca	Sanger	97806	13088	3	0,023
Penaeus monodon	Malacostraca	Sanger	35396	3540	0	0,000
Daphnia pulex	Branchiopoda	Sanger	152659	19264	70	0,363
Artemia franciscana	Branchiopoda	Sanger	37590	2569	0	0,000
Daphnia magna	Branchiopoda	Sanger	13400	1750	3	0,171
Lepeophtheirus salmonis	Maxillopoda	Sanger	129250	16226	3	0,018

Caligus rogercresseyi	Maxillopoda	Sanger	32037	6917	4	0,058
Lernaeocera branchialis	Maxillopoda	Sanger	14927	4048	1	0,025
Tetranychus urticae	Arachnida	Sanger	80855	10918	3	0,027
Rhipicephalus microplus	Arachnida	Sanger	52838	10968	0	0,000
Rhipicephalus appendiculatus	Arachnida	Sanger	19123	2879	0	0,000
ONYCHOPHORA						
Peripatopsis sedgwicki	unassigned	Sanger	10476	1081	0	0
ROTHIFERA						
Brachionus plicatilis	Monogononta	Sanger	52771	8255	0	0
PLATYHELMINTHES						
Schistosoma mansoni	Trematoda	Sanger	205892	14937	1	0,007
Schistosoma japonicum	Trematoda	Sanger	103725	10507	0	0,000
Schmidtea mediterranea	Turbellaria	Sanger	78333	10023	1	0,010
Taenia solium	Cestoda	Sanger	30587	3079	0	0,000
NEMATODA						
Caenorhabditis elegans	Chromadorea	Sanger	393714	23775	0	0
Ancylostoma caninum	Chromadorea	Sanger	80905	10720	0	0
Ascaris suum	Chromadorea	Sanger	56118	3886	0	0
Trichinella pseudospiralis	Enoplea	Sanger	17330	2042	0	0

Table2: Relative abundance of C1qDC transcripts in representative Protostomes. The 3 species with the most representative transcriptomes were selected for a single taxonomic class, with the exception of Mollusca, where all the suitable species with more than 10000 ESTs were analyzed. The number and type of sequences used for the assembly is indicated, and the percentage of C1qDC transcripts or genes is calculated on the total number of the predicted transcripts or genes. *-T: transcriptome sequencing; -G: genomic sequencing.

Transcript/Protein name	*ESTs in Mytibase*	*Protein length (aa)*	*Signal peptide*	*Coiled-coil domain*	*Leucine-zipper domain*	*C1q domain*
MgC1q1	112	169	YES	NO	NO	C-terminal
MgC1q2	26	194	YES	NO	NO	C-terminal
MgC1q3	18	274	YES	YES	NO	C-terminal
MgC1q4	22	182	YES	NO	NO	C-terminal
MgC1q5	10	186	YES	NO	NO	C-terminal
MgC1q6	10	231	YES	YES	NO	C-terminal
MgC1q7	10	231	YES	YES	YES	C-terminal
MgC1q8	17	199	YES	NO	NO	C-terminal

Table 3: Main structural features of the 8 Mytibase (*M. galloprovincialis*) transcript sequences selected for the evaluation of tissue-specific expression.

FIGURES

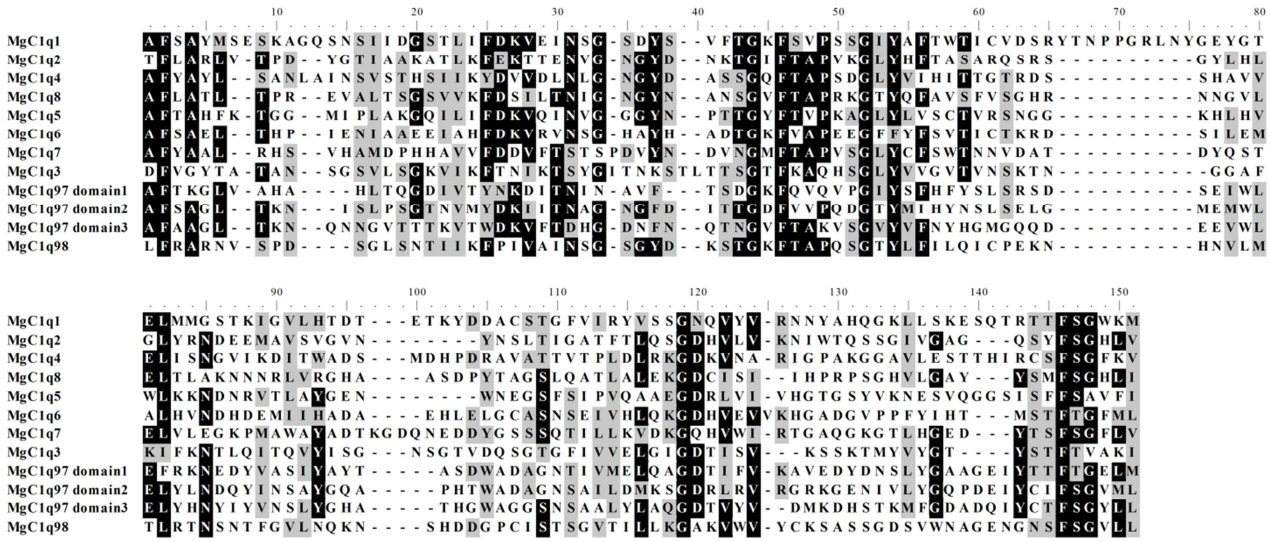

Figure 1: Multiple alignment of c1q domains from the 10 selected *Mytilus galloprovincialis* C1qDC proteins MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7, MgC1q8, MgC1q97 and MgC1q98; all the 3 different MgC1q97 C1q domains are represented in the alignment.

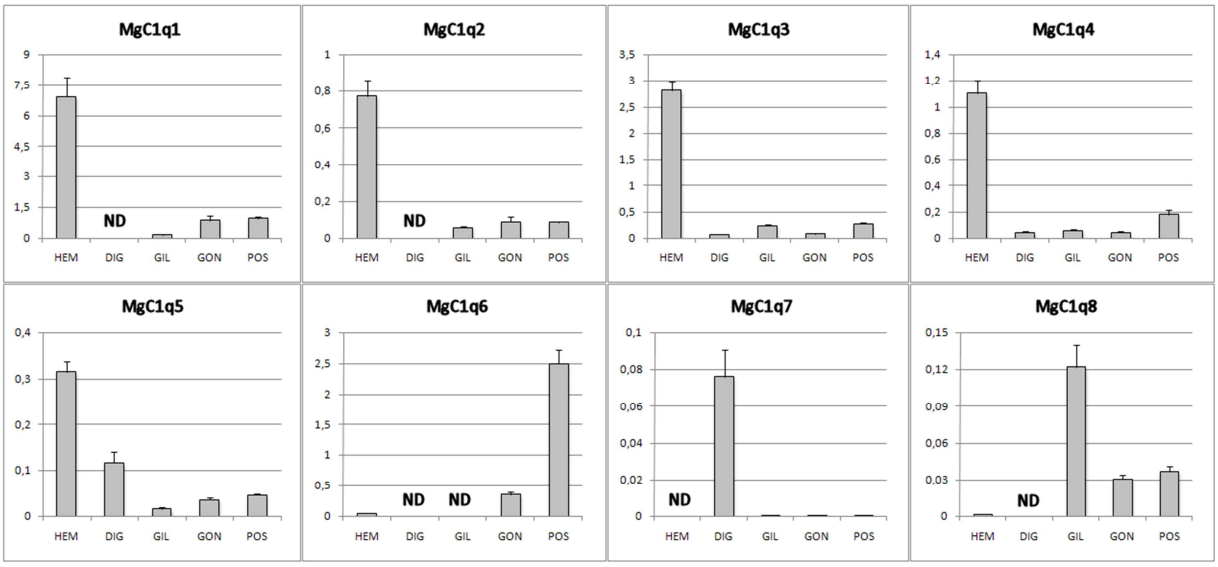

Figure 2: Tissue-specific expression of the mussel C1qDC transcripts MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7 and MgC1q8. Bars depict the transcript expression relative to the elongation factor EF-1. Results are mean ± SD of 3 technical replicates. Y axis of each graph is scaled based on the highest level of expression. HEM: Hemocyte cells, DIG: digestive gland, GIL: gills, GON: gonads, POS: posterior abductor muscle.

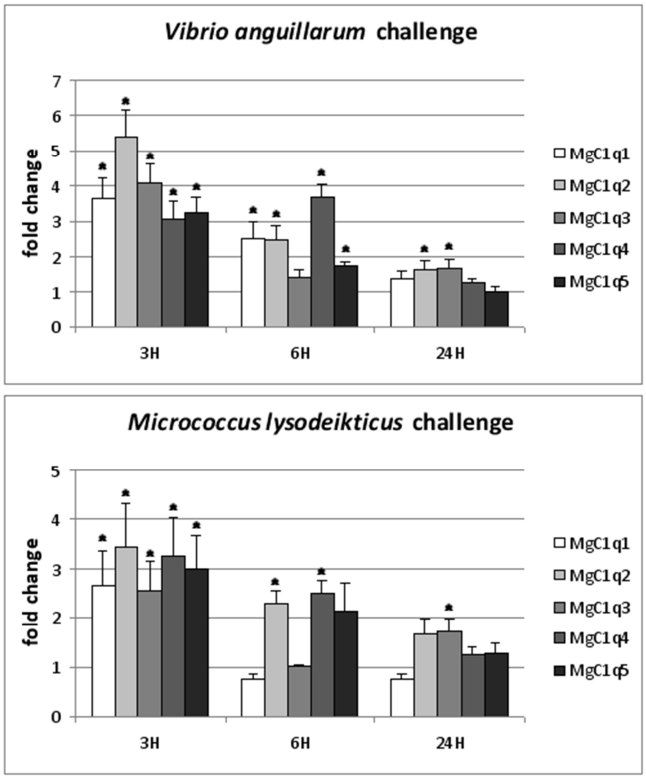

Figure 3: Expression changes of the transcripts MgC1q1, MgC1q2, MgC1q3, MgC1q4 and MgC1q5 in hemocytes sampled at 3, 6 and 24 hours post-injection from mussels challeged with Gram- (*V. anguillarum*, black bars) and Gram+ (*M. lysodeikticus*, white bars) bacteria; error bars represent fold change ± standard deviation of 3 technical replicates relative to the expression levels of untreated mussels, previously normalized to the elongation factor EF-1. Significant differences between challenged group and control group were indicated by an asterisk ($P<0.01$).

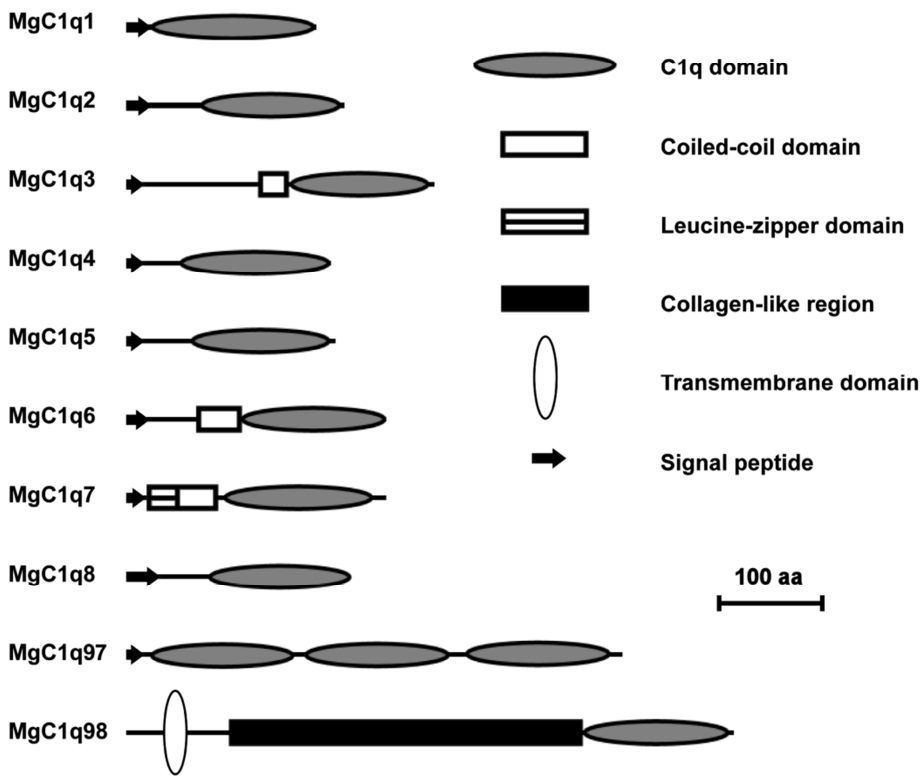

Figure 4: Structural organization of the selected *M. galloprovincialis* C1qDC proteins MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7, MgC1q8, MgC1q97 and MgC1q98.

References

Adema, C. M., Hanington, P. C., Lun, C. M., Rosenberg, G. H., Aragon, A. D., Stout, B. A., Lennard Richard, M. L., Gross, P. S., Loker, E. S., 2010. Differential transcriptomic responses of *Biomphalaria glabrata* (Gastropoda, Mollusca) to bacteria and metazoan parasites, *Schistosoma mansoni* and *Echinostoma paraensei* (Digenea, Platyhelminthes). Mol Immunol. 47, 849-60.

Azumi, K., De Santis, R., De Tomaso, A., Rigoutsos, I., Yoshizaki, F., Pinto, M. R., Marino, R., Shida, K., Ikeda, M., Arai, M., Inoue, Y., Shimizu, T., Satoh, N., Rokhsar, D. S., Du Pasquier, L., Kasahara, M., Satake, M., Nonaka, M., 2003. Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "waiting for Godot". Immunogenetics. 55, 570-81.

Bohlson, S. S., Fraser, D. A., Tenner, A. J., 2007. Complement proteins C1q and MBL are pattern recognition molecules that signal immediate and long-term protective immune functions. Mol Immunol. 44, 33-43.

Bornberg-Bauer, E., Rivals, E., Vingron, M., 1998. Computational approaches to identify leucine zippers. Nucleic Acids Res. 26, 2740-6.

Carland, T. M., Gerwick, L., 2010. The C1q domain containing proteins: Where do they come from and what do they do? Dev Comp Immunol. 34, 785-90.

Doliana, R., Mongiat, M., Bucciotti, F., Giacomello, E., Deutzmann, R., Volpin, D., Bressan, G. M., Colombatti, A., 1999. EMILIN, a component of the elastic fiber and a new member of the C1q/tumor necrosis factor superfamily of proteins. J Biol Chem. 274, 16773-81.

Fujita, T., Matsushita, M., Endo, Y., 2004. The lectin-complement pathway--its role in innate immunity and evolution. Immunol Rev. 198, 185-202.

Gaboriaud, C., Juanhuix, J., Gruez, A., Lacroix, M., Darnault, C., Pignol, D., Verger, D., Fontecilla-Camps, J. C., Arlaud, G. J., 2003. The crystal structure of the globular head of complement protein C1q provides a basis for its versatile recognition properties. J Biol Chem. 278, 46974-82.

Gerlach, D., Schlott, B., Schmidt, K. H., 2004. Cloning and expression of a sialic acid-binding lectin from the snail *Cepaea hortensis*. FEMS Immunol Med Microbiol. 40, 215-21.

Gestal, C., Pallavicini, A., Venier, P., Novoa, B., Figueras, A., 2010. MgC1q, a novel C1q-domain-containing protein involved in the immune response of *Mytilus galloprovincialis*. Dev Comp Immunol. 34, 926-34.

Ghai, R., Waters, P., Roumenina, L. T., Gadjeva, M., Kojouharova, M. S., Reid, K. B., Sim, R. B., Kishore, U., 2007. C1q and its growing family. Immunobiology. 212, 253-66.

Hattan, S. J., Laue, T. M., Chasteen, N. D., 2001. Purification and characterization of a novel calcium-binding protein from the extrapallial fluid of the mollusc, *Mytilus edulis*. J Biol Chem. 276, 4461-8.

Hayward, C. P., Hassell, J. A., Denomme, G. A., Rachubinski, R. A., Brown, C., Kelton, J. G., 1995. The cDNA sequence of human endothelial cell multimerin. A unique protein with RGDS, coiled-coil, and epidermal growth factor-like domains and a carboxyl terminus similar to the globular domain of complement C1q and collagens type VIII and X. J Biol Chem. 270, 18246-51.

Hibino, T., Loza-Coll, M., Messier, C., Majeske, A. J., Cohen, A. H., Terwilliger, D. P., Buckley, K. M., Brockton, V., Nair, S. V., Berney, K., Fugmann, S. D., Anderson, M. K., Pancer, Z., Cameron, R. A., Smith, L. C., Rast, J. P., 2006. The immune gene repertoire encoded in the purple sea urchin genome. Dev Biol. 300, 349-65.

Huang, S., Yuan, S., Guo, L., Yu, Y., Li, J., Wu, T., Liu, T., Yang, M., Wu, K., Liu, H., Ge, J., Huang, H., Dong, M., Yu, C., Chen, S., Xu, A., 2008. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. Genome Res. 18, 1112-26.

Kall, L., Krogh, A., Sonnhammer, E. L., 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338, 1027-36.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J. B., Morris, A., Nichols, S., Richter, D. J., Salamov, A., Sequencing, J. G., Bork, P., Lim, W. A., Manning, G., Miller, W. T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I. V., Rokhsar, D., 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature. 451, 783-8.

Kishore, U., Gaboriaud, C., Waters, P., Shrive, A. K., Greenhough, T. J., Reid, K. B., Sim, R. B., Arlaud, G. J., 2004. C1q and tumor necrosis factor superfamily: modularity and versatility. Trends Immunol. 25, 551-61.

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

Kishore, U., Reid, K. B., 2000. C1q: structure, function, and receptors. Immunopharmacology. 49, 159-70.

Kong, P., Zhang, H., Wang, L., Zhou, Z., Yang, J., Zhang, Y., Qiu, L., Song, L., 2010. AiC1qDC-1, a novel gC1q-domain-containing protein from bay scallop *Argopecten irradians* with fungi agglutinating activity. Dev Comp Immunol. 34, 837-46.

Liu, H. L., Liu, S. F., Ge, Y. J., Liu, J., Wang, X. Y., Xie, L. P., Zhang, R. Q., Wang, Z., 2007. Identification and characterization of a biomineralization related gene PFMG1 highly expressed in the mantle of *Pinctada fucata*. Biochemistry. 46, 844-51.

Livak, K. J., Schmittgen, T. D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 25, 402-8.

Lupas, A., 1996. Coiled coils: new structures and new functions. Trends Biochem Sci. 21, 375-82.

Lupas, A., Van Dyke, M., Stock, J., 1991. Predicting coiled coils from protein sequences. Science. 252, 1162-4.

Matsushita, M., Matsushita, A., Endo, Y., Nakata, M., Kojima, N., Mizuochi, T., Fujita, T., 2004. Origin of the classical complement pathway: Lamprey orthologue of mammalian C1q acts as a lectin. Proc Natl Acad Sci U S A. 101, 10127-31.

Medzhitov, R., Janeway, C. A., Jr., 2002. Decoding the patterns of self and nonself by the innate immune system. Science. 296, 298-300.

Mei, J., Gui, J., 2008. Bioinformatic identification of genes encoding C1q-domain-containing proteins in zebrafish. J Genet Genomics. 35, 17-24.

Pallavicini, A., Costa Mdel, M., Gestal, C., Dreos, R., Figueras, A., Venier, P., Novoa, B., 2008. High sequence variability of myticin transcripts in hemocytes of immune-stimulated mussels suggests ancient host-pathogen interactions. Dev Comp Immunol. 32, 213-26.

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

Prado-Alvarez, M., Gestal, C., Novoa, B., Figueras, A., 2009. Differentially expressed genes of the carpet shell clam *Ruditapes decussatus* against *Perkinsus olseni*. Fish Shellfish Immunol. 26, 72-83.

Ramakers, C., Ruijter, J. M., Deprez, R. H., Moorman, A. F., 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett. 339, 62-6.

Shapiro, L., Scherer, P. E., 1998. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. Curr Biol. 8, 335-8.

Tadokoro, S., Tachibana, T., Imanaka, T., Nishida, W., Sobue, K., 1999. Involvement of unique leucine-zipper motif of PSD-Zip45 (Homer 1c/vesl-1L) in group 1 metabotropic glutamate receptor clustering. Proc Natl Acad Sci U S A. 96, 13801-6.

Taris, N., Lang, R. P., Reno, P. W., Camara, M. D., 2009. Transcriptome response of the Pacific oyster (*Crassostrea gigas*) to infection with *Vibrio tubiashii* using cDNA AFLP differential display. Anim Genet. 40, 663-77.

Tom Tang, Y., Hu, T., Arterburn, M., Boyle, B., Bright, J. M., Palencia, S., Emtage, P. C., Funk, W. D., 2005. The complete complement of C1q-domain-containing proteins in *Homo sapiens*. Genomics. 86, 100-11.

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

Viklund, H., Bernsel, A., Skwark, M., Elofsson, A., 2008. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics. 24, 2928-9.

Yin, Y., Huang, J., Paine, M. L., Reinhold, V. N., Chasteen, N. D., 2005. Structural characterization of the major extrapallial fluid protein of the mollusc *Mytilus edulis*: Implications for function. Biochemistry. 44, 10720-31.

Yu, Y., Huang, H., Wang, Y., Yuan, S., Huang, S., Pan, M., Feng, K., Xu, A., 2008. A novel C1q family member of amphioxus was revealed to have a partial function of vertebrate C1q molecule. J Immunol. 181, 7024-32.

Yuzaki, M., 2008. Cbln and C1q family proteins – New transneuronal cytokines. Cell Mol Life Sci. 65, 1698-705.

Zdobnov, E. M., Apweiler, R., 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17, 847-8.

Zhang, H., Song, L., Li, C., Zhao, J., Wang, H., Qiu, L., Ni, D., Zhang, Y., 2008. A novel C1q-domain-containing protein from Zhikong scallop *Chlamys farreri* with lipopolysaccharide binding activity. Fish Shellfish Immunol. 25, 281-9.

Developmental and Comparative Immunology 36 (2012) 390–399

Contents lists available at SciVerse ScienceDirect

Developmental and Comparative Immunology

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

Big defensins and mytimacins, new AMP families of the Mediterranean mussel *Mytilus galloprovincialis*

Marco Gerdol [a], Gianluca De Moro [a], Chiara Manfrin [a], Paola Venier [b], Alberto Pallavicini [a,*]

[a] Laboratorio di Genetica, Department of Life Sciences, University of Trieste, Via Licio Giorgeri 5, 34126 Trieste, Italy
[b] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

Big defensins and mytimacins, new AMP families of the Mediterranean mussel *Mytilus galloprovincialis*

Marco Gerdol[a], Gianluca De Moro[a], Chiara Manfrin[a], Paola Venier[b], Alberto Pallavicini[a]

[a] Department of Life Sciences, University of Trieste, Trieste, Italy

[b] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

Developmental & Comparative Immunology **36**(2): 390-399.

I peptidi antimicrobici (AMP) giocano un ruolo fondomanetale nell'immunità innata degli invertebrati, prevenendo l'invasione da parte di patogeni. I mitili esprimono un'impressionante gamma di peptidi antimicrobici ricchi in cisteine, appartenenti alle famiglie delle defensine, mitiline, miticine e mitimicine, in modo particolare negli emociti circolanti.
In questo lavoro, partendo dai dati collezionati dall'RNA-seq di *M. galloprovincialis*, sono state identificate due nuove famiglie di AMP, le big defensine e le mitimacine, la cui diversità molecolare ed espressione costitutiva sono discusse. Più nel dettaglio, sono state identificati cinque trascritti appartenenti alla famiglia delle macine (chiamate mitimacine) ed otto appartenenti alla famiglia delle big defensine (chiamate MgBD). I peptidi antimicrobici predetti mostrano un peptide segnale N-terminale, una carica netta positiva ed un elevato contenuto in cisteine, putativamente organizzate in complessi array di ponti disolfuro intramolecolari. Le mitimacine e le big defensine rappresentano quindi due nuova famiglie di AMP in *M. galloprovincialis*, che estendono ulteriormente il già ricco repertorio di AMP ricchi in cisteine dei molluschi bivalvi.
I nuovi peptidi antimicrobici sono stati caratterizzati per tessuto-specificità, dimostrando che, al contrario degli altri AMP descritti finora in mitilo, non sono espressi costitutivamente negli emociti, ma al contrario si ritrovano in altri tessuti, a volte in modo piuttosto uniforme, altre in modo più localizzato.
Nonostante sia le macine che le big defensine siano già state descritte in numerosi organismi, inclusi molluschi bivalvi (anche se soltanto in pochissimi casi), questo studio è stato il primo a descrivere questi AMP come famiglie multigeniche. Più in dettaglio, risulta di particolare interesse

la contemporanea presenza in uno stesso organismo di macine con 8, 10 e 12 cisteine, impegnate in 4, 5 e 6 ponti disolfuro rispettivamente. Nonostante ulteriori studi siano necessari per chiarire, sia a livello strutturale, sia a livello funzionale, gli effetti di un numero variabile di questi legami intramolecolari, è risultato chiaro da un'analisi comparata basata su data mining dei trascrittomi disponibili di altri invertebrati, che le due forme a 8 e 10 cisteine si ritrovano in molti gruppi filogeneticamente distanti. Proprio questa ampia distribuzione tassonomica delle macine (dai poriferi agli echinodermi) è particolarmente peculiare per un peptide antimicrobico, dal momento che nella maggior parte dei casi un rapido tasso di evoluzione molecolare è una delle caratteristiche che contraddistingue gli AMP. Tuttavia un'analisi filogenetica ha messo in luce come la variabilità delle sequenze primarie delle macine sia molto elevata, a fronte di una conservazione della struttura terziaria e della posizione delle cariche positive superficiali molto elevata.

D'altra parte le big defensine rappresentano un caso egualmente interessante, visto che una simile analisi di data mining ha evidenziato come questi AMP siano presenti esclusivamente nei molluschi bivalvi (ma non nelle altre classi di questo phylum), nei Merostomati (ma non in altre classi di artropodi) e nei Cefalocordati, tre gruppi tassonomici lontanamente evolutivamente distanti. L'alto grado di conservazione strutturale suggerisce un'origine comune per le big defensine ed allo stesso tempo implica che i geni codificanti questi AMP siano stati perduti nella maggior parte dei phylum invertebrati e mantenuti selettivamente in alcuni rari casi, inclusi i molluschi bivalvi.

TESTO ORIGINALE

Abstract

Antimicrobial peptides (AMPs) play a fundamental role in the innate immunity of invertebrates, preventing the invasion of potential pathogens. Mussels can express a surprising abundance of cysteine-rich AMPs pertaining to the defensin, myticin, mytilin and mytimycin families, particularly in the circulating hemocytes.

Based on deep RNA sequencing of *M. galloprovincialis*, we describe the identification, molecular diversity and constitutive expression in different tissues of five novel transcripts pertaining to the macin family (named mytimacins) and eight novel transcripts pertaining to the big defensins family (named MgBDs). The predicted antimicrobial peptides exhibit a N-terminal signal peptide, a positive net charge and a high content in cysteines, allegedly organized in intramolecular disulfide bridges. Mytimacins and big defensins therefore represent two novel AMP families of *M. galloprovincialis* which extend the repertoire of cysteine-rich AMPs in this bivalve mollusk.

Keywords: *Mytilus galloprovincialis*; innate immunity; antimicrobial peptides; big defensin; mytimacin.

Abbreviations: AMPs, antimicrobial peptides; MgBDs, *Mytilus galloprovincialis*, big defensins.

1. Introduction

Antimicrobial peptides (AMPs) are humoral components of the innate immunity, present in all metazoans and essential to the immediate defense reactions of invertebrate organisms lacking adaptive immunity. Antibacterial activity was first reported in mollusks in the '80s (Kubota et al., 1985) whereas the isolation and characterization of true AMPs from the mussels *Mytilus galloprovincialis* (Hubert, 1996) and *Mytilus edulis* (Charlet et al., 1996) date back to 1996.

In the Mediterranean mussel *M. galloprovincialis*, cysteine-rich antimicrobial peptides are produced as precursor molecules and processed into mature peptides within the hemocyte granules (Mitta et al., 2000c). All the four AMP classes described so far in mussels, namely defensins (Hubert, 1996; Mitta et al., 2000a; Mitta et al., 1999b), myticins (Mitta et al., 1999a; Pallavicini et al., 2008), mytilins (Mitta et al., 2000a; Mitta et al., 2000b; Roch et al., 2008) and the strictly antifungal mytimycins (Charlet et al., 1996; Sonthi et al., 2011), retain a cysteine array essential to stabilize the mature peptide in a highly compact, cationic and amphipatic structure (Mitta et al., 2000c; Yeaman and Yount, 2007). More in detail, eight cysteine residues defining four intra-molecular disulfide bridges are present in defensins, myticins and mytilins, whereas 12 cysteines and two additional disulfide bridge characterize mytimycins. The structures of mussel defensin (Yang et al., 2000) and mytilin (Roch et al., 2008) have been determined by NMR, confirming the expected pattern of intra-molecular disulfide bonds.

Each of the above mentioned AMP classes comprises several members and the recent identification of 12 additional sequence transcripts sensibly extended the number of mussel AMPs in *M. galloprovincialis* (Venier et al., 2011). New massive sequencing of the *M. galloprovincialis* transcriptome allowed us to prepare a high-coverage transcript collection and to study identity and molecular variability of two classes of previously uncharacterized cysteine-rich mussel AMPs, namely big defensins (MgBDs) and mytimacins.

Big defensins have been originally identified in the horseshoe crab *Tachypleus tridentatus* (Saito et al., 1995), specifically stored in granules within hemocytes (Kawabata and Iwanaga, 1997) likewise many molluscan AMPs (Mitta et al., 2000c). The structure of big defensins typically includes one N-terminal highly hydrophobic region, one C-terminal cysteine-rich and positively charged region, and six cysteine residues arranged to form 1-5, 2-4, 3-6 disulfide bonds in the mature peptide (Saito et al., 1995), in a similar fashion to mammalian β-defensins (Kouno et al., 2008; Selsted et al., 1993; Zhao et al., 2010). The disulfide array is therefore different from the classic 1-4, 2-5, 3-6 cysteines arrangement of arthropod defensins (Dimarcq et al., 1998). Furthermore the cysteine-stabilized a-helix and b-sheet (CSαβ) motif characterizing many plant and invertebrate defensins (including those of mussel) (Cornet et al., 1995) cannot be observed in big defensins.

The two terminal regions of the molecule display remarkable differences in antimicrobial properties, with the N-terminal fragment being more active towards Gram- bacteria and the C-terminal fragment being more effective against Gram+ bacteria (Saito et al., 1995). NMR-based studies indicated that a globular N-terminal hydrophobic domain plays a fundamental role in the dynamic interaction with target membranes (Kouno et al., 2009). To date only two other big defensins have been extensively studied: AiBD of the bay scallop *Argopecten irradians* and VpBD of the clam *Ruditapes philippinarum* were significantly up-regulated in the bivalve hemocytes in response to bacterial challenges and both displayed a broad spectrum of antimicrobial activity (Zhao et al., 2010; Zhao et al., 2007). Transcripts encoding big defensins have been also identified

in the mollusks *Crassostrea gigas*, *Mytilus chilensis* and *Bathymodiolus azoricus* and in the lancelets *Branchiostoma belcheri tsingtauense* and *Branchiostoma floridae*, suggesting a broader taxonomic distribution of this AMP class.

Macins are positively charged secreted peptides which have been first described in the annelids *Theromyzon tessulatum* (Tasiemski et al., 2004) and *Hirudo medicinalis* (Schikorski et al., 2008) and have been later identified in the cnidarian *Hydra magnipapillata* (Jung et al., 2009) and in the mollusk *Hyriopsis cumingii* (Xu et al., 2010). Macins are characterized by a disulfide array of 8 cysteines, with the optional presence of a fifth intra-molecular disulfide bridge involving a C-terminal sequence extension in theromacin. The structure of hydramacin has been determined by NMR, revealing a compact organization with an uneven distribution of positively charged residues which divide the molecular surface into two large hydrophobic hemispheres, characterized by the arrangement of cysteine bonds in a knottin fold, found in all the proteins pertaining the scorpion-toxin-like superfamily members, including mussel defensins (Jung et al., 2009).

Contrary to the majority of cysteine-rich AMPs, macins are not specifically expressed in the circulating cells, being instead localized in the endodermal epithelium (Bosch et al., 2009) or peripheral Large Fat Cells (LFCs) functionally resembling the insect fat body and often in close contact with the coelomic cavity (Tasiemski et al., 2004) or, in the case of neuromacin, in the central nervous system (Schikorski et al., 2008). The only reported exception is represented by the freshwater pearl oyster H. cumingii theromacin-like protein, which was found to be preferentially expressed in hemocytes (Xu et al., 2010). Macin expression is induced after exposure to bacteria (Tasiemski et al., 2004; Xu et al., 2010), and neuromacin localizes especially at the site of tissue injury (Schikorski et al., 2008). Increased expression of a theromacin-like transcript was also observed in response to both infection and tissue injury in the snail *Biomphalaria glabrata* (Ittiprasert et al., 2010).

Macins display membrane aggregating and permeabilizing activity, effective against Gram+ bacteria in theromacin and neuromacin (Schikorski et al., 2008; Tasiemski et al., 2004) and against Gram- bacteria in hydramacin (Jung et al., 2009). On the basis of the tertiary structure of hydramacin determined by NMR, a mechanistic model postulates its interaction with the bacterial membranes, with cell aggregation and microbe morphology changes preceding full permeabilization and their effective killing (Jung et al., 2009).

Although both macins and big defensins have already been reported in mollusks, the knowledge of these two AMP families is still extremely limited and their occurrence and evolutionary relationship in the animal kingdom have not been adequately studied. Here we report the identification in *M. galloprovincialis*, thanks to a whole-transcriptome sequencing approach, of novel transcripts pertaining to the macin family (named mytimacins) and to the big defensins family (named MgBDs) and discuss their molecular diversity and constitutive expression in different tissues.

2. Materials and methods

2.1. Identification of transcripts encoding macins and big defensins from *M. galloprovincialis*

Using second generation sequencing systems (454 Life Sciences and Illumina platforms) we sequenced the transcriptome of Mediterranean mussels (*M. galloprovincialis*) from tissues (hemocytes, gills and digestive gland) of different individuals. Following accurate processing, we could locally assemble a transcript collection which updates and enrich the pre-existing Mytibase (http://mussel.cribi.unipd.it) (Venier et al., 2009). The predicted peptides originated from the assembly process were scanned with HMMER 3 (http://hmmer.janelia.org/) to find mussel transcripts matching the big defensin and macin profiles generated by multiple alignments of the GenBank sequences pertaining to these two AMP classes. Significant hits were cut-off at e-values $<10^{-5}$. The search was re-iterated by including the new results into the alignment and by generating new profiles until no new hits were found.

2.2. dbEST data mining

A similar iterative approach was applied to the NCBI dbEST database (http://www.ncbi.nlm.nih.gov/dbEST) in order to extend the search from Mytilus spp. to other organisms and to assess the taxonomic distribution of the two AMP classes. The EST sequences matching the above mentioned HMMER profiles were assembled into contigs with the CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark) to remove redundancy. Only complete sequences were considered for further analysis.

2.3. Sequence analysis

All transcript sequences related to macins and big defensins were translated with the Expasy Translate tool (http://expasy.org/tools/dna.html) to obtain the virtual encoded peptides: signal peptides were predicted using SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP), isoelectric point and molecular weight were calculated with the Expasy Compute pI\MW tool (http://expasy.org/tools/pi_tool.html) and functional role was evaluated with the antimicrobial peptide predictor APD2 (http://aps.unmc.edu/AP/prediction/prediction_main.php). Structural homology with the *T. tridentatus* big defensin (2RNG) and the *H. magnipapillata* hydramacin-1 (2K35) was evaluated by automated tridimensional modeling with Phyre2 (Kelley and Sternberg, 2009).

An analysis of SNPs frequency in the transcript sequences of mytimacins and big defensins was performed with the CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark). To exclude potential sequencing errors, sites with low-coverage (less than 10 sequencing reads) were not analyzed and SNPs occurring with very low frequency (<2%) or not covered by at least 3 independent reads were not considered reliable.

2.4. Phylogenetic analysis

Multiple sequence alignments displayed in figures and those used for the generation of HMMER profiles and Bayesian phylogenetic analysis were produced with MEGA5.02 (Tamura et al., 2011) using the MUSCLE algorithm (Edgar, 2004), with gap opening and extension penalties of -2 and -1, respectively.

Phylogenies of big defensins and macins were estimated with MrBayes 3.2 (Ronquist and Huelsenbeck, 2003) starting from an alignment of the entire mature predicted peptides. The GTR substitution model of molecular evolution with a proportion of invariable sites, and a Gamma-shaped distribution of rates across sites (GTR + γ + I), was chosen as the best-fitting model for our datasets with ProtTest (Abascal et al., 2005). We ran two independent analyses with four chains each (one "cold, three "warm") for 1,000,000 generations, sampling a single tree each 1,000 generations. The first 25% of the generated trees were discarded for the burn-in procedure and the remaining trees were used to calculate the posterior probability for each node in a 50% consensus trees.

2.5. Mussel samples

To evaluate the tissue-specific expression of Mytimacin-1, Mytimacin-2, Mytimacin-3, MgBD1, MgBD3 and MgBD6 transcripts, mussels of 6.5-7 cm shell length were collected from the Gulf of Trieste, Italy. Total RNA was individually purified from hemolymph and from digestive gland mantle, posterior abductor muscle, gill, and foot, previously homogenized in RNATidy G according to the manufacturer's instructions (AppliChem, Darmstadt, Germany). Following extraction, the RNA quality was assessed by electrophoresis on denaturing agarose gel and its quantity was estimated by UV-spectrophotometry. Complementary DNA was prepared by retro-transcription with the iScript™cDNA Synthesis Kit (Bio-Rad) from pooled RNA samples representing five individuals.

2.6. Quantitative PCR expression analysis

The expression levels of the Mytimacin-1, Mytimacin-2, Mytimacin-3, MgBD1, MgBD3 and MgBD6 transcripts were assessed in samples representing hemolymph, digestive gland, mantle, posterior abductor muscle, gills and foot of five adult mussels. Primer pairs were designed (Table 1) and used to obtain specific PCR amplicons. The primers for MgBD3 were specifically designed to co-amplify the three sequences MgBD3a, MgBD3b and MgBD3c.

All the PCR assays were performed using a Bio-Rad CFX96 system. The 15 µL reaction mix included 7.5 µl of 2X IQ™ SYBR Green® Supermix (Biorad), 0.3 µl of each 10 µM primer and 2 µL of a 1:10 cDNA dilution. The following thermal profile was used: an initial 3' denaturation step at 95°C, followed by 40 cycles at 95° for 20", 60° for 15" and 72° for 20". Amplification products were analyzed with a 65°/95°C melting curve. The expression levels of the selected transcripts were determined using the comparative Ct method (2-ΔΔCt method) (Livak, 2001). Ct values used for quantification were corrected based on PCR efficiencies using LinRegPCR (Ramakers et al., 2003). The expression values were normalized using the elongation factor EF-1 as housekeeping gene (EF-1 primers are shown in Table 1). Results are given as the mean with standard deviation of three technical replicates.

3. Results and discussion

3.1. Big defensins

3.1.1. Computational identification and sequence features of MgBDs

The mussel transcriptomic collection, assembled starting from a total of 24901 Sanger, 150857 454 Life Sciences and 108620377 Illumina sequencing reads, conprises110259 contigs (with an average length of 590 nucleotides; the N50 parameter of the assembly was 658). In this transcriptomic mussel collection we could identify 8 different sequences encoding big defensins, named MgBD 1-6 and deposited at EMBL under the accession IDs FR873266-FR873273. Three sequences showing remarkable similarity are indicated as MgBD3a, MgBD3b and MgBD3c.

The 8 inter-related sequences differ quite widely for their representation in the transcript collection, with MgBD1 showing a very high coverage in respect with the average of mussel transcripts and MgBD3b and MgBD3c showing, in contrast, an extremely low relative abundance (Table 2). An open reading frame (ORF) encoding the full-length peptide precursor was identified in all the eight different nucleotidic sequences (the alignment of the full length precursor peptides is shown in Figure 1).

The virtual translation yielded aminoacid sequences ranging from 114 and 122 residues in lenght (the shortest being MgBD3b and the longest one being MgBD5). A short N-terminal signal peptide was predicted in all cases, and the alignment with the big defensin isolated from *T. tridentatus* revealed that MgBDs are produced as prepropeptides. Following the cleavage of the signal peptide, a second proteolytic cleavage could result in the mature peptides, whose molecular weight range from 8.64 to 9.70 KDa.

In the C-terminal region of all big defensin transcripts, six conserved cisteines define the motif C-X6-C-X3-C-X13-C-X4-C-C, essential for the disulfide bridge formation (see Figure 1). Six out of the eight predicted mature peptides show a basic isoelectric point (8.3-9.6) meaning a a positive net charge at neutral pH whereas the isolectric point of MgBD1 and MgBD6 is closer to neutrality. As reported in Table 2, the prediction of antimicrobial features performed with the APD2 software also revealed a rather high percentage of hydrophobic residues, comparable to those of mytilins and defensins (Roch et al., 2008).

The tertiary structure modeling by Phyre 2 was successful for all the 8 MgBDs and a percentage of residues ranging from 88 (MgBD3a) to 96% (MgBD4) were modeled with >90% confidence based on the tertiary structure of the *T. tridentatus* big defensin (PDB accession: 2RNG), denoting a high structural conservation within this AMP family (see Figure 2).

Overall, the analysis of the 8 MgBDs sequences highlights properties common to many AMPs, such as a basic isoelectric point and a high hydrophobicity ratio. Furthermore, the conserved cysteine array, N-terminal hydrophobic domain and high sequence and predicted structural similarity with big defensins previously characterized in other organism further suggest that they represent genuine big defensins of *M. galloprovincialis*.

3.1.2. Constitutive tissue expression of MgBDs

The expression analysis of MgBD1, MgBD3 (with primers co-amplifying the 3 isoforms MgBD3a, MgBD3b and MgBD3c) and MgBD6 revealed very low or negligible constitutive expression in most of the six tissues analyzed but each transcript resulted to be selectively expressed in a given

tissue (Figure 3). MgBD1, the sequence represented with the highest sequence coverage in our collection (see Table 2) was expressed only in the digestive gland whereas MgBD3 and MgBD6 expression was mainly traced in gills and mantle, respectively (Figure 3). Almost no expression was evident in hemolymph for any MgBD. These data are somewhat surprising since the few big defensins described so far have been isolated in hemocytes, likewise the other known mussel AMPs (defensins, mytilins, myticins and mytimycins). Nevertheless the knowledge about big defensins expression pattern in mollusks is still deficient, as it has only been investigated in AiBD (evidencing specificity to hemocytes and, to a lesser extent, to gills (Zhao et al., 2007)), whereas VpBD was isolated from hemocytes, but its expression in other tissues was not assessed (Zhao et al., 2010). The tissue-specific expression of MgBD1 and other MgBDs would indicate their involvement in localized protection towards invading pathogens. Further studies should point out whether any MgBDs display a positive regulation of expression in response to immune-stimulating challenges, likewise VpBD and AiBD (Zhao et al., 2010; Zhao et al., 2007).

3.1.3. Evolution of big defensins in animals

The Bayesian phylogenetic analysis (Figure 4) grouped the 8 MgBDs in a highly supported monophyletic clade together with a close relative to MgBD3a from *M. chilensis* and the 3 big defensins of *C. gigas*. This clade is well separated from the other molluscan big defensins from *B. azoricus*, *A. irradians* and *R. philippinarum*.

The 3 MgBD subgroups MgBD1\4, MgBD2\6 and MgBD3a\b\c underline the close similarity of their amino acidic sequences. In particular, MgBD3a, MgBD3b and MgBD3c are almost identical in the N-terminal hydrophobic region and in the C-terminal region with the cysteine array (high similarity is also retained at nucleotidic level), but diverge substantially in the region bridging the two portions (see Figure 1) and, as expected, in the UTRs.

The dbEST data mining revealed close relatives to MgBD2, MgBD5 and MgBD6 in M. californianus in addition to the M. chilensis MgBD3a ortholog (see Figure 5 for details), and several big defensin sequences in other bivalves (combined data from GenBank and dbEST permitted to identify big defensins in 11 different bivalve species besides *M. galloprovincialis*), but the overall taxonomic distribution of this AMP family seems to be strictly restricted to bivalve mollusks, horseshoe crabs, and amphioxus, whereas no big defensins were detected in many other large invertebrate classes. Such a distribution is quite unusual, as mollusks and horseshoe crabs (phylum Arthropoda, subphylum Chelicerata, class Merostomata) are distantly related and no big defensins could be identified neither in other large Arthropoda subgroups (crustaceans, insects, etc.) or Lophotrocozoans. Nevertheless, the presence of big defensins in early chordates like amphioxus suggests a broader taxonomical distribution for this AMP class.

As the search for homologous sequences was exclusively based on ESTs, one could argue that the absence of big defensins in the classes bridging mollusks, horsheshoe crabs and lancelets could be merely the result of a either very low or of a very specialized tissue expression (as the case of MgBD1, MgBD3 and MgBD6 in *M. galloprovincialis*, see section 3.1.2), not sufficient to guarantee an homogenous representation of these sequences in transcriptomic sequencing projects based on Sanger sequencing. In order to test this hypothesis, we used a similar approach to the genomic data available for any representative species of the main invertebrate families available, revealing that the lack of big defensin-like sequences in dbEST effectively depends on the absence of these gene models in most genomes. Therefore, such a taxonomic distribution would imply gene loss in some

invertebrate classes and selective retention of big defensin genes in other classes. Retention and expansion of genes encoding big defensins in the Mediterranean mussel could explain the evidence of various MgBD transcripts as products of different genetic loci, likewise the previously characterized AMP families of defensins, mytilins, myticins and mytimycins.
A lower-scale diversity at a SNP level is still detectable in our new sequencing data, since the processed transcript sequences derived from many individuals of *M. galloprovincialis*, although most SNPs are located in the UTR regions and therefore don't cause amino acid substitutions. Only the availability of complete genomic data from mussel will reveal whether the diversity observed is the product of inter-individual variability or rather the result of highly similar paralogs, likewise other invertebrate AMPs, such as oyster and tick defensins (Schmitt et al., 2010; Wang and Zhu).

3.2 Mytimacins

3.2.1. Computational identification and sequence features of mytimacins

Five different transcripts encoding macins, named mytimacin-1-5, were identified as reported for the MgBDs. Nucleotidic sequences were deposited at EMBL under the accession IDs FR873274-FR873278. Sequence representation in the mussel transcript collection was variable, with mytimacin-1 showing the highest coverage and mytimacin-5, displaying the lowest one (see Table 3).
An open reading frame (ORF) encoding a full-length peptide was identified in four out of five nucleotidic sequences. Mytimacin-5 appears incomplete, since no stop codon was identified at the 3'end of the sequence; its full length can nevertheless be predicted to be 105 amino acids by the comparison with the *M. edulis* homologue (EST AM879320.1, see Figure 5). The multiple alignment of the deduced amino acidic sequences of mytimacins is shown in Figure 1.
The predicted mytimacins are characterized by a length of 85-101 residues (the complete sequence of mytimacin-5 is unavailable). A N-terminal signal peptide was predicted with high probability, suggesting that mytimacins are produced as precursors targeted to the secretory pathway. Predicted molecular weights of mature peptides range between 6.79 and 9.17 KDa. The eight cysteine residues, arranged in four intramolecular disulfide bridges characterizing all macins, are conserved also in mytimacins. The two additional cysteines engaged in the fifth, optional, disulfide bond typical of the longer, theromacin-like AMPs, can be identified only in mytimacin-1, -4 and -5, which present, indeed, a remarkable extension at their C-terminus, likewise the theromacins of the segmented worm *T. tessulatum* (Tasiemski et al., 2004) and of the mollusk *H. cumingii* (Xu et al., 2010). On the contrary, mytimacin-2 and -3 lack this portion, therefore more closely structurally resembling hydramacin (Jung et al., 2009) and neuromacin (Schikorski et al., 2008). The mytimacin-2 sequence is nevertheless substantially different from that of mytimacin-3, as it is characterized by a peculiar, potentially highly flexible, glycine-rich stretch at the N-terminus of the mature peptide, which cannot be observed in any other macin reported so far.
The predicted mature peptides of mytimacins are characterized by basic isoelectric points, thus carrying a positive net charge at neutral pH. Their analysis performed with the antimicrobial peptide predictor APD2 revealed additional typical AMPs characteristics, i.e. a positive net charge and a rather high percentage of hydrophobic residues, which is also in this case comparable to those of other mussel AMPs (Roch et al., 2008). The main features of mytimacins are detailed in Table 3.

The tertiary structure modeling by Phyre 2 based on the tertiary structure of hydramacin-1 (Protein database accession: 2K35) was successful for all the 5 mytimacins and a high percentage of residues were modeled with >90% confidence (the lowest one being mytimacin-5, with 69%). In particular, the predicted tertiary structure of mytimacin-3 resulted highly similar to hydramacin-1, as 97% residues were modeled with high confidence and both molecules are characterized by the presence of 8 cysteines. Figure 6 displays the highly conserved positions of lysine and arginine residues on the molecular surfaces of hydramacin-1 and mytimacin-3. The distribution of these residues, forming a positively charged "belt" dividing two hydrophobic hemispheres is postulated to be essential for the antimicrobial activity of hydramacin-1 (Jung et al., 2009), and the retention of this feature in mytimacin-3 suggests that this molecule may exert a similar mode of action. Given their cationic nature, the presence of a conserved knottin-like disulfide array and the highly significant predicted structural similarity with hidramacin-1, the 5 *M. galloprovincialis* mytimacins are likely to act as AMPs.

3.2.1 Constitutive expression of mytimacins

The expression analysis of mytimacin-1 highlighted its constitutive expression, at comparable levels, in all the tissues analyzed, with the exception of hemocytes where the transcript expression was much lower (Figure 3). Similarly, mytimacin-2 was not expressed at all in hemocytes, whereas it showed a rather specific localization to the gills and, to a lesser extent, to the foot. Mytimacyn-3 was almost exclusively detected in the mantel, although its expression level was particularly low also in this tissue.

Our data therefore point out that mytimacins, unlike the vast majority of known molluscan AMPs, are not specifically synthesized and stored in circulating hemocytes. This is not surprising, considering that most macins are produced in highly specialized cells, called LFCs, located in tissues in contact with the coelomic cavity, with the intestinal epithelium and with the epidermis in segmented worms (Tasiemski et al., 2004), or in the secondary endoderm in *Hydra* (Bosch et al., 2009). The theromacin of the freshwater mussel *H. cumingii* represents the only reported exception, as it is mainly expressed in hemocytes (Xu et al., 2010). The completely different expression pattern of mytimacin-1 in respect with Hc theromacin is consistent with the presence of specialized producing cells evenly distributed in the whole animal body, likewise segmented worms and Hydra.

On the contrary, mytimacin-2 and -3 resulted to be expressed in specific tissues, and hypothesizing the reasons for such a specificity is particularly tricky, considering that they don't show any striking similarity with other macins which have been described so far, although mytimacin-3 could be linked to neuromacin and hydramacin, considering its similar molecular organization.

3.2.1. Evolution of macins in animals

The Bayesian phylogenetic analysis revealed that macins are highly heterogeneous sequences (Figure 7). Macins of segmented worms and cnidarians were grouped in highly supported clades, whereas molluscan macins couldn't be grouped together, but instead formed several, distantly related, subgroups, reflecting the sequence diversity we observed in *M. galloprovincialis*. While no obvious hortologues to mytimacin-2, -3 and-5 could be identified in other organisms, mytimacin-1 and -4, which share a high identity percentage at an amino acidic level (77%), are grouped in a strongly supported clade with the theromacin-like sequences of the freshwater mussels *Alasmidonta*

heterodon and *Alasmidonta varicosa*, similarly characterized by C-terminal extensions and the presence of two additional cysteines in conserved positions.
Our dbEST data mining strategy revealed orthologues sequences to mytimacin-3 and mytimacin-5 in *M. edulis* and *M. californianus* (for details, see Figure 5). More importantly, the data mining also evidenced that macins represent an ancient and widespread AMP family. Indeed, a 8-cysteines macin can be identified in the sponge *Leucetta chagoensis*, suggesting that macins were already exploited in antimicrobial defense in primitive multicellular animals. Globally, our analysis identified macins in more than 40 different species, pertaining to Cnidarians, to most of the major groups of protostomes, including Mollusca, Insecta, Arachnida, Crustacea, Nematoda, Annelida and Tardigrada, ranking up to the basal Deuterostomes *Patiria miniata* and *Asterina pectinifera* (phylum Echinodermata, class Asteroidea), although they seem to be more extensively represented in some groups (i.e. Cnidaria and Lophotrochozoa) and just sporadically in others (i.e. Ecdysozoa). A phylogenetic analysis of the whole set of sequences evidenced extremely complex relationships between the macins of different organism and was not able to shed definitive light on this topic (data not shown). The picture is made even more complex by the main structural differences observed, which can, in turn, be helpful to categorize macins into four subclasses as follows: a) short macins with 4 disulfide bridges (8-Cys macins); b) short macins with 4 disulfide bridges and a N-terminal glycine-rich stretch (8-Cys + poly-Gly macins); d) long macins with 5 disulfide bridges (10-Cys macins); e) long macins with 6 disulfide bridges (12-Cys macins).
The distribution of the four subclasses of macins in metazoans is exemplified in Figure 8. While all the four subclasses are represented in *M. galloprovincialis*, only 8-Cys and 10-Cys macins seem to be widespread in animals, whereas the diffusion of the previously uncharacterized 8-Cys + poly-Gly and 12-Cys macins seem to be restricted to a few classes only. The presence of four disulfide bonds combined with a poly-Glycine N-terminal stretch observed in mytimacin-2 has never been described before, although peptides with an astounding similarity can be detected in the distantly related phylum of Cnidaria (predicted peptides retrieved from EST data of *Clitya hemisphaerica* and *Podocoryna carnea* showed, respectively, 74% and 65% identity with mytimacin-2). The comparison of mytimacin-5 with its orthologues in *M. edulis* and *Pinctada maxima* was useful to reveal the presence of two additional cysteines in conserved position (the first one located immediately before the C4, the second one in the C-terminal extension), suggesting that they may be involved in the creation of an additional, 6th disulfide bond, adding even more structural complexity to this subclass of 12-Cys macins, which was only identified in Mollusca (see Figure 8).
Our data are consistent enough to affirm that the five mytimacins are the product of different genetic loci, revealing for the first time macins as a multi-genic family within a single species. The presence of a minor inter-individual variability was evidenced by the SNP analysis performed with the CLC Genomic Workbench 4.5.1, although the variability observed was rather low.

Conclusions

The advent of next generation sequencing technologies provided a valuable resource for the bioinformatic identification of previously uncharacterized protein families in non-model organisms on a transcriptomic or on a genomic scale (Patrzykat and Douglas, 2003). Such methodologies have also been successfully used also in the identification of potential AMPs in plants (Belarmino and Benko-Iseppon, 2010; Graham et al., 2008) and more recently also in invertebrate genomes (Tian et al., 2010; Wang and Zhu).

We chose to use a similar approach in the identification of members of two previously uncharacterized mussel AMPs families, big defensins and macins. Our analysis revealed the presence of eight novel big defensins (MgBDs) and five novel macins (mytimacins) in the transcriptome of the Mediterranean mussel *M. galloprovincialis*, which further extend the rich and complex antimicrobial peptides repertoire of this organism (Venier et al., 2011). Our data point out that most of these sequences are the products of multi-genic families, suggesting that a strategy of gene expansion similar to the one described for oyster and thick defensins (Schmitt et al., 2010; Wang and Zhu) has been implied in MgBDs and mytimacins. Furthermore, the data mining analysis revealed a widespread distribution of macins in invertebrates and, on the contrary, a very restricted distribution of big defensins to a few taxonomic classes.

Further studies should be focused on the investigation of the physiological role and the sites of synthesis and storage of these two newly discovered AMP families in mussel, as well as on their spectrum and mode of action against invading pathogens.

Acknowledgements

This work was supported by the project FP7-KBBE-2010-4-266157 (Bivalife) and by Regione Friuli Venezia Giulia, Direzione Centrale Risorse Agricole, Naturali, Forestali e Montagna, L.R. 26/2005 prot. RAF/9/7.15/47174.

TABLES

Primer name	Primer sequence
EF-1 FOR	cctcccaccatcaagaccta
EF-1 REV	ggctggagcaaaggtaacaa
Mytimacin-1 FOR	ctcctgcaaattcccacatc
Mytimacin-1 REV	atcttttgttccgccagaga
Mytimacin-2 FOR	gtggtggtggaagtggaagt
Mytimacin-2 REV	tccaagctctttgcatctgt
Mytimacin-3 FOR	acaatcaccaatgggaccac
Mytimacin-3 REV	tttgggcagcaaattctctc
MgBD1 FOR	gcgtagattccatatgcagca
MgBD1 REV	tgttgatactccctgctcag
MgBD3 FOR	ccgattctaggacgagttgtggca
MgBD3 REV	ggcaactttccaagcgccatatgc
MgBD6 FOR	agcatcatacgcaggattgtc
MgBD6 REV	tagctctacaccatcctctg

Table 1: Primers designed for assessing the tissue-specific levels of Mytimacin-1, Mytimacin-2, Mytimacin-3, MgBD1, MgBD3 and MgBD6 transcripts

	relative abundance*	precursor/ mature peptide length (aa)	disulfide bridges	pI of the mature peptide	MW of the mature peptide	hydrophobicity ratio
MgBD1	36.65	115/79	3	7.09	8.64	45%
MgBD2	1.31	116/82	3	9.02	8.96	40%
MgBD3a	0.05	119/85	3	9.61	9.70	40%
MgBD3b	<0.01	114/80	3	9.35	9.16	41%
MgBD3c	<0.01	118/84	3	9.47	9.66	39%
MgBD4	0.15	115/79	3	8.30	8.78	44%
MgBD5	5.65	122/87	3	8.87	9.73	42%
MgBD6	0.92	116/82	3	6.02	8.85	42%

Table 2: Sequence representation in the transcript collection and main predicted features of the big defensin peptides of *M. galloprovincialis*. * value representing the rate between the expression level of each transcript and the average expression value of all other transcripts in the whole transcript collection (measured in RPKM).

	relative abundance*	precursor/ mature peptide length (aa)	disulfide bridges	pI of the mature peptide	MWof the mature peptide	hydrophobicity ratio
Mytimacin-1	4.87	101/78	5	9.10	9.11	38%
Mytimacin-2	0.78	92/64	4	8.65	8.12	32%
Mytimacin-3	1.00	85/61	4	9.06	6.79	34%
Mytimacin-4	0.17	101/78	5	8.04	9.17	38%
Mytimacin-5	0.03	100+/78+	6?	?	?	?

Table 3: summary of mytimacins sequence coverage in Mytibase and main features of the corresponding predicted peptides. * value representing the rate between the expression level of each transcript and the average expression value of all other transcripts in the whole transcript collection (measured in RPKM).

FIGURES

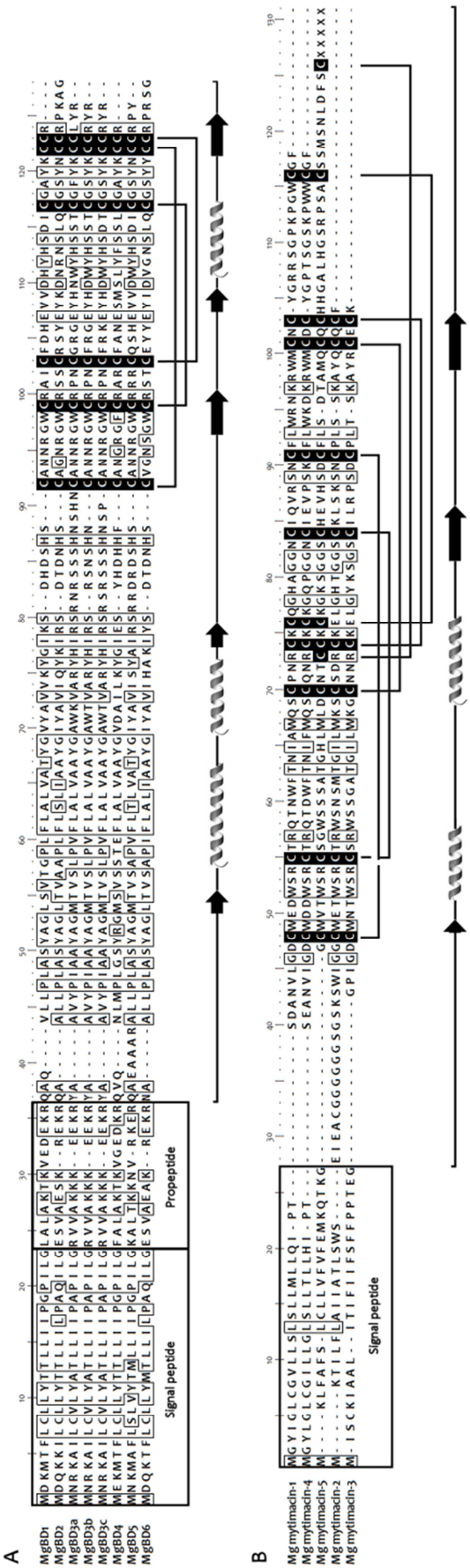
A
B
Signal peptide
Propeptide
Signal peptide
MgBD1
MgBD2
MgBD3a
MgBD3b
MgBD3c
MgBD4
MgBD5
MgBD6
Mg mytimacin-1
Mg mytimacin-4
Mg mytimacin-5
Mg mytimacin-2
Mg mytimacin-3

Figure 1: Panel A: multiple alignment of deduced amino acid sequences of the 8 mussel big defensins. Panel B: multiple alignment of deduced amino acid sequences from the five mytimacins.
Conserved residues are outlined, cysteine residues engaged in disulfide bridges are *black boxed* and the organization of the disulfide arrays are schematically shown. The signal peptide and propeptide regions are shown. The predicted secondary structure is shown below the sequence alignment (β-sheet: arrow; α-helix: helix).

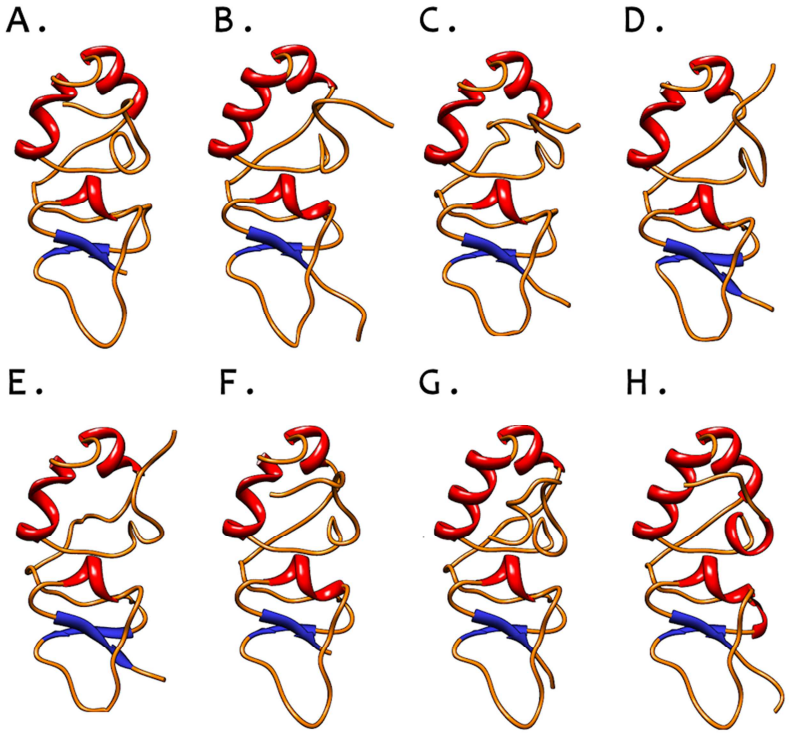

Figure 2: Predicted ribbon structures of mussel big defensins, obtained by Phyre2 modeling. A: MgBD1; B: MgBD2; C: MgBD3a; D: MgBD3b; E. MgBD3c; F: MgBD4; G: MgBD5; H: MgBD6.

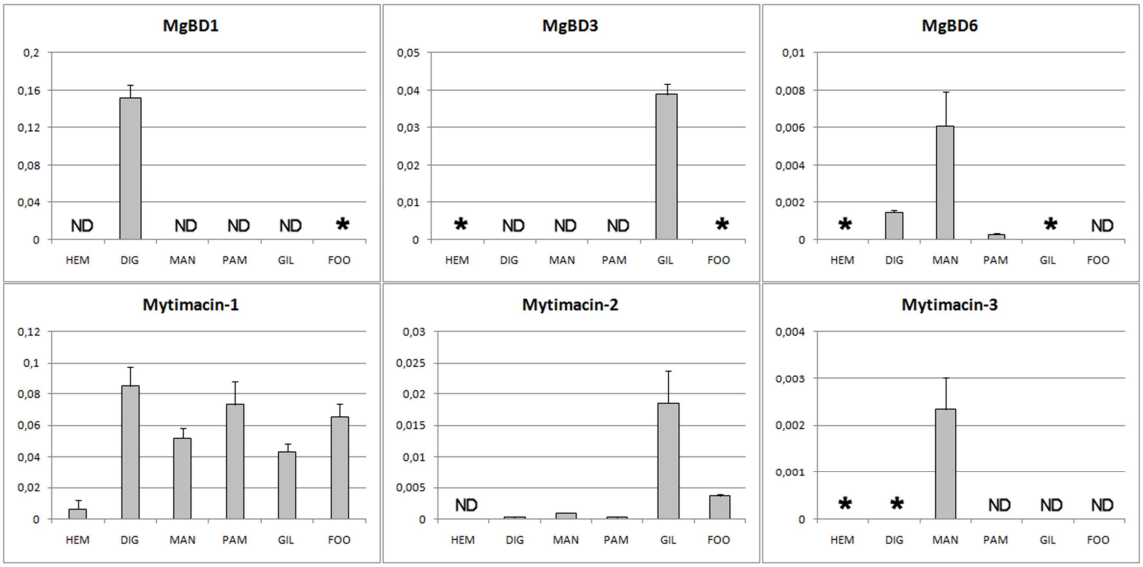

Figure 3: Tissue-specific expression of the transcripts mytimacin-1, mytimacin-2, mytimacin-3, BD1, BD3 and BD6. The expression values (bars) are relative to the elongation factor EF-1. Results are mean ± SD of 3 technical replicates. The Y axis of each graph is scaled based on the highest level of expression. HEM: Hemolymph, DIG: digestive gland, MAN: mantle, PAM: posterior abductor muscle, GIL: gills, FOO: foot. ND: not detected (fluorescence did not reach threshold after 40 cycles of PCR or the melting peak analysis did not reveal any specific product). *: not quantifiable (fluorescence did not reach threshold after 40 cycles of PCR but the melting peak analysis revealed a limited production of the specific amplicon).

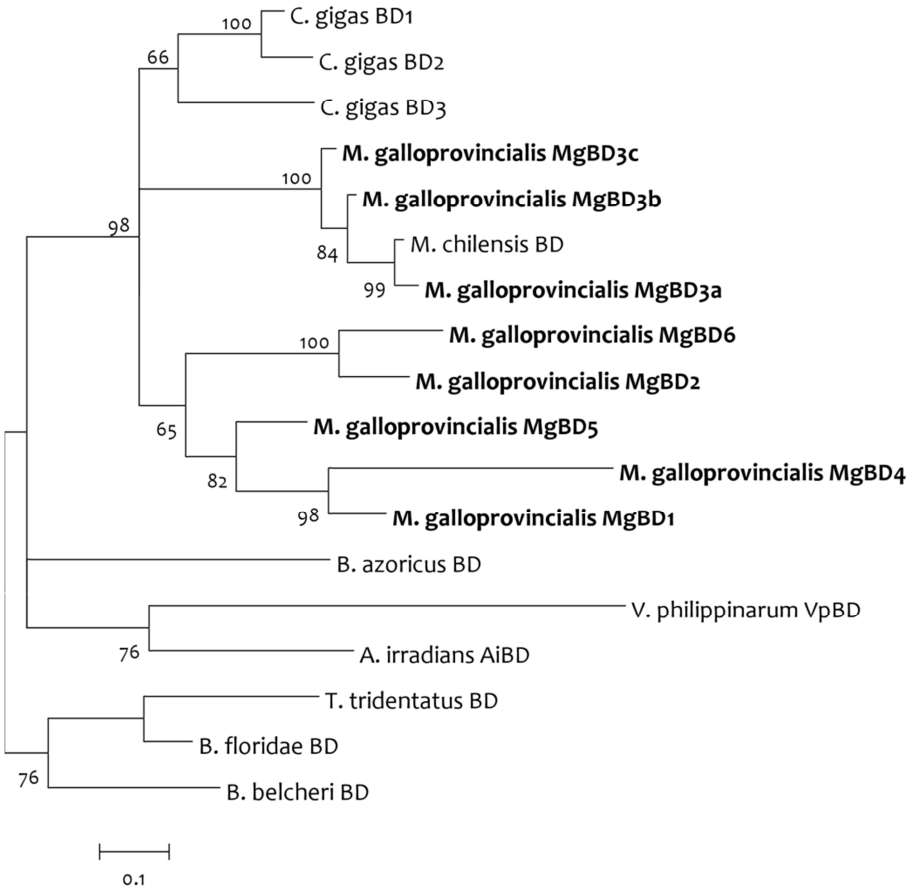

Figure 4: Bayesian phylogeny of big defensins inferred from the alignment of the predicted mature peptides. Posterior probabilities are shown for each branch. Entry IDs: *T. tridentatus* BD: P80957.2; *B. belcheri* BD: Q86QN6.1; *B. floridae* BD: ADH03419.1; *A. irradians* AiBD: Q0H293.1; *V. philippinarum* BD: ADM25826.1); *M. chilensis* BD: AEE60906.1; *B. azoricus* BD: HM756150.1; *C. gigas* BD1, BD2 and BD3: AEE92785.1, AEE92787.1 and AEE92790.1.

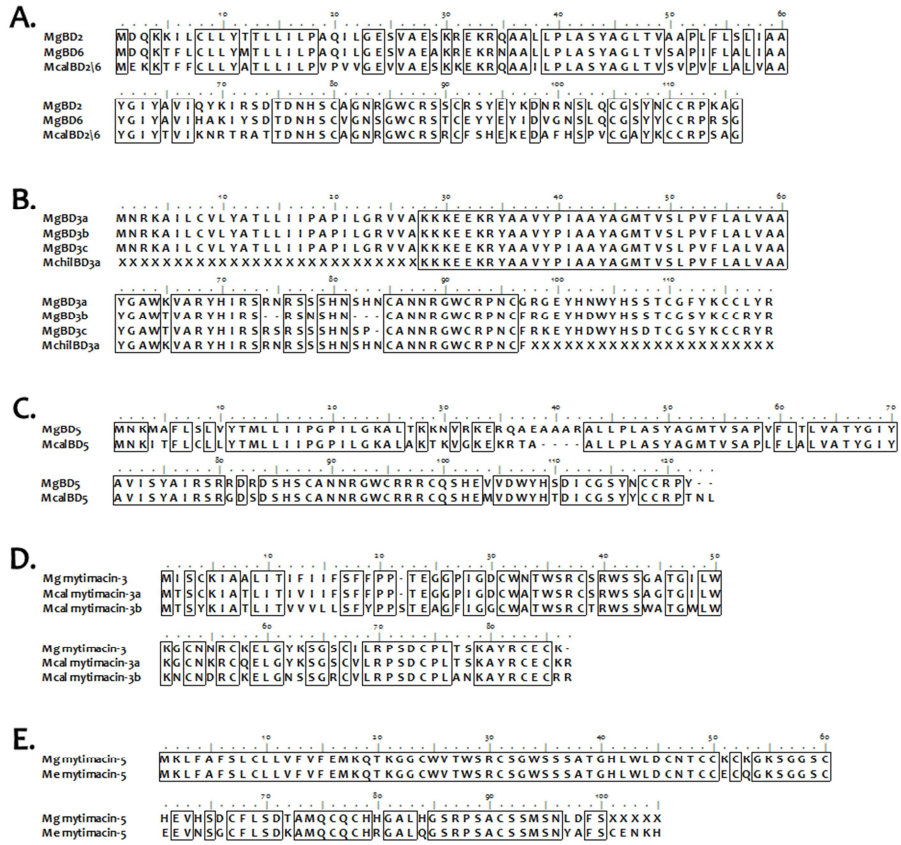

Figure 5: Alignment of *M. galloprovincialis* big defensins and mytimacins and their orthologs in other *Mytilus* species. Peptides sequences were inferred from the following ESTs: *Mytilus californianus* BD2\6: GE761911.1, GE763207.1, GE764803.1, GE756683.1, GE749104.1, GE753537.1; *Mytilus chilensis* BD3a: AEE60906.1; *Mytilus californianus* BD5: ES398618.1, GE759807.1, GE760702.1; *Mytilus californianus* mytimacin-3a: GE754022.1, GE749772.1, GE747980.1, GE749598.1; *Mytilus californianus* mytimacin-3b: GE752669.1, GE750343.1; *Mytilus edulis* mytimacin-5: AM879320.1.

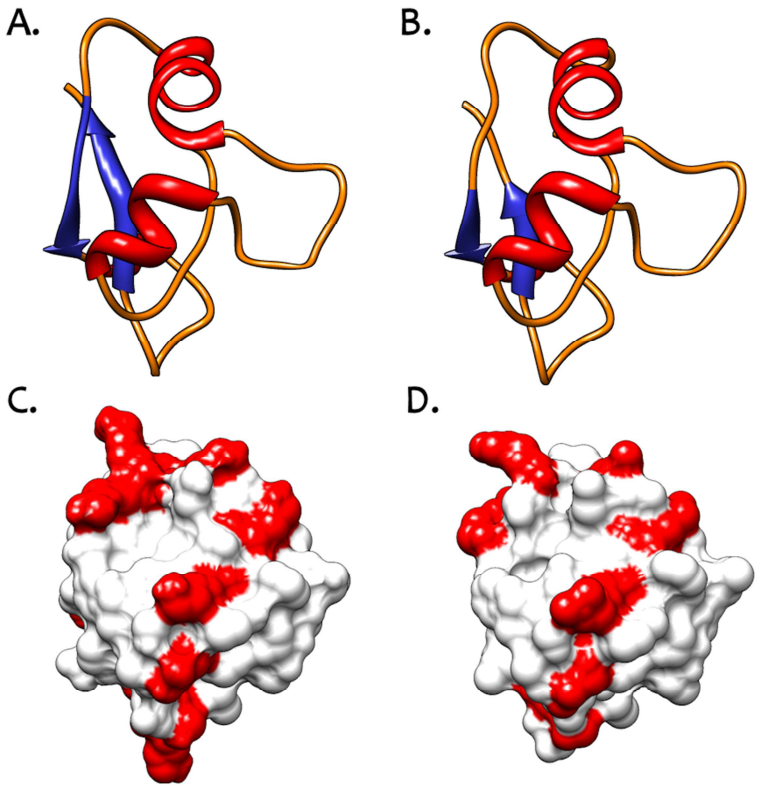

Figure 6: Ribbon structures and molecular surfaces of hydramacin-1 (panels A and C, PDB accession: 2K35) and mytimacin-3 (panels B and D, obtained by Phyre 2 modeling). Positive charges of lysine and arginine are colored, highlighting the high conservation of the positively charged residues distribution.

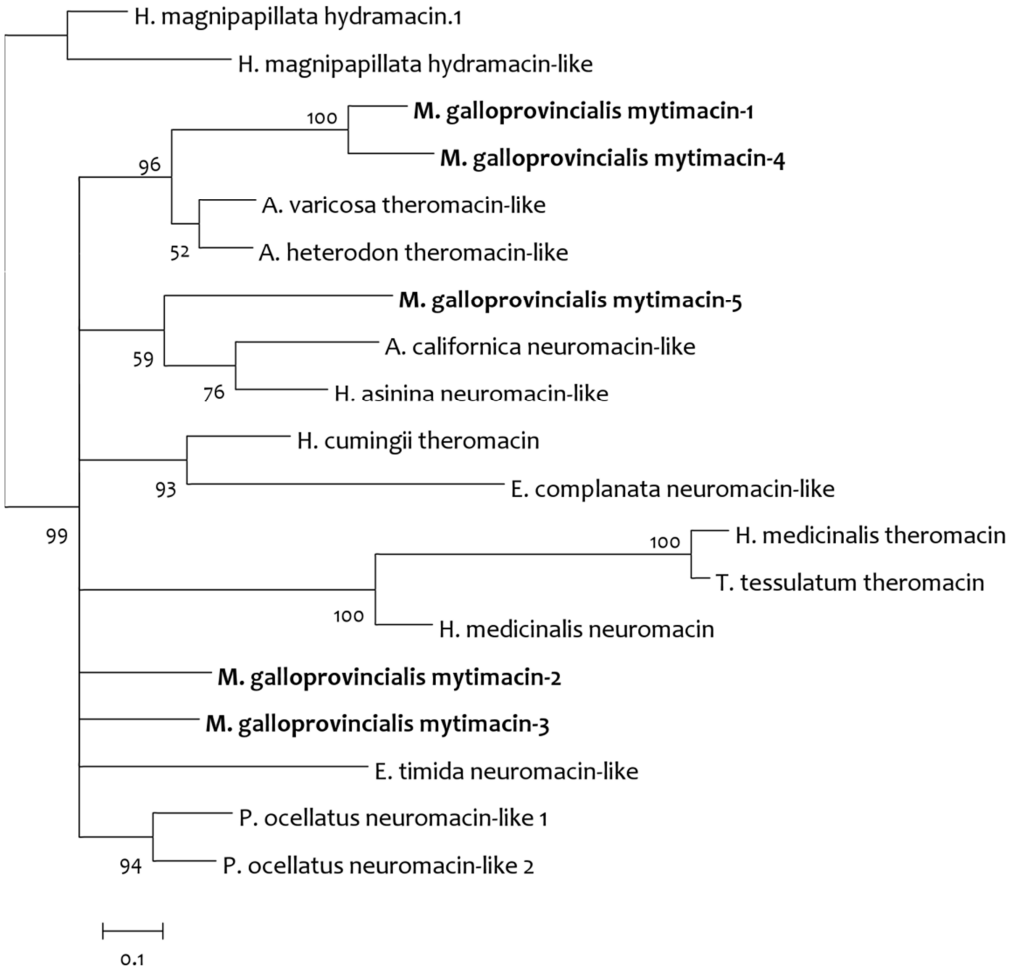

Figure 7: Bayesian phylogeny of macins inferred from the alignment of the predicted mature peptides. Posterior probabilities are shown for each branch.
(Entry IDs: *H. cumingii* theromacin: ADK94899.1; A. *varicosa* theromacin-like: HP640944.1; *A. heterodon* theromacin-like: HP640617.1; *H. medicinalis* theromacin: ABV56207.1; *T. tessulatum* theromacin: Q6T6C2.1; *A. californica* neuromacin-like: A5GZY1.1; *E. complanata* neuromacin-like: HP640944.1; *H. asinina* neuromacin-like: EZ420620.1; *E. timida* neuromacin-like: HP157026.1; *H. magnipapillata* hydramacin: B3RFR8.1; *H. magnipapillata* hydramacin-like: XP_002163468.1; *H. medicinalis* neuromacin: A8V0B3.1; *P. ocellatus* neuromacin-like 1: HP232903.1; *P. ocellatus* neuromacin-like 2: HP231655.1).

TAXONOMIC RANKS					8-Cys	8-Cys + poly-Gly	10-Cys	12-Cys
Parazoa				Porifera	■			
Eumetazoa	Radiata			Cnidaria	■	■	■	
	Bilateralia	Protostomia	Ecdysozoa	Nematoda	■		■	
				Tardigrada	■			
				Arthropoda	■		■	
			Lophotrochozoa	*Mytilus galloprovincialis*	■	■	■	■
				other Mollusca	■		■	■
				Annelida	■		■	
		Deuterostomia		Echinodermata	■		■	

Figure 8: taxonomic distribution of macins, inferred from the dbEST data mining analysis. 8-cys: short macins with 4 disulfide bridges; 8-Cys + poly-Gly: short macins with 4 disulfide bridges and a N-terminal glycine-rich stretch; 10-Cys: long macins with 5 disulfide bridges; 12-Cys: long macins with 6 disulfide bridges.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-2105.

Belarmino, L.C., Benko-Iseppon, A.M., 2010. Data bank based mining on the track of antimicrobial weapons in plant genomes. Current Protein and Peptide Science 11, 195-198.

Bosch, T.C.G., Augustin, R., Anton-Erxleben, F., Fraune, S., Hemmrich, G., Zill, H., Rosenstiel, P., Jacobs, G., Schreiber, S., Leippe, M., Stanisak, M., Grötzinger, J., Jung, S., Podschun, R., Bartels, J., Harder, J., Schröder, J.M., 2009. Uncovering the evolutionary history of innate immunity: The simple metazoan *Hydra* uses epithelial cells for host defence. Developmental and Comparative Immunology 33, 559-569.

Charlet, M., Chernysh, S., Philippe, H., Hetru, C., Hoffmann, J.A., Bulet, P., 1996. Innate immunity: Isolation of several cysteine-rich antimicrobial peptides from the blood of a mollusc, *Mytilus edulis*. Journal of Biological Chemistry 271, 21808-21813.

Cornet, B., Bonmatin, J.-M., Hetru, C., Hoffmann, J.A., Ptak, M., Vovelle, F., 1995. Refined three-dimensional solution structure of insect defensin A. Structure (London, England : 1993) 3, 435-448.

Dimarcq, J.-L., Bulet, P., Hetru, C., Hoffmann, J., 1998. Cysteine-rich antimicrobial peptides in invertebrates. Peptide Science 47, 465-477.

Edgar, R.C., 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5.

Graham, M.A., Silverstein, K.A.T., VandenBosch, K.A., 2008. Defensin-like genes: Genomic perspectives on a diverse superfamily in plants. Crop Science 48, S3-S11.

Hubert, F., 1996. A member of the arthropod defensin family from edible Mediterranean mussels (*Mytilus galloprovincialis*). European Journal of Biochemistry 240, 302-306.

Ittiprasert, W., Miller, A., Myers, J., Nene, V., El-Sayed, N.M., Knight, M., 2010. Identification of immediate response genes dominantly expressed in juvenile resistant and susceptible *Biomphalaria glabrata* snails upon exposure to *Schistosoma mansoni*. Molecular and Biochemical Parasitology 169, 27-39.

Jung, S., Dingley, A.J., Augustin, R., Anton-Erxleben, F., Stanisak, M., Gelhaus, C., Gutsmann, T., Hammer, M.U., Podschun, R., Bonvin, A.M.J.J., Leippe, M., Bosch, T.C.G., Grötzinger, J., 2009. Hydramacin-1, structure and antibacterial activity of a protein from the basal metazoan hydra. Journal of Biological Chemistry 284, 1896-1905.

Kawabata, S., Iwanaga, S., 1997. Big defensin and tachylectins-1 and -2. Methods in molecular biology (Clifton, N.J.) 78, 51-61.

Kelley, L.A., Sternberg, M.J.E., 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protocols 4, 363-371.

Kouno, T., Fujitani, N., Mizuguchi, M., Osaki, T., Nishimura, S.I., Kawabata, S.I., Aizawa, T., Demura, M., Nitta, K., Kawano, K., 2008. A novel β-defensin structure: A potential strategy of big defensin for overcoming resistance by gram-positive bacteria. Biochemistry 47, 10611-10619.

Kouno, T., Mizuguchi, M., Aizawa, T., Shinoda, H., Demura, M., Kawabata, S.I., Kawano, K., 2009. A novel β-defensin structure: Big defensin changes its N-terminal structure to associate with the target membrane. Biochemistry 48, 7629-7635.

Kubota, Y., Watanabe, Y., Otsuka, H., Tamiya, T., Tsuchiya, T., Matsumoto, J.J., 1985. Purification and characterization of an antibacterial factor from snail mucus. Comp Biochem Physiol C 82, 345-348.

Livak, K., 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. Methods 25, 402-408.

Mitta, G., Hubert, F., Dyrynda, E.A., Boudry, P., Roch, P., 2000a. Mytilin B and MGD2, two antimicrobial peptides of marine mussels: Gene structure and expression analysis. Developmental and Comparative Immunology 24, 381-393.

Mitta, G., Hubert, F., Noël, T., Roch, P., 1999a. Myticin, a novel cysteine-rich antimicrobial peptide isolated from haemocytes and plasma of the mussel *Mytilus galloprovincialis*. European Journal of Biochemistry 265, 71-78.

Mitta, G., Vandenbulcke, F., Hubert, F., Roch, P., 1999b. Mussel defensins are synthesised and processed in granulocytes then released into the plasma after bacterial challenge. Journal of Cell Science 112, 4233-4242.

Mitta, G., Vandenbulcke, F., Hubert, F., Salzet, M., Roch, P., 2000b. Involvement of mytilins in mussel antimicrobial defense. Journal of Biological Chemistry 275, 12954-12962.

Mitta, G., Vandenbulcke, F., Noel, T., Romestand, B., Beauvillain, J.C., Salzet, M., Roch, P., 2000c. Differential distribution and defence involvement of antimicrobial peptides in mussel. Journal of Cell Science 113, 2759-2769.

Pallavicini, A., del Mar Costa, M., Gestal, C., Dreos, R., Figueras, A., Venier, P., Novoa, B., 2008. High sequence variability of myticin transcripts in hemocytes of immune-stimulated mussels suggests ancient host-pathogen interactions. Developmental and Comparative Immunology 32, 213-226.

Patrzykat, A., Douglas, S.E., 2003. Gone gene fishing: How to catch novel marine antimicrobials. Trends in Biotechnology 21, 362-369.

Ramakers, C., Ruijter, J.M., Deprez, R.H.L., Moorman, A.F.M., 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neuroscience Letters 339, 62-66.

Roch, P., Yang, Y., Toubiana, M., Aumelas, A., 2008. NMR structure of mussel mytilin, and antiviral-antibacterial activities of derived synthetic peptides. Dev Comp Immunol 32, 227-238.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Saito, T., Kawabata -i, S., Shigenaga, T., Tokayenoki, Y., Cho, J., Nakajima, H., Hirata, M., Iwanaga, D., 1995. A novel big defensin identified in horseshoe crab hemocytes: Isolation, amino acid sequence, and antibacterial activity. Journal of Biochemistry 117, 1131-1137.

Schikorski, D., Cuvillier-Hot, V., Leippe, M., Boidin-Wichlacz, C., Slomianny, C., Macagno, E., Salzet, M., Tasiemski, A., 2008. Microbial challenge promotes the regenerative process of the injured central nervous system of the medicinal leech by inducing the synthesis of antimicrobial peptides in neurons and microglia. Journal of Immunology 181, 1083-1095.

Schmitt, P., Gueguen, Y., Desmarais, E., Bachere, E., de Lorgeril, J., 2010. Molecular diversity of antimicrobial effectors in the oyster *Crassostrea gigas*. BMC Evolutionary Biology 10, 23.

Selsted, M.E., Tang, Y.Q., Morris, W.L., McGuire, P.A., Novotny, M.J., Smith, W., Henschen, A.H., Cullor, J.S., 1993. Purification, primary structures, and antibacterial activities of beta-defensins, a new family of antimicrobial peptides from bovine neutrophils. Journal of Biological Chemistry 268, 6641-6648.

Sonthi, M., Toubiana, M., Pallavicini, A., Venier, P., Roch, P., 2011. Diversity of Coding Sequences and Gene Structures of the Antifungal Peptide Mytimycin (MytM) from the Mediterranean Mussel, *Mytilus galloprovincialis*. Mar Biotechnol (NY).

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution.

Tasiemski, A., Vandenbulcke, F., Mitta, G., Lemoine, J., Lefebvre, C., Sautière, P.E., Salzet, M., 2004. Molecular characterization of two novel antibacterial peptides inducible upon bacterial challenge in an annelid, the leech *Theromyzon tessulatum*. Journal of Biological Chemistry 279, 30973-30982.

Tian, C., Gao, B., Fang, Q., Ye, G., Zhu, S., 2010. Antimicrobial peptide-like genes in *Nasonia vitripennis*: a genomic perspective. BMC Genomics 11, 187.

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

Venier, P., Varotto, L., Rosani, U., Millino, C., Celegato, B., Bernante, F., Lanfranchi, G., Novoa, B., Roch, P., Figueras, A., Pallavicini, A., 2011. Insights into the innate immunity of the Mediterranean mussel *Mytilus galloprovincialis*. BMC Genomics 12, 69.

Wang, Y., Zhu, S., The defensin gene family expansion in the tick *Ixodes scapularis*. Developmental & Comparative Immunology In Press, Accepted Manuscript.

Xu, Q., Wang, G., Yuan, H., Chai, Y., Xiao, Z., 2010. CDNA sequence and expression analysis of an antimicrobial peptide, theromacin, in the triangle-shell pearl mussel *Hyriopsis cumingii*. Comparative Biochemistry and Physiology - B Biochemistry and Molecular Biology 157, 119-126.

Yang, Y.S., Mitta, G., Chavanieu, A., Calas, B., Sanchez, J.F., Roch, P., Aumelas, A., 2000. Solution structure and activity of the synthetic four-disulfide bond Mediterranean mussel defensin (MGD-1). Biochemistry 39, 14436-14447.

Yeaman, M.R., Yount, N.Y., 2007. Unifying themes in host defence effector polypeptides. Nat Rev Micro 5, 727-740.

Zhao, J., Li, C., Chen, A., Li, L., Su, X., Li, T., 2010. Molecular characterization of a novel big defensin from clam *Venerupis philippinarum*. PLoS ONE 5.

Zhao, J., Song, L., Li, C., Ni, D., Wu, L., Zhu, L., Wang, H., Xu, W., 2007. Molecular cloning, expression of a big defensin gene from bay scallop *Argopecten irradians* and the antimicrobial activity of its recombinant protein. Molecular Immunology 44, 360-368.

Environ. Sci. Technol. **2010**, *44*, 8276–8283

Mediterranean Mussel Gene Expression Profile Induced by Okadaic Acid Exposure

CHIARA MANFRIN, RENE' DREOS,[†]
SILVIA BATTISTELLA, ALFRED BERAN,[‡]
MARCO GERDOL, LAURA VAROTTO,[‡]
GEROLAMO LANFRANCHI,[||]
PAOLA VENIER,[§] AND
ALBERTO PALLAVICINI*

Department of Life Sciences, Università di Trieste, P.le Valmaura, 9, Trieste, Italy 34148

Received November 19, 2009. Revised manuscript received September 15, 2010. Accepted September 15, 2010.

Mediterranean Mussel Gene Expression Profile Induced by Okadaic Acid Exposure

Chiara Manfrin[1], Renè Dreos[2], Silvia Battistella[1], Alfred Beran[3], Marco Gerdol[1], Laura Varotto[4], Gerolamo Lanfranchi[4], Paola Venier[4], Alberto Pallavicini[1]

[1] Department of Life Sciences, University of Trieste, Trieste, Italy

[2] John Innes Centre Norwich Research Park,Colney, Norwich, UK NR47UH

[3] Istituto nazionale di Oceanografia e di Geofisica sperimentale, Dipartimento di Oceanografia Biologica, via Auguste Piccard 54, 34151 Santa Croce, Trieste, Italy

[4] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

Environmental Science & Technology **44**(21): 8276-8283

Gli incrementi stagionali della temperatura delle acque marine definiscono condizioni ottimali per la crescita di varie specie di alghe Dinoflagellate, che possono raggiungere alte concentrazioni nella colonna d'acqua e anche negli organismi filtratori, come i mitili.
L'acido okadaico (OA) ed i suoi analoghi, comunqmente prodotti da *Dynophisis* e *Prorocentrum* spp., sono responsabili della sindrome DSP (Diarrheic Shellfish Poisoning) nell'uomo. La chiusura delle molluschicolture è pertanto raccomandata quando i livelli di tossine diarreiche è maggiore di 16 μg OA 100 g^{-1} di carne. Nonostante non siano responsabili di morte né nell'uomo, ne nei mitili,

gli eventi di DSP sono considerati eventi naturali che provocano problemi legati alla salute pubblica ed all'economia data la loro frequenza. Sfruttando le opportunità offerte dagli studi di espressione genica, nel corso di questo esperimento è stato utilizzato un microarray a cDNA specificamente disegnato su mitilo, al fine di poter quantificare i cambiamenti nell'espressione genica nella ghiandola digestiva di mitili nutriti per 5 settimane con cibo contaminato con acido okadaico.
Tra i geni differenzialmente espressi è stata osservata una up-regolazione generalizzata dei trascritti codificanti proteine legate allo stress, coinvolte nella sintesi cellulare ed alcune ancora non annotate. Globalmente, nel primo punto sperimentale analizzato, sono stati identificati 58 trascritti candidati come marker per lo stress indotto dall'acido okadaico. La metà di questi è tuttora a funzione ignota, vista l'assenza di similarità con altre sequenze depositate nei database pubblici.
In questo lavoro è stato presentata la prima analisi di espressione genica condotta su mitili esposti ad acido okadaico. La caratterizzazione dei trascritti responsivi all'OA potrebbe risultare di grande utilità per l'identificazione di una rapida risposta fisiologica all'esposizione a questa biotossina algale.

ISJ 8: 179-189, 2011 ISSN 1824-307X

REVIEW

How gene expression profiles disclose vital processes and immune responses in *Mytilus* spp.

S Domeneghetti[1*], C Manfrin[2*], L Varotto[1], U Rosani[1], M Gerdol[2], G De Moro[2], A Pallavicini[2], P Venier[1]

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

*Equal contribution

Accepted September 08, 2011

How gene expression profiles disclose vital processes and immune responses in *Mytilus* spp.

Stefania Domeneghetti[a], Chiara Manfrin[b], Laura Varotto[a], Umberto Rosani[a], Paola Venier[a], Marco Gerdol[b], Gianluca De Moro[b] Alberto Pallavicini[b]

[a] Department of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

[b] Department of Life Sciences, University of Trieste, Trieste, Italy

ISJ - Invertebrate Survival Journal **8**(2).

Gli studi di espressione genica supportano largamente la comprensione delle interazioni tra geni ed ambiente nell'uomo e negli altri organismi viventi, ma la mancanza di informazioni genomiche e genetiche spesso complica l'analisi delle risposte funzionali in specie modello non tradizionali. Cionondimeno, il rapido avanzamento delle tecnologie di DNA microarray e di sequenziamento ora rende le analisi globali di espressione genica possibili in virtualmente qualsiasi specie di interesse. Per quanto riguarda il genere *Mytilus*, decine di migliaia di Expressed Sequence Tags (ESTs) sono al momento disponibili per *M. californianus* e *M. galloprovincialis*, e microarray a DNA sono stati sviluppati per entrambe le specie.
Tra queste, l'Immunochip 1.0 include specificamente 1,820 sonde di geni centralmente coinvolti o modulati nella risposta immunitaria innata del mitilo Mediterraneo. Questa review richiama le peculiarità e le applicazioni dei microarray a DNA esistenti e riassume le conoscenze attuali riguardo ad una varietà di trascritti con ogni probabilità coinvolti nell'immunità del mitilo. Oltre ai microarray a DNA, le tecnologie di sequenziamento di nuova generazione (NGS) oggi offrono prospettive di ricerca nuove e più ampie, a partire dalla copertura completa del trascrittoma per arrivare al sequenziamento completo del genoma di mitilo.

PARTECIPAZIONI A CONVEGNI E CONFERENZE

Gerdol M, Bernante F, Venier P, Pallavicini A: **A survey on *Mytilus galloprovincialis* C1q-domain containing proteins**. In: *ESF-FWF Conference in Partnership with LFUI: The Impact of the Environment on Innate Immunity: The Threat of Diseases.* Universitätszentrum Obergurgl, Austria; 2009.

Gerdol M, Bernante F, Venier P, Pallavicini A: **A survey on *Mytilus galloprovincialis* C1q-domain containing proteins**. In: *11th Congress Int Soc Dev Comp Immunol.* Prague, Czech Republic; 2009: 32.

Varotto L, Rosani U, Domeneghetti S, Procopio G, Manfrin C, Gerdol M, Pallavicini A, Venier P: **Expression and diversity of AMP and other immune-related molecules in *M. galloprovincialis***. In: *27th Congress of the newEuropean Society of Comparative Biochemistry and Physiology: 2010; Alessandria, Italy*; 2010.

Gerdol M, Manfrin C, Venier P, Pallavicini A: **Characterization of Mytimacins, a novel cationic antimicrobial peptides family in the Mediterranean mussel *Mytilus galloprovincialis***. In: *45th European Marine Biology Symposium.* Heriot-Watt University, Edinburgh, Scotland; 2010.

Manfrin C, Gerdol M, Sosa S, Venier P, Tubaro A, Pallavicini A: **A first glimpse towards genes putatively involved in okadaic acid-contamination in Mediterranean mussel**. In: *45th European Marine Biology Symposium 2010.* Heriot-Watt University, Edinburgh, Scotland; 2010.

Gerdol M, De Moro G, Manfrin C, Beran A, Pallavicini A: **The transcriptomic analysis of *Mytilus galloprovincialis* by RNA-seq reveals a specific response to toxic *Alexandrium minutum***. In: *46th European Marine Biology Symposium.* Rovinj, Croatia; 2011.

Gerdol M, Casagrande R, De Moro G, Manfrin C, Venier P, Pallavicini A: **Comparative sequence analysis of antimicrobial peptides in *Mytilus galloprovincialis* and *Ruditapes philippinarum***. In: *XIII Meeting of the Italian Association of Developmental and Comparative Immunobiology (IADCI).* San Benedetto del Tronto; 2012.

BIBLIOGRAFIA

1. Coustau C, Renaud F, Delay B: **Genetic characterization of the hybridization between *Mytilus edulis* and *M. galloprovincialis* on the Atlantic coast of France**. *Marine Biology* 1991, **111**(1):87-93.
2. Seed R, Suchanek TH: **Population and community ecology of Mytilus**. In: *The mussel Mytilus: ecology, physiology, genetics and culture.* Edited by Gosling E; 1992: 87-169.
3. Daguin C, Borsa P: **Genetic relationships of *Mytilus galloprovincialis* Lamarck populations worldwide: Evidence from nuclear-DNA markers**. In.; 2000: 389-397.
4. Hilbish TJ, Mullinax A, Dolven SI, Meyer A, Koehn RK, Rawson PD: **Origin of the antitropical distribution pattern in marine mussels (*Mytilus* spp.): Routes and timing of transequatorial migration**. *Marine Biology* 2000, **136**(1):69-77.
5. Gosling EM: **Bivalve molluscs: biology, ecology culture**. Oxford: Fishing New Books; 2003.
6. Lutz A, Kennis MJ: **Ecology and morphology of larval and early postlarval mussels**. In: *The mussel Mytilus: ecology, physiology, genetics and culture.* Edited by Gosling EM; 1992: 87-169.
7. Juhel G, Davenport J, O'Halloran J, Culloty S, Ramsay R, James K, Furey A, Allis O: **Pseudodiarrhoea in zebra mussels *Dreissena polymorpha* (Pallas) exposed to microcystins**. *Journal of Experimental Biology* 2006, **209**(5):810-816.
8. Liviero A: **La produzione di mitili e vongole nelle regioni Alto Adriatiche**. In: *La pesca in numeri.* Edited by Adriatico OSEdPdA. Chioggia (VE): Veneto Agricoltura; 2007.
9. Bower SM: **Diseases and parasites of mussels**. In: *The mussel Mytilus: ecology, physiology, genetics and culture.* Edited by Gosling EM; 1992: 1-20.
10. Riedl R: **Fauna e Flora del Mediterraneo - dalle alghe ai mammiferi: una guida sitematica alle specie che vivono del Mar Mediterraneo**: Franco Murzio ed.; 1991.
11. Smayda TJ: **Novel and nuisance phytoplancton blooms in the sea: Evidence for a global epidemic**. In: *Toxin Marine Phytoplancton.* Edited by Edler L, Anderson DM. New York: Elsiver; 1990: 29-40.
12. Hallegraff GM: **Harmful algal blooms: a global overview**. In: *Manual on harmful marine microalgae.* UNESCO; 1995: 1-22.
13. Shumway SE: **Toxic algae: a serious threat to shellfish aquaculture**. *World Aquaculture* 1989, **20**:65-74.
14. Satake M, Ofuji K, Naoki H, James KJ, Furey A, McMahon T, Silke J, Yasumoto T: **Azaspiracid, a new marine toxin having unique spiro ring assemblies, isolated from irish mussels, *Mytilus edulis* [14]**. *Journal of the American Chemical Society* 1998, **120**(38):9967-9968.
15. Hallegraeff GM, Steffensen DA, Wetherbee R: **Three estuarine Australian dinoflagellates that can produce paralytic shellfish toxins**. *J PLANKTON RES* 1988, **10**(3):533.
16. Mee LD, Espinosa M, Diaz G: **Paralytic shellfish poisoning with a *Gymnodinium catenatum* red tide on the Pacific coast of Mexico**. *Marine Environmental Research* 1986, **19**(1):77-92.
17. Terlau H, Heinemann SH, Stühmer W, Pusch M, Conti F, Imoto K, Numa S: **Mapping the site of block by tetrodotoxin and saxitoxin of sodium channel II**. *FEBS Letters* 1991, **293**(1–2):93-96.
18. García C, del Carmen Bravo Ma, Lagos M, Lagos N: **Paralytic shellfish poisoning: post-mortem analysis of tissue and body fluid samples from human victims in the Patagonia fjords**. *Toxicon* 2004, **43**(2):149-158.
19. Bricelj VM, Lee JH, Cembella AD: **Influence of dinoflagellate cell toxicity on uptake and loss of paralytic shellfish toxins in the northern quahog *Mercenaria mercenaria***. *Marine Ecology Progress Series* 1991, **74**(1):33-46.
20. MacQuarrie SP, Bricelj VM: **Behavioral and physiological responses to PSP toxins in *Mya arenaria* populations in relation to previous exposure to red tides**. *Marine Ecology Progress Series* 2008, **366**:59-74.
21. Bricelj VM, Shumway SE: **Paralytic shellfish toxins in bivalve molluscs: Occurrence, transfer kinetics, and biotransformation**. *Reviews in Fisheries Science* 1998, **6**(4):315-383.
22. Connell LB, MacQuarrie SP, Twarog BM, Iszard M, Bricelj VM: **Population differences in nerve resistance to paralytic shellfish toxins in softshell clam, *Mya arenaria*, associated with sodium channel mutations**. *Marine Biology* 2007, **150**(6):1227-1236.

23. Shumway SE, Pierce FC, Knowlton K: **The effect of Protogonyaulax Tamarensis on byssus production in *Mytilus edulis* l., *Modiolus modiolus* linnaeus, 1758 and *Geukensia demissa* dillwyn**. *Comparative Biochemistry and Physiology -- Part A: Physiology* 1987, **87**(4):1021-1023.
24. Galimany E, Hégaret H, Ramón M, Sunila I, Wikfors GH: **Experimental exposure of the blue mussel (*Mytilus edulis*, L.) to the toxic dinoflagellate *Alexandrium fundyense*: Histopathology, immune responses, and recovery**. *Harmful Algae* 2008, **7**:702-711.
25. Francisco CJ, Hermida MA, Santos MJ: **Parasites and Symbionts from *Mytilus galloprovincialis* (Lamark, 1819) (Bivalves: Mytilidae) of the Aveiro Estuary Portugal**. *Journal of Parasitology* 2009, **96**(1):200-205.
26. Watermann BT, Herlyn M, Daehne B, Bergmann S, Meemken M, Kolodzey H: **Pathology and mass mortality of Pacific oysters, *Crassostrea gigas* (Thunberg), in 2005 at the East Frisian coast, Germany**. *Journal of Fish Diseases* 2008, **31**(8):621-630.
27. 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**. *Reviews in Fisheries Science* 2008, **16**(SUPPL.1):131-154.
28. Venier P, Varotto L, Rosani U, Millino C, Celegato B, Bernante F, Lanfranchi G, Novoa B, Roch P, Figueras A *et al*: **Insights into the innate immunity of the Mediterranean mussel *Mytilus galloprovincialis***. *BMC Genomics* 2011, **12**.
29. Gorbushin AM, Iakovleva NV: **A new gene family of single fibrinogen domain lectins in *Mytilus***. *Fish and Shellfish Immunology* 2011, **30**(1):434-438.
30. Gestal C, Pallavicini A, Venier P, Novoa B, Figueras A: **MgC1q, a novel C1q-domain-containing protein involved in the immune response of *Mytilus galloprovincialis***. *Developmental and Comparative Immunology* 2010, **34**(9):926-934.
31. Weiss IM, Kaufmann S, Mann K, Fritz M: **Purification and characterization of perlucin and perlustrin, two new proteins from the shell of the mollusc *Haliotis laevigata***. *Biochemical and Biophysical Research Communications* 2000, **267**(1):17-21.
32. Li H, Parisi MG, Parrinello N, Cammarata M, Roch P: **Molluscan antimicrobial peptides, a review from activity-based evidences to computer-assisted sequences**. *Invertebrate Survival Journal* 2011, **8**(1):85-97.
33. Gueguen Y, Bernard R, Julie F, Paulina S, Delphine DG, Franck V, Philippe B, Evelyne B: **Oyster hemocytes express a proline-rich peptide displaying synergistic antimicrobial activity with a defensin**. *Molecular Immunology* 2009, **46**(4):516-522.
34. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.
35. Hudson ME: **Sequencing breakthroughs for genomic ecology and evolutionary biology**. *Molecular Ecology Resources* 2008, **8**(1):3-17.
36. Jones W: **High-Throughput Sequencing and Metagenomics**. *Estuaries and Coasts* 2010, **33**(4):944-952.
37. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
38. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA *et al*: **The Genome Portal of the Department of Energy Joint Genome Institute**. *Nucleic Acids Research* 2012, **40**(D1):D26-D32.
39. Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW: **Resolving the evolutionary relationships of molluscs with phylogenomic tools**. *Nature* 2011, **480**(7377):364-367.
40. Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B *et al*: **Phylogenomics reveals deep molluscan relationships**. *Nature* 2011, **477**(7365):452-456.