



UNIVERSITA' DEGLI STUDI DI TRIESTE

**XXV CICLO DEL DOTTORATO DI RICERCA IN
BIOLOGIA AMBIENTALE**

**ANALISI E GESTIONE INFORMATICA
DI SEQUENZE TRASCritte
IN ORGANISMI NON-MODELLO**

Settore scientifico-disciplinare BIO/18

DOTTORANDO

GIANLUCA DE MORO

RELATORE

PROF. ALBERTO PALLAVICINI

ANNO ACCADEMICO 2011/2012

INDICE

INDICE	2
INTRODUZIONE	5
Le tecnologie di sequenziamento di nuova generazione e la loro applicazione in ambito trascrittomico di organismi non-modello	5
Problematiche legate all'elaborazione di dati RNA-seq	9
Software di assemblaggio e analisi	10
PRODUZIONE SCIENTIFICA DEL CANDIDATO	17
<i>Mytilus galloprovincialis</i>	20
RNA sequencing and <i>de novo</i> assembly of the digestive gland transcriptome in <i>Mytilus galloprovincialis</i> fed with toxinogenic and non-toxic strains of <i>Alexandrium minutum</i>	24
Abstract	24
Introduction	25
Results and Discussion	27
Conclusion	33
Methods	34
Tables	39
Figures	44
References	53
The C1q domain containing proteins of the Mediterranean mussel <i>Mytilus galloprovincialis</i>: A widespread and diverse family of immune-related molecules	61
Abstract	61
Introduction	62
Materials and methods	63
Results	65
Discussion	67
Conclusions	70
Tables	71
Figures	75
References	78
Big defensins and mytimacins, new AMP families of the Mediterranean mussel <i>Mytilus galloprovincialis</i>	81
Introduction	82
Materials and methods	83
Results and discussion	85
Mytimacins	87

Conclusions	90
Tables.....	91
Figures	93
References	100
How gene expression profiles disclose vital processes and immune responses in <i>Mytilus</i> spp.	103
Abstract	103
Introduction	104
Concluding remarks	113
Acknowledgments.....	113
Figures	114
References	120
Physiological and molecular responses of bivalves to toxic dinoflagellates	125
Abstract	125
Introduction	126
Conclusion	140
Tables.....	141
Figures	144
References	146
<i>Latimeria menadoensis</i>:	156
Analysis of the transcriptome of the Indonesian coelacanth <i>Latimeria menadoensis</i>.....	159
Abstract	159
Introduction	160
Results.....	162
Discussion	167
Conclusion	173
Methods	175
Figures	180
Tables.....	188
References	194
Comparative analysis of the genome of the African coelacanth, <i>Latimeria chalumnae</i>, sheds light on tetrapod evolution	200
Abstract	200
Introduction	201
Discussion	210
Author contributions.....	212
References	217

Characterization of sex determination and sex differentiation genes in <i>Latimeria</i>	222
Abstract	222
Introduction	223
Methods	225
Results.....	227
Discussion	232
Conclusions	236
Figures	237
Tables.....	250
References	254
Characterization of purine catabolic pathway genes in coelacanth	260
Abstract	260
<i>Pontastacus leptodactylus</i>:.....	261
Hepatopancreatic transcriptome in the crayfish <i>Pontastacus leptodactylus</i> reveals peptidase activation and glycolysis suppression following injection of D-crustacean Hyperglycemic Hormone.	263
Abstract	263
Introduction	264
Materials and Methods	266
Results and Discussion	268
Conclusions	274
Figures	277
Tables.....	281
References	285
Altri organismi:	292
Bibliografia	293

INTRODUZIONE

Le tecnologie di sequenziamento di nuova generazione e la loro applicazione in ambito trascrittomico di organismi non-modello

Prima dell'avvento delle nuove tecnologie di sequenziamento denominate di “nuova generazione” (NGS, dal termine inglese “Next Generation Sequencing”), la tecnologia di sequenziamento storicamente più utilizzata è stata la tecnologia Sanger.

Questa tecnologia, chiamata anche metodo della terminazione della catena, è stata sviluppata da Frederick Sanger nel 1977 (Sanger et al., 1977) e, data l'alta efficienza e la bassa radioattività, ha riscosso subito un enorme successo.

Negli anni questa tecnologia è stata inoltre migliorata con l'utilizzo dell'elettroforesi capillare che ha permesso una maggior accuratezza e velocità di sequenziamento.

Solamente da quando Roche Diagnostic rilasciò, dopo l'acquisizione della 454 Life Scienze, il primo sequenziatore GS20 nel 2005 le tecnologie di sequenziamento di nuova generazione si sono rese disponibili.

Oltre ad un perfezionamento e miglioramento della tecnologia 454, negli anni immediatamente successivi si sono affacciate sul mercato anche altre piattaforme come Illumina, con il Genome Analyzer I nel 2007, e Solid della Applied Biosystem nel 2008.

Queste innovative tecnologie, oltre ad aumentare la velocità di sequenziamento riducendone simultaneamente i costi, hanno anche portato un elevato aumento di produttività.

Questa maggior produttività, motivo per cui queste tecniche vengono definite spesso “high throughput sequencing” (HTS), deriva dal sequenziamento in parallelo di un gran numero di molecole di DNA.

Queste piattaforme inoltre, seppur diverse nella biochimica di sequenziamento, utilizzano un protocollo concettualmente simile.

Innanzitutto vi è una fase di ancoraggio, in cui le singole molecole di DNA vengono immobilizzate su di un supporto solido, seguita da diversi cicli di amplificazione tramite PCR su fase solida, con cicli ripetuti di lavaggio e scansione (“wash-and-scan”),

Nel processo di sequenziamento “wash-and-scan” le molecole ancorate al supporto solido sono immerse in reagenti, come nucleotidi marcati, in modo che i nucleotidi vengano incorporati nelle eliche di DNA. Dopo aver fermato l'incorporazione l'eccesso di reagente viene eliminato, viene scansionato il supporto per identificare quale base è stata incorporata e, infine, la nuova base incorporata viene trattata in modo da preparare il *template* di DNA per il successivo ciclo “wash-and-scan”.

Una descrizione dettagliata della chimica di questi metodi di sequenziamento esulano dallo scopo di questa tesi e, per un approfondimento delle tecnologie NGS, si rimanda quindi a specifiche *review* sull'argomento. (Shendure and Ji, 2008)

Questi supporti, dove il DNA è ancorato, possono avere un elevatissima densità di frammenti di DNA portando una processività che, ad esempio nel caso dello strumento Illumina HiSeq 2500, può generare in un'unica corsa fino a 600 gigabasi di dati di sequenze.

A causa di questi cicli, e a seconda della metodica di sequenziamento scelto, le operazioni possono impiegare da alcune ore a diversi giorni (Tabella 1).

Poiché il rendimento per ogni passaggio è inferiore al 100%, una popolazione di molecole diventa più “asincrona” per ogni base aggiunta. Questa perdita di sincronicità, chiamata “*dephasing*” causa un incremento nel rumore di fondo ed un aumento di errori di sequenziamento durante l'estensione delle *reads*.

Il *dephasing*, oltre a rendere la gestione dei dati più difficoltosa, è anche alla base della limitata lunghezza delle sequenze prodotte dai sequenziamenti NGS rispetto al sequenziamento Sanger. (Schadt et al., 2010)

Recentemente si sono rese disponibili tecnologie di sequenziamento denominate di “terza generazione” che, seppur basate anch'esse sulla metodica “wash-and-scan”, non necessitano di una fase di PCR prima del sequenziamento e si basano principalmente sul fatto che il segnale è catturato in tempo reale ed è monitorato durante la reazione enzimatica di aggiunta di nucleotidi nell'elica complementare.

Grazie a queste caratteristiche le tecnologie di terza generazione, tra le quali ad esempio la Ion Torrent, permettono di avere una velocità di esecuzione maggiore e, essendo più stabili, sembrano avere un'accuratezza maggiore rispetto alle tecniche di seconda generazione.

Anche in questo caso, per un approfondimento della chimica delle tecnologie di terza generazione rimando a pubblicazioni specifiche. (Schadt et al., 2010)

Sequenziatore	454 Gs FLX+	Illimina HiSeq 2500	Solidv4
Metodica di Sequenziamento	Pyrosequencing	Suquencing by synthesis	Ligation and two-base coding
Lunghezza di Reads	1000pb	100-150pb paired-end	50+50pb
Accuratezza	99,997%	98%	99,94%
Numero di Basi	700M	600Gb (100Pe) 90Gb (150Pe)	120 Gb
Tempo di Esecuzione	23 Ore	Da 3 a 10 giorni	Da 7 a 14 giorni
Costo per milione di basi	10\$	0.07\$	0.14\$
Vantaggi	Velocità e lunghezza delle reads	Numero di basi prodotte	Qualità delle Reads
Svantaggi	Costo, processività	Lunghezza delle reads	Lunghezza delle reads

Tabella 1: Vantaggi, costi e meccanismi di sequenziamento delle tecnologie di sequenziamento di nuova generazione

Queste tecniche di sequenziamento sono state recepite dalla comunità scientifica molto rapidamente ma se nei primi anni le tecnologie 454 e Illumina si sono divise quasi equamente il mercato, ultimamente appare chiaro come quest'ultima sia la tecnologia che più di ogni altra viene utilizzata (Figura 1).

Le metodiche NGS non solo hanno portato degli enormi vantaggi nel sequenziamento genomico ma anche, soprattutto grazie alla tecnologia denominata RNA-seq, una vera e propria rivoluzione in campo trascrittomico.

Questa tecnologia denominata anche “Whole Transcriptome Shotgun Sequencing” (WTSS), basata sul sequenziamento del cDNA, permette di ottenere informazioni sull'mRNA espresso e di assemblare *de novo* interi trascrittomi.

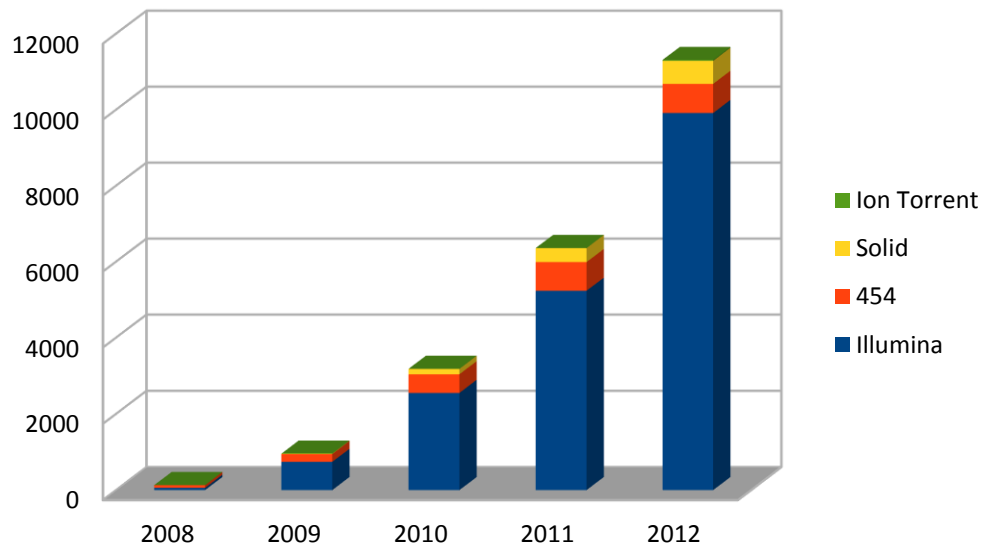


Figura 1 Confronto tra le 4 principali piattaforme NGS (ABI Solid, 454 Life Science, Illumina e Ion-torrent) negli ultimi 6 anni. Il numeri di dati sono ricavati dagli esperimenti depositati negli archivi SRA

La possibilità di produrre un assemblaggio trascrittomico *de novo*, senza la necessità di avere informazioni genomiche pregresse, ne ha fatto uno strumento di eccellenza nello studio di organismi non-modello.

Il sequenziamento trascrittomico per gli organismi non-modello è, inoltre, diventata negli ultimi anni molto utilizzato grazie ai suoi costi relativamente contenuti e alla minor necessità di potenza di calcolo rispetto alla gestione di un sequenziamento genomico.

Nonostante i dati ottenibili siano inferiori rispetto al sequenziamento genomico, le informazioni trascrittomiche possono essere utilizzate in una grande varietà di studi biologici come lo studio dei livelli di espressione genica, studi riguardanti i profili d'espressione dopo uno specifico trattamento, l'individuazione di isoforme di splicing alternativi, l'identificazione di trascritti di fusione o di espressioni *strand-specifiche*. (Dillies et al., 2012)

Oltre a permettere di avere un accesso diretto alle sequenze codificanti di molti geni, infatti, la tecnologia RNA-seq, ci consente di avere informazioni sui loro livelli di espressione relativi.

La WTTS permette anche di superare alcune problematiche tipiche della tecnologia *microarray*.

Quest'ultima, infatti, necessita di informazioni pregresse sulla sequenza genomica e di avere a disposizione una quantità elevata di sequenze per effettuare la cross-ibridazione.

Rispetto all'RNA-seq, il *microarray* presenta un inferiore range dinamico di rilevamento, a causa sia del segnale di background che della saturazione del segnale, e pone dei problemi di comparazione nei livelli di espressione tra differenti esperimenti che rendono necessario l'utilizzo di appositi metodi di normalizzazione.

L'RNA-Seq permette, al contrario, di avere un'intera panoramica su tutti i trascritti espressi in un tessuto, senza dover necessariamente avere informazioni genomiche pregresse, e permette di calcolare il livello di espressione assoluto di un trascritto senza segnali di background.

Oltre a questi vantaggi l'RNA-Seq, richiede una minor quantità di RNA di partenza, ha un range dinamico per lo studio dell'espressione genica maggiore e possiede un elevato livello di riproducibilità. (Wang et al., 2009)

Problematiche legate all'elaborazione di dati RNA-seq

Le sequenze prodotte da NGS necessitano, nel caso in cui non sia disponibile una genomica o un trascrittoma di riferimento, di essere assemblate tra loro per ottenere le predizioni geniche.

L'assemblaggio delle corte sequenze derivanti dal sequenziamento (*reads*) pone numerosi problemi, non ultimo l'enorme potenza di calcolo necessaria per gestirle.

Se da un lato sembrerebbe essere più semplice la gestione di sequenze trascrittomiche rispetto a sequenze genomiche, poiché i livelli di ripetizioni sono minori nelle regioni codificanti, in realtà l'assemblaggio e l'analisi di questo tipo di sequenze porta con sé un elevato numero di problematiche.

Innanzitutto le sequenze trascrittomiche sono molto meno informative delle genomiche, a causa della mancanza di regioni introniche, e questo pone particolari problemi nello studio di famiglie multigeniche che risultano essere più difficili da identificare.

Poiché i geni meno espressi sono molto meno rappresentati da corrispondenti *reads* vi è una differente copertura dei vari trascritti che pone dei problemi durante l'assemblaggio in cui le sequenze con bassa copertura sono più difficilmente assemblabili rispetto ai trascritti altamente espressi.

L'assemblaggio di dati trascrittomici è ulteriormente complicata dalla presenza di *splicing* alternativi che, a seconda dei parametri scelti per l'assemblaggio, possono essere messi più o meno in evidenza.

Un altro problema tipico della gestione di trascrittomi è dato dai software di assemblaggio.

La quasi totalità di *assembler* disponibili sono stati creati per l'assemblaggio di genomi e devono quindi essere utilizzati con particolare attenzione per l'assemblaggio trascrittomico.

E' possibile evidenziare tre fattori che maggiormente incidono sulla difficoltà d'utilizzo di *assembler* non specifici.

Innanzitutto, i programmi di sequenziamento genomico utilizzano la differenza di copertura per discernere le regioni altamente ripetute e quindi non riescono a gestire correttamente le ampie differenze di copertura dei dati trascrittomici.

Le sequenze di RNA-Seq, inoltre, sono solitamente *strand*-specifiche mentre gli *assembler* genomici sono sviluppati per gestire sequenze *both-stands* e non utilizzano, quindi, le informazioni sull'orientamento per risolvere i problemi di *overlapping*.

Infine, gli *assembler* genomici hanno difficoltà nel gestire strutture ripetute portando quindi notevoli problemi nell'analisi trascrittomico in cui varianti di uno stesso gene possono condividere lo stesso esone.

Tutti questi problemi portano, nell'assemblaggio *de novo* da sequenze di RNA-seq, la produzione di vari tipi di errori di predizione, tra i quali: chimere, frammenti genici, alleli non assemblati e assemblaggi di paraloghi.

Risulta quindi chiaro come la gestione di queste sequenze debba essere effettuata con particolare cura.

Software di assemblaggio e analisi

Nel corso di questo dottorato è stata messa a punto una *pipeline* che permette, fondendo assieme i *contig* derivanti da diversi programmi di sequenziamento, di ovviare parte dei problemi citati.

Non avendo lavorato su organismi modello, né su organismi dei quali era già disponibile un genoma o un trascrittoma di riferimento, in una prima fase le *reads* di tutti gli organismi studiati sono state assemblate.

Prima di poter procedere all'assemblaggio le sequenze sono state filtrate utilizzando il software CLC Genomic Workbench (<http://www.clcbio.com>), in modo da eliminare sequenze a bassa qualità o parti di sequenze contenenti adattatori utilizzati nel protocollo di sequenziamento.

In un primo momento gli assemblaggi sono stati quasi esclusivamente elaborati da una workstation con 74gb di RAM, presente all'interno del laboratorio in cui ho svolto il dottorato.

Nonostante l'ingente quantitativo di RAM in dotazione alla macchina solamente alcuni degli *assembler* disponibili riuscivano a completare la fase d'assemblaggio a causa dell'elevato numero di sequenze a nostra disposizione,

Inizialmente abbiamo testato alcuni *assembler* sviluppati appositamente per le corte *reads* di nuova generazione, alcuni basati sull'algoritmo Overlap-Layout-Consensus (OLC), come Newbler e Mira, altri sull'utilizzo di grafici de Bruijn (DBG), come il CLC Genomic Workbench.

I test di assemblaggio sono stati fatti utilizzando *reads* ibride, provenienti, cioè, da diverse tecnologie di sequenziamento.

Dopo alcuni test è stato subito evidente come solamente il software CLC Genomic Workbench riuscisse a gestire le sequenze a nostra disposizione nelle nostra *workstation* con risultati soddisfacenti.

Una delle caratteristiche peculiari di questo software proprietario è, oltre all'utilizzare un quantitativo di RAM inferiore rispetto ai software concorrenti, la sua velocità di assemblaggio che ci ha permesso di elaborare fino a 450 milioni di *reads* in poche ore.

Un altro vantaggio di questo software è dato dall'immediatezza d'utilizzo, della facilità con cui è possibile visualizzare i *mapping* delle sequenze sui *contig* creati e dalla possibilità di creare facilmente delle statistiche d'assemblaggio. Queste caratteristiche permettono di avere subito una stima della qualità dell'assemblaggio ottenuto.

Sfortunatamente per riuscire a garantire una così alta velocità d'esecuzione, unita ad una modesta richiesta di potenza di calcolo, il software CLC Genomic Workbench ha, nei confronti di altri *assembler*, una sensibilità più bassa e non riesce a riconoscere alcuni *overlaps* tra le *reads*.

Pur utilizzando una metodica DBG il CLC Genomic Workbench non permette di modificare la dimensione dei *K-mers* per cui non è possibile scegliere il rapporto tra sensibilità e velocità.

Questo software infine gestisce con difficoltà sequenze che presentano un elevato numero di ripetizioni e tende a ridurre la complessità degli assemblaggi fondendo varianti simili (possibili paraloghi, varianti alleliche o splicing alternativi) in un'unica sequenza consensus.

Rispetto ad altri *assembler*, inoltre, non garantisce lo stesso numero di *contig* creati né l'utilizzo dello stesso numero di *reads*. (Kumar and Blaxter, 2010)

Poiché la scelta dell'*assembler* era stata limitata dalla potenza di calcolo dei nostri computer è stato richiesto un accesso al cluster DIAG, una griglia di computer in cloud fondata dalla National Science Foundation, messo a disposizione dall'università del Maryland.

Grazie alla potenza di calcolo a disposizione (125 nodi bi-processore con 48Gb di ram per nodo) sono riuscito ad utilizzare un *assembler* specificatamente sviluppato per la gestione di sequenze di RNA-Seq, sviluppato dal Broad Institute, chiamato Trinity (Grabherr et al., 2011)

Questo software, basato anch'esso sull'utilizzo di grafi de Bruijn e già preinstallato in DIAG, garantisce, rispetto al CLC, una maggiore percentuale di *reads* mappate e la possibilità di scegliere con più libertà i singoli parametri d'assemblaggio.

Quando si effettua un sequenziamento trascrittomico, i livelli d'espressione dei differenti geni sono determinati dal conto del numero di *reads* mappate in un'entità biologica (per esempi un gene) e dalla normalizzazione di questo numero di *reads*, basata sulla lunghezza del gene preso in esame e il numero totale di *reads* mappate sul campione.

Solitamente i livelli di espressione sono quindi indicati in “Reads Per Kilobase per Million mapped reads” (RPKM), ossia le conte sono divise per la lunghezza dei trascritti in kilobasi e moltiplicate per il numero totale delle *reads* mappate, espresse in milioni.

Questo dovrebbe permettere la comparazione dei livelli di espressione sia tra geni di lunghezza differente che tra campioni di diversa profondità di sequenziamento.

Recentemente questo tipo di normalizzazione è stata messa in discussione in alcune pubblicazioni per cui nello studio di cambiamenti di livelli d'espressione fatti durante questo dottorato sono stati utilizzati anche altri metodi statistici. (Wagner et al., 2012)

Per effettuare un'analisi dei livelli d'espressione è necessario che le *reads* vengano allineate sulle sequenze contigue create dall'assemblaggio.

A momento sono disponibili più di 60 di questi software che permettono l'allineamento, denominati *mapper*.

La maggior parte di questi software sono stati rilasciati dopo il 2008, e tra questi solamente 9 sono specifici per il *mapping* di dati trascrittomici. (Fonseca et al., 2012)

Questo gran numero di software disponibili deriva dal fatto che ognuno di questi deve adattarsi alla crescente quantità di dati generati dall'NGS cercando di seguire lo sviluppo dei nuovi protocolli e delle nuove tecnologie.

La scelta del miglior *mapper* deve tener conto non solo della tecnologia specifica per cui un determinato *mapper* è stato creato (DNA, RNA, miRNA) ma anche deve tener conto della piattaforma di sequenziamento che ha generato tali dati.

Come già spiegato in precedenza, a causa del “dephasing”, nella piattaforma Illumina, ad esempio, l'accuratezza di sequenziamento decresce con l'aumentare del numero di cicli, quindi verso l'estremità 3' di ogni sequenza sono presenti basi meno affidabili.

Alcuni *mapper*, come ad esempio Bowtie2, tenendo conto dell'inferiore affidabilità nelle estremità, possono tagliare alcune basi per cercare di contrastare questo problema (Langmead and Salzberg, 2012).

Non tutti i *mapper* riescono, inoltre, a gestire le sequenze *paired* e, quindi, scegliendo il software sbagliato si rischierebbe di perdere quest'utile informazione che permetter di migliorare la rilevazione degli errori di allineamento e di migliorarne sensibilità e specificità.

Visti gli enormi numeri di sequenze che vengono generate dalle più recenti tecnologie di sequenziamento è, inoltre, preferibile che i *mapper*, così come gli *assembler*, possano nativamente essere eseguiti in parallelo, come ad esempio in cluster composti da molti computer .

Durante il mio lavoro di dottorato si è scelto di utilizzare due diversi *mapper*.

Se il numero di sequenze era gestibile da una sola macchina ho scelto di utilizzare il software CLC Genomic Workbench data la sua velocità e la facilità con cui è subito possibile avere una visione d'insieme della qualità del *mapping* e dei possibili errori.

Quando il numero di *reads* non era gestibile dalla *workstation* del nostro laboratorio, ho scelto di utilizzare Bowtie2 sul cluster DIAG.

Questo software gratuito oltre a poter essere eseguito usando computer con memoria condivisa permette, a differenza di altri *mapper*, che pongono delle limitazioni sul numero di mismatch/gaps

per aumentare l'efficienza computazionale, di avere una piena libertà di scelta sui parametri di *mapping*.

Per identificare i geni differenzialmente espressi, sono stati utilizzati DEGSeq (Wang et al., 2010) e edgeR (Anders and Huber, 2010) due pacchetti del software di analisi bioinformatiche Bioconductor (<http://www.bioconductor.org>) che utilizza il linguaggio di programmazione statistico R (<http://www.r-project.org>). Sia DEGseq che EdgeR, pur con alcune differenze, usano un test statistico simile, basato sulla distribuzione binomiale negativa, chiamata anche distribuzione di Pascal.

I dati provenienti da questi due programmi, inoltre, sono facilmente utilizzabili e confrontabili usando una pipeline che automatizza i passaggi di analisi e le comparazione dei risultati denominata DEB (Yao and Yu, 2011).

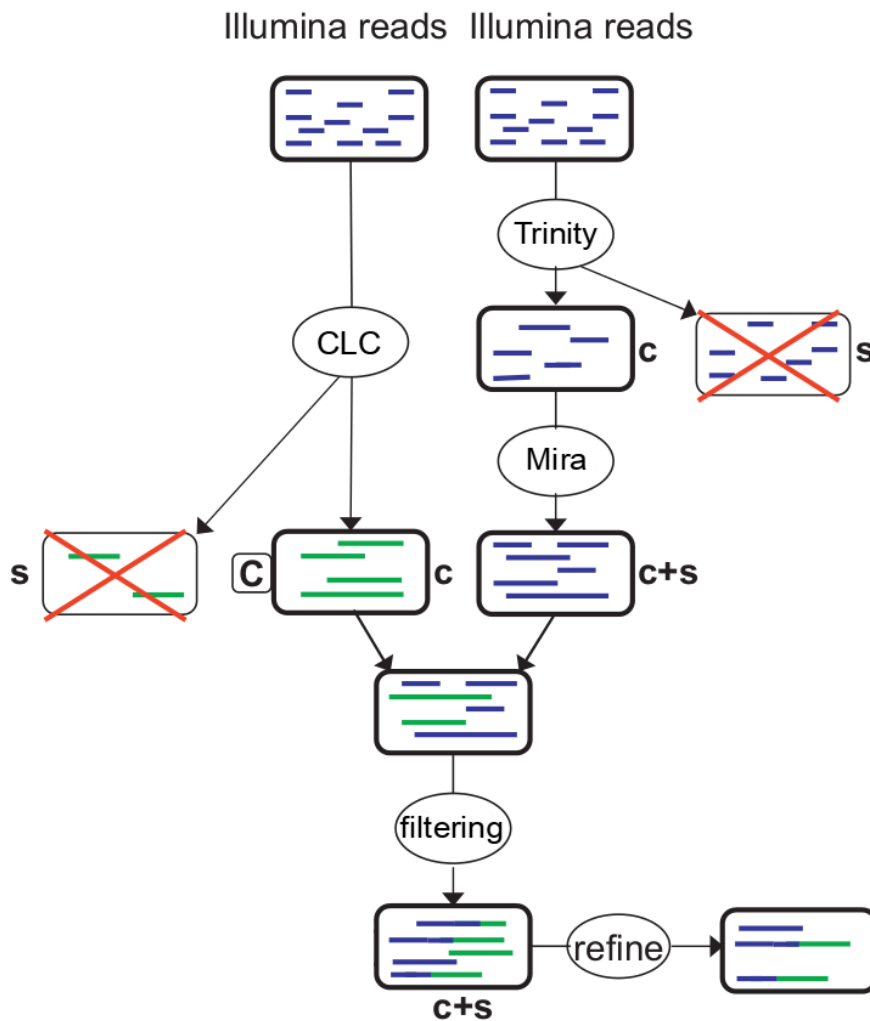


Figura 2 Rappresentazione grafica di come varie strategie di assemblaggio e filtraggio sono stati applicate per ottenere trascritti non ridondanti di alta qualità. C = contig, S = singoli

Un'ulteriore fase nell'analisi di sequenze NGS è l'annotazione delle sequenze ottenute dall'assemblaggio per cercare di dar loro un significato biologico.

I software utilizzati in questa fase sono basati sulla ricerca di analogie di sequenze (es. Blast) tra le sequenze e i dati depositati su vari database biologici e su modelli probabilistici basati su profili Markoviani (es. HMMER).

Per la gestione dei file risultanti da queste analisi si è resa necessario la creazione di programmi e script *ad-hoc* per la loro gestione. Per la quasi totalità degli script creati sono stati utilizzati il linguaggio di programmazione Python (<http://www.python.org>) e Perl (<http://www.perl.org>), e le corrispondenti librerie bioinformatiche Biopython (<http://biopython.org>) e BioPerl (<http://www.bioperl.org>).

Per la durata del dottorato, inoltre, ho fatto un largo uso della metodica chiamata bash-scripting che sfrutta le potenzialità della shell unix Bash.

Se non diversamente specificato tutti gli script e i programmi in Python e Perl citati in questa tesi sono stati da me sviluppati appositamente durante il dottorato.

Con l'utilizzo degli script sviluppati è stato possibile creare nuove strategie di annotazione e testare vari metodi di assemblaggio confrontando e filtrando i vari risultati ottenuti. (Figura 2)

	<i>Mytilus galloprovincialis</i>	<i>Ruditapes philippinarum</i>	<i>Procambarus clarkii</i>	<i>Astacus leptodactylus</i>	<i>Latimeria menadoensis</i>
NGS	Sanger, 454, Illumina	454	Illumina	Illumina	Illumina
Reads	297.948.875	1.288.514	83.170.732	445.265.969	145.435.156
Contigs	206377*	81.410	81.231	91.732	66.308
Media	555*	647	1.036	754	626
N50	586*	755	1.860	1.277	1.761
Trascritto più lungo	21101*	8.748	20.419	23.528	20.815

Tabella 2: Tabella riassuntiva degli organismi studiati, dati ottenuti dal sequenziamento e statistiche d'assemblaggio. * I dati si riferiscono ad un assemblaggio non ancora filtrato

In questa tesi si descrivono le applicazioni di next-generation sequencing a 5 organismi non-modello.

Gli organismi presi in esame sono stati: *Mytilus galloprovincialis*, *Latimeria menadoensis*, *Pontastacus leptodactylus*, *Procambarus clarkii* e *Ruditapes philippinarum*.

Per la maggior parte degli organismi i dati provenivano da sequenziamento Illumina, tranne per *R. philippinarum*, per cui è stato eseguito un sequenziamento 454 e *M. galloprovincialis* di sono stati generati dati derivanti da sequenziamento Sanger, 454 e Illumina. (Tabella 2)

Questi lavori hanno permesso di sviluppare dei database di trascritti espressi nei vari organismi tramite assemblaggio *de novo* e hanno permesso di effettuare degli studi di espressione genica negli individui sottoposti a stimoli diversi.

I database creati rappresentano inoltre un'enorme risorsa che permetterà di allargare gli argomenti di studio sugli organismi sequenziati.

PRODUZIONE SCIENTIFICA DEL CANDIDATO

In questa sezione sono presentati i testi originali e integrali, in lingua inglese, che sono stati prodotti a partire dal lavoro del candidato.

Per ciascuno degli organismi verrà presentata una breve introduzione in italiano del lavoro di analisi informatica svolta.

Di seguito la lista dei lavori allegati, divisi per organismo.

Mytilus galloprovincialis:

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* 2013.

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.

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).

Manfrin,C., De Moro,G., Torboli1, V., Venier, P., Pallavicini, A., Gerdol, **M. Physiological and molecular responses of bivalves to toxic dinoflagellates. ISJ - Invertebrate Survival Journal 9; 2012, 9(2).**

***Latimeria menadoensis*:**

Pallavicini A., Canapa, A., Barucca, M., Alfoldi, J., Biscotti M.A., Buonocore F., De Moro, G., Di Palma, F., Fausto, A.M., Forconi, M., Gerdol, M., Makapedua, D.M., Turner-Meier, J., Olmo, E., Scapigliati, G., **Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. BMC Genomics, submitted. 2013**

Forconi,M., Canapa, A., Barucca, M., Biscotti M.A., Buonocore, Fausto, A.M., Makapedua, D.M., Pallavicini, A., Gerdol,M., De Moro, G., Scapigliati, G., Olmo, E., Schartl, M., **Characterization of sex determination and sex differentiation genes in *Latimeria*. PLoS ONE, accepted. 2013.**

Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Brinkmann, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J.J., Robinson, M., Dorrington, R.A., Gerdol, M., Aken, B., Biscotti, M.A., Barucca, M., Baurain, D., Berlin, A.M., Blatch, G.L., Buonocore, F., Burmester, T., Campbell, M.S., Canapa, A., Christoffels, A., De Moro, G., Edkins, A.L., Fan, L., Fausto, A.M., Feiner, N., Forconi, M., Gamielien, J., Gnerre, S., Haerty, W., Hahn, M.E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S.I., Lara, M., Levin, J., Litman, G.W., Mauceli, E., Miyake, T., Mueller, M.G., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C.P., Prohaska, S.J., Przybylski, D., Saha, N.R., Ravi, V., Ribeiro, F., Sauka-Spengler, T., Scapigliati, G., Searle, S.M.J., Sharpe, T., Simakov, O., Stadler, P.F., Sumiyama, K., Tafer, H., Turner-Maier, J., van Heusden, P., White, S., Yandell, M., Philippe, H., Volff, J.-N., Tabin, C.J., Shubin, N., Schartl, M., Jaffe, D., Postlethwait J.H., Venkatesh, B., Palma, F.D., Lander, E.S., Meyer, A., Lindblad-Toh K., **Comparative analysis of the genome of the African coelacanth, *Latimeria chalumnae*, sheds light on tetrapod evolution, Nature, submitted. 2013**

Forconi M., Biscotti M. A., Barucca M., Buonocore F, De Moro G., Fausto A. M., Gerdol M., Pallavicini A., Scapigliati G., Schartl M., Olmo E., Canapa A., **Characterization of purine catabolic pathway genes in coelacanths, Journal of Experimental Zoology Part B, accepted. 2013**

Pontastacus leptodactylus:

Manfrin, C., Tom, M., De Moro, G., Gerdol, M., Mosco, A., Pallavicini, A., Giulianini1, P.G.,

Hepatopancreatic transcriptome in the crayfish *Pontastacus leptodactylus* reveals peptidase activation and glycolysis suppression following injection of D-crustacean Hyperglycemic Hormone. *Manuscript in preparation*

Mytilus galloprovincialis

Nei primi due anni di dottorato l'attenzione è stata rivolta quasi esclusivamente nella gestione e nell'elaborazione di sequenze trascritte di *Mytilus galloprovincialis* ottenute mediante diverse tecniche di sequenziamento.

Il materiale da cui sono partito era composto da 24.896 sequenze Sanger, 150.857 sequenze ottenute con la metodica di pirosequenziamento 454 e 108.556.255 sequenze Illumina.

In questo lavoro sono state sfruttate le più recenti tecnologie di sequenziamento 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 dinoflagellato è comunemente associato all'avvelenamento di origine marina chiamato Paralytic Shellfish Poisoning (PSP). *M. galloprovincialis* è in grado di accumulare questo tipo di tossine a livelli molto elevati e potenzialmente tossici per il consumo umano.

Oltre a permettere l'identificazione di un'eventuale risposta trascrizionale nel mitilo, successivamente ad un accumulo di tossine paralizzanti, questo lavoro ha permesso di costruire un database di sequenze espresse nelle ghiandola digestiva di questo organismo.

A causa della mancanza di un trascrittoma o genoma di riferimento di *M. galloprovincialis*, infatti, si è reso necessario utilizzare tutte le sequenze a nostra disposizione per creare un set come riferimento per l'analisi di espressione genica.

Il materiale genetico per il sequenziamento è stato estratto dalla ghiandola digestiva in due distinti tempi sperimentali (24 e 48 ore dall'inizio dell'esperimento) da gruppi di 3 mitili trattati con un ceppo di *Alexandrium minutum* produttore di tossine oppure con l'alga non tossica, oltre ad un terzo gruppo di mitili di controllo.

Per prima cosa le sequenze sono state pulite da eventuali adattatori ed è stato fatto un *trimming* per poter eliminare sequenze, o parti di sequenze, a bassa qualità. Questa analisi ha permesso di scartare 45.254 sequenze per la maggior parte proveniente dalla metodica 454, probabilmente a causa dell'intrinseca minor qualità delle sequenze ottenute con questa tecnologia di sequenziamento.

Le 108.686.754 sequenze risultanti sono state assemblate utilizzando il software CLC Genomic Workbench.

In questa fase del lavoro, infatti, non avevo ancora a disposizione l'accesso al cluster DIAG motivo per il quale le scelte dei possibili *assembler* erano piuttosto limitate.

Da questa prima fase di assemblaggio ho ottenuto 110.972 *contig* con una lunghezza media di 590 basi.

Un'ulteriore analisi è stata fatta per ricercare eventuali *contig* ridondanti tramite l'utilizzo di BlastClust, software che permette di raggruppare sequenze tra loro simili. Da questa analisi solamente 203 *contig* sono risultati essere ridondanti denotando, così, una buona qualità dell'assemblaggio ottenuto.

E' seguita una fase in cui si sono cercati eventuali contaminanti all'interno del set di sequenze creato. Per prima cosa sono state cercate, con il programma BLASTn, eventuali similarità in un database contenente sequenze provenienti da dinoflagellati, funghi e piante.

I potenziali contaminanti sono stati ulteriormente sottoposti ad un'analisi utilizzando un set di sequenze del regno Metazoa e successivamente un database di sequenze batteriche.

Quest'ultimo passaggio si è reso necessario per evitare di scartare sequenze di geni altamente conservate che avrebbero potuto essere presenti nel regno animale, vegetale o batterico.

Anche in questa fase abbiamo ottenuto risultati confortanti poiché meno dell'1% delle sequenze risultavano essere potenziali contaminanti.

Per cercare di eliminare *contig* derivanti da un errato assemblaggio, sono stati scartati i trascritti troppo corti o con una copertura molto bassa.

In particolare sono stati mantenuti i *contig* con una copertura media superiore a 10 *reads*, formati da un almeno 50 *reads* e con una lunghezza non inferiore a 250 basi.

Successivamente, i 39.289 *contig* rimanenti, di lunghezza media pari a 689 pb e con un N50 di 814, sono stati annotati utilizzando software pubblicamente disponibili (Blast e InterproScan, HMMER) e script da me creati.

Come prima analisi è stato effettuato un BLASTx mantenendo come *cut-off* un *e-value* pari a $1e^{-5}$ per ricercare eventuali similarità tra i *contig* e le sequenze contenute nel database proteico, rilasciato pubblicamente dall'NCBI, NT/NR.

Da questa analisi 34.157 sequenze hanno avuto un riscontro positivo.

I *contig* sono stati successivamente tradotti in peptidi potenziali usando FrameDP, un software che permette di predire sequenze codificanti partendo da sequenze trascrittomiche.

Poiché nelle sequenze lunghe e con una bassa copertura può presentarsi il problema della frammentazione dei trascritti è stato applicato un metodo chiamato "Ortholog Hit Ratio", metodo che permette di stimare l'integrità dei trascritti.

I risultati ottenuti dal precedente BLASTx contro database NR sono stati processati dividendo la lunghezza della sequenza senza gap della *query* con la lunghezza del *subject*.

A causa della mancanza di genomi completi di bivalvi una corretta predizione della frammentazione dei trascritti non è facilmente ottenibile, ciononostante circa il 50% dei *contig* sono stati assemblati con una lunghezza corrispondente ad almeno il 75% dei loro ortologi.

Un'ulteriore analisi è stata effettuata cercando similarità all'interno del database Superfamily, database di annotazioni funzionali e strutturali di proteine costruito mediante profili markoviani.

Per la ricerca all'interno di questo database è stato usato HMMER, software che si basa sull'utilizzo di modelli probabilistici anch'essi basati su profili di modelli markoviani nascosti.

Dai risultati ottenuti da Superfamily è stato possibile, usando un programma Python appositamente sviluppato, risalire alle annotazioni corrispondenti di Gene Ontology.

Questo metodo per l'annotazione Gene Ontology ha permesso di ottenere i risultati in modo molto piuttosto rapido, sono stati sufficienti infatti solo poche ore di elaborazione.

Le sequenze peptidiche sono state, inoltre, analizzate usando una versione locale di InterproScan in modo da ricercare eventuali motivi conservati e informazioni funzionali e strutturali.

Dopo aver terminato l'annotazione delle sequenze nucleotidiche e peptidiche è stato creato un database MySQL in modo che tutte le informazioni ricavate fossero facilmente utilizzabili.

Successivamente si è cercato di identificare eventuali alterazioni nell'espressione genica, in modo da evidenziare potenziali geni candidati per il monitoraggio della contaminazione in un evento naturale di bloom algale.

Il *mapping* che ha permesso di ottenere la lista dei geni con i loro valori d'espressione è stato fatto utilizzando il software CLC Genomic Workbench e sono stati utilizzati, come valore d'espressione, gli RPKM.

Per aver una maggior affidabilità statistica le differenze di espressione nei campioni sono state analizzate utilizzando edgeR e DEGseq tramite l'utilizzo della pipeline DEB.

Da questa analisi è stato possibile identificare solamente 16 geni come differenzialmente espressi nella ghiandola digestiva degli animali nutriti con il ceppo tossigenico.

Questi 16 geni, assieme a campioni provenienti da un set più di tempi sperimentali, sono stati oggetto di un'analisi con esperimenti di real-time PCR.

I profili di espressione derivanti da questa analisi non sono compatibili con i profili che ci aspetteremmo da geni responsivi all'accumulo di tossine dimostrando, così, che i geni derivanti dall'analisi di RNA-seq sono probabilmente dei falsi positivi.

L'assemblaggio *de novo* e l'annotazione del trascrittoma di mitilo ottenuto ha permesso di implementare enormemente il precedente database di sequenze trascrittomiche di *M. galloprovincialis*, Mytibase.

Questo database, formato da 7.112 *contig* assemblate dalle sole sequenze Sanger, è stato, al momento della pubblicazione nel Febbraio 2009, il più grande database di sequenze trascritte di mitilo disponibile. (Venier et al., 2009)

Questo database è stato alla base degli ulteriori studi, presentati in questa tesi, principalmente volti allo studio di specifiche famiglie geniche, in particolar modo famiglie proteiche coinvolte nell'immunità innata di mitilo, e rappresenta un'importante risorsa come supporto per eventuali studi genomici futuri.

Successivamente alla stesura delle pubblicazioni presentate è stato fatto sequenziare RNA proveniente da branchia che ha prodotto 189.392.620 *reads*.

Con un totale di 298.124.628 di sequenze è stato effettuato un sequenziamento *de-novo* con Trinity che ha portato alla creazione di 222.325 *contig*.

Questi *contig* sono, al momento della stesura di questa tesi, ancora in fase di elaborazione.

RNA sequencing and *de novo* assembly of the digestive gland transcriptome in *Mytilus galloprovincialis* fed with toxinogenic and non-toxic strains of *Alexandrium minutum*

Gerdol Marco¹, De Moro Gianluca¹, Manfrin Chiara¹, Milandri Anna², Riccardi Elena², Beran Alfred³, Venier Paola⁴, Pallavicini Alberto^{1*}

Abstract

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 been only marginally investigated at a molecular level. We used deep RNA sequencing to compare gene expression profiles of the digestive gland in *Mytilus galloprovincialis* fed for 5 days with toxic and non-toxic strains of the dinoflagellate *Alexandrium minutum*.

The gene expression analysis performed in mussel digestive gland indicated that paralytic shellfish toxins (PSTs) scarcely affected mussels, as the few genes identified as possibly differentially expressed in response to toxin accumulation were revealed to be false positives by real-time PCR. Although not conclusive, the overall absence of gene expression changes supports the classification of mussels as bivalves refractory to PSTs and points out that the identification of PSP molecular biomarkers in this organism is problematic. Comprehensive *de novo* assembly of the pre-existing mussel ESTs with the new dataset, and bulk re-annotation of the mussel transcriptome, yielded a collection of 39,289 consensus sequences with an average length of 689 bp, a basic resource for expanding functional genomics investigations in the Mediterranean mussel.

Keywords: *Mytilus galloprovincialis*, *Alexandrium minutum*, paralytic shellfish poisoning, harmful algal bloom

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

Introduction

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 (Humpage et al., 1994), the organisms most commonly associated with PSP are dinoflagellates, such as *Alexandrium catenella*, *A. tamarense* (Cembella et al., 1987) *A. minutum* (Hallegraeff et al., 1988), *A. cohorticula* (Kodama et al., 1988), *A. fundyense* (Schwinghamer et al., 1994), *A. ostenfeldii* (Hansen et al., 1992), *Gymnodinium catenatum* (Mee et al., 1986) and *Pyrodinium bahamense* (Gacutan et al., 1985). Filter-feeding organisms such as bivalve mollusks can accumulate paralytic shellfish toxins (PSTs) at very high concentrations and act as lethal vectors of toxins for organisms at higher trophic levels, including humans.

PSTs are structurally similar to STX (saxitoxin) and their paralytic effects depend on their high affinity to the neuronal voltage-gated sodium ion channels (Terlau et al., 1991). The binding of STX to the channel blocks action potentials, in a similar fashion to tetrodotoxin (Narahashi et al., 1967). The symptoms of intoxication in humans are mainly of a neurological nature and include numbness, tingling, weakness, shortness of breath and ataxia (James et al., 2010). While recovery is generally complete and uncomplicated, in some cases respiratory paralysis and death may occur, especially with the consumption of heavily contaminated mollusks (García et al., 2004).

The widespread occurrence of PSP reflects the broad distribution of the causative algae. In fact, cases of PSP toxicity have been extensively reported in Japan (Hashimoto et al., 2002, Okumura et al., 1994, Takatani et al., 1998), both in the eastern and western coast of Northern America (Gessner and Middaugh, 1995, Jester et al., 2009, Shumway et al., 1994) and in Southern America (Álvarez et al., 2009, Montebruno, 1993), in Britain (Ayres, 1975), and along the Atlantic coasts of the Iberian peninsula (Anderson et al., 1989, Bravo et al., 1999) and France (Amzil et al., 1999). Sporadic cases have also been described elsewhere, i.e. in the Mediterranean Sea (Lilly et al., 2002, Honsell et al., 1996, Ujević et al.). 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 (James et al., 2010), but also causes severe economic damage to the molluscan industry because of the closure of farming areas affected by the algal blooms (Conte, 1984, Anderson et al., 1989). 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 800 μg STX eq kg^{-1} meat. Considering the

ingestion of large quantities of shellfish meat, the European Food Safety Authority established that the limit concentration in shellfish for human consumption should be reduced to 75 $\mu\text{g STX eq kg}^{-1}$, in order not to exceed the acute reference dose (ARfD) of 5 $\mu\text{g STX eq kg}^{-1}$ body weight (EFSA, 2009).

Likewise humans and other vertebrates (Coulson et al., 1968, Geraci, 1989, Kvittek et al., 1991, Cembella et al., 2002), certain bivalve species suffer 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 other species such as *Spisula solidissima* and *Modiolus modiolus* seem to be completely unaffected (Bricelj et al., 1991). Although different species display different behavioral responses to PSP blooms, there is a broad negative relationship between the susceptibility to PSTs and the ability to feed on toxigenic algae and to consequently bioaccumulate toxins (Bricelj et al., 1991, MacQuarrie and Bricelj, 2008). One of the most common behavioral modifications observed in susceptible bivalves is the reduction of filtration rate (Gainey Jr and Shumway, 1988, Basti et al., 2009, Nagai et al., 2006) which could be either interpreted as a paralytic effect or as a strategy adopted to avoid contamination (Tran et al., 2010, Haberkorn et al., 2011). Other mechanisms adopted by susceptible species to reduce the intoxication involve accumulation of PSTs in specific tissues (Sagou et al., 2005, Kitts et al., 1992), binding to sequestering proteins (Takati et al., 2007), enzymatic or chemical transformation and degradation reactions (Tian et al., 2010, Oshima, 1995, Sullivan et al., 1983), even though it is not clear whether this latter processes depend on bivalve metabolism or on symbiotic bacteria (Smith et al., 2001, Donovan et al., 2008).

Electrophysiological studies indicated that mussel nerves are insensible to the paralytic effects of STX (Twarog et al., 1972, Twarog, 1974). Such a resistance may reflect adaptive evolution to recurrent toxic algal blooms (a direct link between the sensitivity to PSTs and frequency of red tides has been observed in clam populations) and may be explained by sodium channel mutations leading to a decreased affinity to PSTs in resistant populations (Bricelj et al., 2005, Connell et al., 2007).

Due to the substantial lack of physiological and behavioral changes in response to the feeding with PSP-producing dinoflagellates, mussels are considered refractory to PSP (Bricelj et al., 1990, Marsden and Shumway, 1993). On the other hand, increased valve closure, decreased filtration rates and reduced byssus production have been occasionally observed and these symptoms have been associated to the increased mortality of *M. edulis* fed with toxic *A. tamarensis* (Shumway and Cucci, 1987, Shumway et al., 1987) and to the extensive histopathological modifications described in blue mussels exposed to *A. fundyense* (Galimany et al., 2008).

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. To the best of our knowledge the only study about the effects of an algal toxin in bivalves ever performed at a whole-transcriptome scale concerns a purified okadaic acid (Manfrin et al., 2010). 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.

The advent of next generation sequencing has definitely expanded large-scale molecular studies to non-model invertebrates (Pérez-Enciso and Ferretti, 2010). Based on the 454 (Milan et al., 2011, Hou et al., 2011, Clark et al., 2010, Craft et al., 2010, Bettencourt et al., 2010, Joubert et al., 2010, Philipp et al., 2012), SOLiD (Gavery and Roberts, 2012) and Illumina (Ghiselli et al., 2012) technologies, the massive sequencing of bivalve transcriptomes is revealing the molecular basis of the functional responses to environmental changes and paving the way to an improved view of the evolutionary relationships within mollusks (Smith et al., 2011, Kocot et al., 2011).

In the present study, we investigated the response of the Mediterranean mussel to PSTs bioaccumulated in vivo by comparing the transcription profiles of digestive gland samples from animals fed with toxigenic or non-toxigenic strains of the dinoflagellate *A. minutum* via Illumina RNA sequencing. The transcriptional analysis performed on the digestive gland did not reveal useful biomarkers of mussel exposure to PST but the overall assembly of the new sequencing reads with transcript sequences previously obtained (Venier et al., 2009) significantly enriched the overall knowledge of *M. galloprovincialis* transcriptome, thus helping us to create one of the most relevant sequence collection existing to date in the Mollusca phylum.

Results and Discussion

Toxin accumulation

Concentrations of *A. minutum* varying from 1 to 47×10^6 cells L⁻¹ have been reported in toxic blooms (Delgado et al., 1990, Maguer et al., 2004, Garcés et al., 2004, Galluzzi et al., 2004, Van Lenning et al., 2007). We exposed adult *M. galloprovincialis* individuals for five days to 5×10^6 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. A third group of animals, not subjected to a forced diet

based on dinoflagellates, was used as a control. Mussels were sacrificed before daily feeding at seven time points during 5 days of intoxication and 6 days of depuration in order to collect digestive glands for gene expression analysis and estimate the bioaccumulation of PSTs at selected time points. The experimental design is summarized in **Table 1** and detailed in Methods.

According to the HPLC analyses, the *A. minutum* strain AL9T produced an average concentration of 76,4 fg STX diHCleq\cell, whereas the strain AL1T did not produce any toxins, as expected. The estimate of toxin bioaccumulation was performed on the soft mussel tissues, after the digestive gland was taken apart for RNA extraction. The levels of PSTs in the remaining tissues and estimated on 3 individuals, resulted to be about 100 μg STX eq / kg of meat at T4 (5 days from the start of the experiment). Visceral organs are known to accumulate approximately 95% of PSTs in mussel (Bricelj et al., 1990): considering the removal of the digestive gland, the accumulation of PSTs at T4 could be estimated around 2,000 μg STX eq kg^{-1} of meat, well above the EU and US limits (set at 800 μg STX eq kg^{-1}). Although accumulation of PSTs 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, time-course studies previously published pointed out that mussels accumulate paralytic toxins at very high rates, resulting high toxic in the matter of a few hours (Blanco et al., 2003, Bricelj and Shumway, 1998, Navarro and Contreras, 2010).

***De novo* assembly of the digestive gland transcriptome**

The Illumina sequencing of the digestive gland samples (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 sequenced reads per sample. The raw Illumina reads have been deposited at the NCBI Sequence Read Archive (study ID: SRP011280.2). Aiming to refine the data set, the trimmed Illumina reads from the whole sample series were preliminarily assembled together with the pre-existing Sanger and 454 Life Sciences sequences from various tissues and challenges (18,788 and 115,557, respectively) plus an additional set of Illumina reads from the digestive gland of naïve mussels (28,186,684). The processing of sequence consensus was carefully performed to overcome the creation of short and low quality or misassembled contigs, a problem commonly arising from the assembly of next generation sequencing data (Feldmeyer et al., 2011), and to remove contaminant sequences (mainly originated by ingested *A. minutum* cells). Contig filtering provided a remarkable improvement of the assembly quality (**Figure 1**), producing a shift of the contig length towards higher ranges, hence reducing the bias towards short, incomplete contigs.

Overall, the final assembly yielded a high quality collection of *M. galloprovincialis* transcripts (39,289 consensus sequences) (**Table 2**). The average contig length was 689 bp (from 250, the minimum length allowed, to 13,211 bp) and the N50 statistic of the assembly was 814.

RNA-seq expression analysis

The *de novo* assembly described above generated the reference contigs used for the subsequent mapping of RNA-seq data for the expression analysis. Nevertheless, a further filtering step was applied prior to analysis in order to remove transcripts whose expression was too low. As a result, only 5,523 contigs with a global average coverage higher than 5 were selected. The representation of poorly expressed transcripts is highly dependent on the total number of reads obtained from a sample and it has been demonstrated that not negligible random variability can occur even between technical replicates when the sequence coverage is particularly low (McIntyre et al., 2011). Overall, the removal of these contigs guaranteed the achievement of a less noisy dataset, less prone to false positive detection.

Since several different methods for differential expression detection in RNA-seq experiments have been developed, based on different mathematical assumptions (Oshlack et al., 2010), we chose to perform the statistical analyses with 3 different algorithms, namely EdgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010a) and the Baggerly's test on proportions included in the CLC Genomic Workbench (Baggerly et al., 2003). The analyses revealed a rather low number of genes as significantly differentially expressed (FDR <0.01) in the AL9T toxigenic strain-fed mussels both at T1 and T4. More in detail, only 20 genes were identified by EdgeR, 65 genes by DESeq and 258 by the Baggerly's test. A schematic representation of the results is summarized in the Venn diagram in **Figure 2**.

The comparison of the results revealed that only five transcripts were found differentially expressed by all the 3 algorithms. These sequences, all up-regulated in the PSP-contaminated mussels (**Table 3**), were selected as the most likely PSP-responsive candidate genes. Nevertheless, the functional classification could not directly link any of these 5 sequences to PSTs accumulation. More in detail, 3 contigs were found to have no similarity with other known sequences (and were therefore named "transcripts of unknown function"), reflecting the low representation of molluscan sequences in public databases and the difficult annotation of poorly conserved sequences. One contig pertained to the C1qDC family which is an extremely large class of immunity-related lectin-like molecules, possibly including several hundred genes in bivalves (Gerdol et al., 2011) and representing the largest group of sequences expressed in *M. galloprovincialis*. In this study, the overall abundance of C1qDC transcripts is close to 1.5%, with 580 contigs being annotated as containing the C1q domain

IPR001073 (**Table 4**). Finally, one contig was an IMAP family GTPase characterized by the presence of AIG1, another domain very common in mussel (IPR006703) (**Table 4**).

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 analysis was performed on samples from all the available time points (T0 to T7, see **Table 1**) to monitor the expression trends of 6 selected transcripts. We selected 3 out of the 5 sequences identified as differentially expressed by EdgeR, DEseq and the Baggerly's test (transcript of unknown function I and II and C1q domain containing protein I). Furthermore, since the involvement in PSP contamination of other genes identified by just one or two of the algorithms (see **Figure 2**), although less probable, couldn't be completely ruled out, we also selected 1 sequence identified by both EdgeR and DEseq (hemicentin 1-like) and 2 exclusively detected by the Baggerly's test (retinol dehydrogenase and C1q domain containing protein II).

The results confirmed the data obtained by the RNA-seq experiment at T1 and T4, showing significantly different expression values in the AL9T strain-fed mussels in all cases, except from retinol dehydrogenase at T1 (**Figure 3**). On the other hand, the analysis also revealed remarkable fluctuations in the expression levels of these genes throughout the intoxication and depuration phases, apparently independent from the bioaccumulation of PSTs (data not shown). Therefore, the real-time PCR analyses, while confirming the experimental data obtained with RNA-seq at T1 and T4, also showed that the changes observed were likely the result of random expression fluctuations, leading to an apparent responsiveness to PSTs at the two time points analyzed by RNA-seq (T1 and T4).

Overall, these results suggest that most, if not all, the genes identified as differentially expressed by the expression analyses were not connected to PSP accumulation in any way. Therefore, given the absence of trustworthy PSP-responsive genes, the accumulation of toxins achieved in our experiment likely didn't produce any remarkable effect on the transcriptomic profile of digestive gland in mussel.

Transcriptome annotation

Following global assembly, 34,157 contigs (86,94%) found BLASTx hit similarity in the NCBI nr protein database, with the number of unknown sequences being significantly lower than those previously reported for Mytibase (Venier et al., 2009). Not surprisingly, the most represented species identified by the top BLAST hits are invertebrates whose genome has been fully sequenced and released, namely the cephalocordate *Branchiostoma floridae*, the hemicordate *Saccoglossus kovalenskii*, the echinoderm *Strongylocentrotus purpuratus* and the cnidarian *Nematostella vectensis*

(**Figure S1**). Despite the few *Mytilus* sequences available in the nr database, *M. galloprovincialis* stands as the fifth top hit species.

Due to the limited sequence resources available for mollusks in public databases, homologies have been investigated in the only fully sequenced molluscan genome available to date, the gastropod snail *Lottia gigantea* (Grigoriev et al., 2012). About half (54%) of the assembled contigs displayed significant similarity (e-value $< 10^{-6}$) to proteins predicted from the snail genome, with highly significant e-value ($< 10e^{-50}$) in the 16% of the cases (**Figure S2**).

To address the problem of transcript fragmentation, an ortholog hit ratio analysis was performed (O'Neil et al., 2010). Since this measure is strongly influenced by the availability of sequence data from closely related organisms, due to evolutionary divergence resulting in sensible underestimations, we modified the test of O'Neil as described in the Methods section, by only considering “true orthologs”. The ortholog hit ratios distribution (**Figure 4**) shows that approximately 25% contigs were assembled to a length corresponding to $>90\%$ of their ortholog. About 40% contigs were assembled to less than 50% of their hypothetical full length, pointing out that increased sequencing depth and RNA-seq experiments from additional tissues would be required to improve full-length transcript reconstruction.

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 GO assignments (molecular function) several of the most abundant domains (i.e. immunoglobulin-like, ankyrin, C1q, etc.) are characterized by marked binding properties.

On the basis of the sequence homologies identified with BLAST and InterPro domains annotations, 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 summary of GO classification is shown in **Figure S3**. 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, evidence supported by the relevance of “cellular processes” and “metabolic processes” among other biological processes, with most transcripts located within the cell while only a minority resulted to be located in organelles, macromolecular complexes or in the extracellular environment. The high metabolic activity of the digestive gland was also confirmed by exploring the level 3 GO terms, as the “primary metabolic process”, “cellular metabolic process” and “macromolecule metabolic process” resulted to be the three most abundant categories (**Table S4**).

Comparison with Mytibase

The sequence data generated by RNA-seq in the present study provide a limited view of the entire complement of transcripts expressed in different tissues, different life stages and in response to many and fluctuating biotic and abiotic stimuli in *M. galloprovincialis*, as it has also been highlighted by the fragmentation of a relevant proportion of contigs (**Figure 4**), linked to low coverage. Since the digestive gland has already been reported as a tissue characterized by the expression of transcripts having all together a broad spectrum of functions, we mapped the RNA-seq data from the digestive gland of mussels fed with *A. minutum* on Mytibase, a collection of 24,937 Sanger ESTs, assembled into 7,112 consensus sequences, derived from different tissues of normal, treated and immunostimulated mussels (Venier et al., 2009).

An overview of the mapping statistics is shown in **Figure 5**. A large proportion of Mytibase contigs displayed an average coverage higher than 100X (33 % of the total) or comprised between 10X and 100X (39%) whereas a limited number of the Mytibase transcripts was not present (7%) or expressed at very low levels (5%, average coverage <1X) in the new dataset, indicating that the large majority of the sequence data generated by the previous Sanger sequencing efforts were virtually included within the new digestive gland sequence dataset. Compared to the 51% of the Illumina reads finding a match on Mytibase sequences, 49% the sequence data could not be mapped and were therefore originated from transcripts not included in Mytibase, contributing to the *de novo* assembly of novel transcripts.

Overall, the sequencing depth applied to this study was high enough to obtain a good coverage also of genes expressed at relatively low levels in the mussel digestive gland. Nevertheless, a certain number of Mytibase transcripts (7%) not found in the new dataset could be the product of genes whose expression is strictly regulated or extremely specific of tissues other than the digestive gland. This data, together with the ortholog hit ratio results, suggests that RNA-seq should be performed from additional tissues in order to obtain a comprehensive overview of the *M. galloprovincialis* transcriptome.

The comparison between the relative abundance of specific functional domains within the two datasets (*de novo* assembly vs Mytibase) revealed a perceivable enrichment of many common Interpro signatures of Metazoans (e.g. immunoglobulin-like and zinc-finger C2H2, as reported in **Table 4**), most closely approaching the expected frequencies from a complete transcriptome. On the contrary, Interpro signatures closely associated to immunity-related functions, such as C1q, C-type lectin-like and fibrinogen C-terminal globular domain were under-represented in the new dataset compared to Mytibase (see **Table 4**), additionally confirming Mytibase as a valuable source of immune- and defense-related transcripts (Venier et al., 2011). Therefore, the new transcript collection

obtained from the RNA-sequencing of the digestive gland integrates and substantially enriches Mytibase and provides the basis for specific studies as illustrated by the recent description of novel defense peptides using a whole-transcriptome mining approach (Gerdol et al., 2012).

Conclusion

This study provides the first comprehensive analysis of the transcriptional effects of bioaccumulated PSTs in a molluscan species (*M. galloprovincialis*).

The analysis of the expression profiles revealed that paralytic toxins did not affect mussels, at least not at the concentrations reached in the experiment (about 2,000 $\mu\text{g STX eq kg}^{-1}$ 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 (Bricelj et al., 1990) or just observed a mild early response followed by an extremely rapid acclimatization (Blanco et al., 1997, Blanco et al., 2003, Fernández-Reiriz et al., 2008) (even though this response could be merely related to the adaptation to a different alimentation regime). Our study provided the first molecular lines of evidence supporting the classification of mussels as organisms not responsive to PSP, as no significant alteration of gene expression was observed in the digestive gland.

The occasional reports of PSP adverse effects on mussels (Shumway and Cucci, 1987, Galimany et al., 2008) did not find any confirmation in our result. Nevertheless, these observations are not necessarily contradictory, as different responses could be linked to inter-population variability in the sensitivity to PSTs, in a similar fashion to other mollusk species (Connell et al., 2007).

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 (EFSA, 2009), and as a strategy to minimize the possibility of PSP contamination in the aquaculture sector (Desbiens and Cembella, 1993). 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.

The new sequencing data allowed a novel global assembly of the *M. galloprovincialis* transcriptome. RNA deep sequencing had already been applied to a few bivalve mollusks species (Clark et al., 2010, Craft et al., 2010, Gavary and Roberts, Hou et al., 2011, Milan et al., 2011, Philipp et al., 2012), but

this is the first Illumina technology-based sequencing effort ever reported in *Mytilus*. The resulting transcript sequence collection remarkably improves the existing database Mytibase, further revealing the variety of genes expressed in the digestive gland tissue. Nevertheless, further RNA sequencing of different tissues would be needed to obtain a comprehensive overview of the mussel transcriptome. The newly annotated sequence set will certainly provide an important resource for improving the molecular knowledge of this species and will be the basis for further studies requiring whole-transcriptome mining approaches.

Methods

Mussel specimens

Adult *Mytilus galloprovincialis* (Lamarck, 1819) were obtained from a commercial producer of 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 $2,48 \pm 0,42$ g) were acclimated at 15°C and 32‰ salinity for one week in running prefiltered seawater and for 3 days in bacteria-free filtered seawater (Millipore Durapore GV $0,22 \mu\text{m}$, hydrophile PVDF) at 12:12 h dark:light regime. Mussels were tested by HPLC before the start of the experiment and were found free of PSP toxins.

Alexandrium minutum cultures

The AL1T (non-toxicogenic) and AL9T (toxin producing) strains of *A. minutum*, previously isolated from the Gulf of Trieste, were cultured in medium B (Agatha et al., 2004) in a suitable number of aerated 1 L batch cultures. The cultures were maintained at 15°C at 10:14 h dark:light regime with an irradiance of $60 \mu\text{E m}^{-2} \text{s}^{-1}$. Algal cells were harvested in the late exponential phase of growth.

Both strains were tested at the time points T1 and T4 (relevant for RNA-seq analysis) for the production of PSTs as described below in Toxin analysis: 100 ml of culture were filtered on Millipore Durapore GV $0,22 \mu\text{m}$ filters and immediately frozen at -18°C for HPLC analysis. The typical toxin profile of the AL9T strain is shown in **Figure S5**.

Experimental design

Mussels were maintained in standard conditions in glass tanks containing 0.4 L of $0,22 \mu\text{m}$ filtered seawater per mussel. Water was renewed every morning at 9 AM with filtered bacteria-free seawater. The overall work plan is outlined in **Table 1**.

A total of 6 tanks were prepared for the exposure to *A. minutum*: 3 sets hosted the AL1T (non-toxicogenic) cells and the remaining 3 the AL9T (toxicogenic) cells. During the 5 days of intoxication, a

dose of 2×10^6 cells of *A. minutum* per mussel was added every 2 hours, 5 times a day, beginning at 10 AM. Then, mussels were allowed to depurate with regular water renewal but without food supply for other six days. At selected time points, always at 9.00 AM, one mussel per aquarium was sacrificed for further analyses. Namely at 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).

Three additional tanks were kept as a “standard diet” controls. Mussels were fed once a day at 9 AM with 36 mg marine invertebrate feed (Brightwell Reef Snow) per animal. One mussel per tank was sacrificed at T1 and T4 to provide the control material for the RNA-seq analysis.

Toxin analysis

The analysis of the PSTs was performed on the *A. minutum* cells and soft mussel tissues at the time points T0, before the first feeding dose, and T4 when the maximum bioaccumulation of toxins was supposedly achieved. The PSTs detection was based on pre-column oxidation and High Performance Liquid Chromatography coupled to Fluorescence Detection (HPLC-FLD) according to the protocol AOAC 2005.06 (Lawrence et al., 2004).

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

From each single mussel, whole body tissues deprived of the digestive gland (used in parallel for RNA extraction) were homogenised and tissue aliquots equivalent to 1.7 g were analysed. Following preliminary sample oxidation with both peroxide and periodate, the HPLC-FDL 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 (Quilliam et al., 1993). Toxins were quantified against linear calibrations of all currently-available PSP toxin certified reference standards and the toxicity equivalence factors (TEFs) proposed by the CONTAM Panel (EFSA, 2009) were used to calculate STX-equivalent concentrations and to estimate the concentration of PSTs in the whole mussel tissues.

RNA extraction and analysis

Digestive glands were excised from 1 mussel per aquarium at each of the selected time points during the exposure and recovery period (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 was prepared by retro-transcription with the iScript™ cDNA Synthesis Kit (Bio-Rad) and used for Real Time quantitative PCR. RNA extracted from the 3 individual mussels sampled at each experimental time point from the two groups (the AL1T and AL9T *A. minutum* strains-fed mussels) were pooled in equal quantities and used for the RNA-seq analysis and for the expression analysis by real-time PCR, according to the scheme outlined in **Table 1**. RNA pools, comprising 3 individuals each, were also prepared from the 3 control aquariums at T1 and T4 respectively and used for the RNA-seq analysis.

Sequencing and *de novo* transcriptome assembly

cDNA libraries were prepared and subjected to massive sequencing at the Biotechnology Center of the University of Illinois, using an Illumina GAII sequencing platform. The output sequencing reads were further processed for adapter removal and trimming, according to the base calling quality. The resulting sequences were assembled with the CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark) assuming a distance of 100-350 bp between paired reads, 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 process included the 18,788 Sanger sequences of Mytibase (Venier et al., 2009); (mismatch\insertion\deletion cost set at 3\3\2, length fraction and similarity at 0.2\0.9), additional 115,557 reads obtained from different tissues of mussels by 454 Life Sciences sequencing (gap\insertion\mismatch penalties set at 2\2\2, length fraction\similarity at 0.4\0.8) and also 28,186,684 Illumina reads obtained from the digestive gland of naive mussels (same settings stated above for Illumina reads). The minimum contig length allowed in the assembly was set at 250 base pairs.

The resulting contigs were filtered to eliminate sequences originated from ingested *A. minutum* cells and contaminants such as symbiont bacteria and parasites 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 at least a 10^{-5} factor were discarded as probable contaminants. Furthermore, a BLASTn analysis was conducted against an assembly of the *A. catenella* ESTs obtained by Toulza et al. (2010), discarding the contigs showing

identity and possibly originated from the ingestion of *A. minutum* cells. Contigs created by less than 50 sequencing reads were not included in the final “high quality” set of expressed sequences. Finally, all the transcripts without an open reading frame of at least 50 codons, the possible result of fragmentation or misassembly of longer transcripts, were discarded before the annotation step.

Transcripts annotation

The BLASTx algorithm (Altschul et al., 1997) was used to determine the contig 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 (Conesa and Götz, 2008), and Gene Ontology mapping and InterPro domains (Hunter et al., 2009) annotation were performed using the default settings. The Gene Ontology mappings 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 from this organism (<http://genome.jgi-psf.org/Lotgi1/>).

Ortholog hit ratios were calculated using a modified version of the method of O’Neil et al. (O’Neil et al., 2010), based on the BLASTx output, analyzing only the contigs displaying identities >90%, in order to select only conserved orthologs, thus balancing for the evolutionary divergence and the low representation of mollusk sequences in public databases.

Expression analysis by RNA-seq

The filtered contigs were further processed prior to the expression analysis to generate a suitable reference set for the RNA-seq mapping of the reads originated from each of the six analyzed samples (T1 toxic, T1 non-toxic, T4 toxic, T4 non-toxic, T1 control and T4 control). Contigs displaying a global coverage lower than 5 (calculated by the mapping of all the sequencing reads from all the six samples) were discarded prior to the analysis, as they could be subject to expression fluctuations due to insufficient coverage (McIntyre et al., 2011).

Raw counts from the six samples were used in the statistical analysis to identify differentially expressed transcripts with DEB (Yao and Yu, 2011), which simultaneously analyzes data with edgeR (Robinson et al., 2010) and DEseq (Anders and Huber, 2010b) and with the Baggerly’s test on proportions tool included in the CLC Genomic Workbench (Baggerly et al., 2003). The analysis aimed at the identification of differentially expressed genes in response to PSP, independently from the mussel diet.

To this purpose, the two “standard diet” and the two AL1T strain-fed experimental samples 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.

Quantitative PCR expression analysis

Six transcripts among those identified as differentially regulated in response to PSP contamination by the edgeR, DEseq and Baggerly’s test analyses were selected to perform the expression analysis via real-time quantitative PCR. Namely, transcript of unknown function I and II, C1q domain-containing protein I and II, hemicentin 1-like and retinol dehydrogenase 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 of mussels fed with the AL1T and AL9T strains.

All the PCR assays were performed using a Bio-Rad CFX96 system. The 15 μ L reaction mix included 7.5 μ L of 2x IQ™ SYBR Green® Supermix (Bio-Rad), 0.3 μ L of each 10 μ M primer and 2 μ L 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- $^{-Ct}$ Ct method) (Livak and Schmittgen, 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 5**). Results are given as the mean with standard deviation of three technical replicates.

Tables

day	experimental time point	feeding	RNA-seq*	Real Time PCR*	phase
0	T0	x		X	intoxication 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: Experimental plan. Mussels were kept in 6 tanks and subjected to alternative feeding regimes (3 tanks with the *A. minutum* toxigenic strain AL1T vs 3 tanks with the non-toxigenic strain AL9T) and collected at 7 time points during the intoxication and detoxification phases. Three additional tanks were used as controls for the gene expression analysis by RNA-seq (mussels were fed with a “standard diet”, see the Methods section). *One mussel per aquarium was sacrificed at each time point before the feeding and analyses 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 length 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
Additional sequences used for the assembly	
Illumina (digestive gland)	28,186,684
454 (various tissues)	115,557
Sanger (various tissues, Mytibase collection)	18,788
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 of the Illumina sequencing output, number of reads per sample and *de novo* assembly statistics.

Transcript name	EdgeR FDR	DEseq FDR	Baggerl y's Test FDR	expression level (RPKM)							
				T1		T4		T1		T4	
				cont rol	T4 control	T1 non- toxic	T4 non- toxic	T1 toxic	T4 toxic		
Transcript of unknown function (I)*	4.72e-06	1.66e-05	0	0	0	0	0	37,71	25,45		
Transcript of unknown function (II)*	2.20e-04	7.41e-05	1.47e- 05	0	0	0	0	18,36	19,82		
Transcript of unknown function (III)	2.63e-04	1.40e-04	4.49e- 05	0	0	0	0	22,33	24,11		
C1q domain containing protein*	3.69e-03	9.79e-03	1.59e- 04	0	0	2,1	0	33,37	25,26		
IMAP family GTPase	3.00e-04	2.10e-04	0	0	2,65	0	0	82,42	47,86		

Table 3: List of the 5 differentially expressed genes identified by the edgeR, DEseq and the Baggerly's test analyses. The related expression levels detected by RNA-seq in the control, AL1T and AL9T-fed mussels are also shown. RPKM = Reads Per Kilobase per Million reads mapped

*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 Major facilitator superfamily domain, general substrate	234	58	0,73
IPR016196	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 Fibrinogen, alpha/beta/gamma chain, C-terminal	208	38	0,99
IPR014716	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 *de novo* assembly according to the Interproscan assignments. *This value represents the rate between the number of contigs observed and the number of contigs expected in the assembly; expected numbers were calculated based on the relative abundances observed in Mytibase . 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	cctcccacatcaagaccta	ggctggagcaaaggtaacaa
transcript of unknown function I	tcagcgtagcacctttacca	ccatctggcaaagccttact
transcript of unknown function II	acagcttgaaacggaccttc	tattcacgtgccttgcctc
C1q domain containing protein I	gacaactcaaggcgcattgt	ttcaaaggtagaccctca
C1q domain containing protein II	catacatgccgaacatagc	gataccaagaccaggagca
retinol dehydrogenase	aggagcaggcatagcgtagt	aaagctcgttaccgggtgtg
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.

Figures

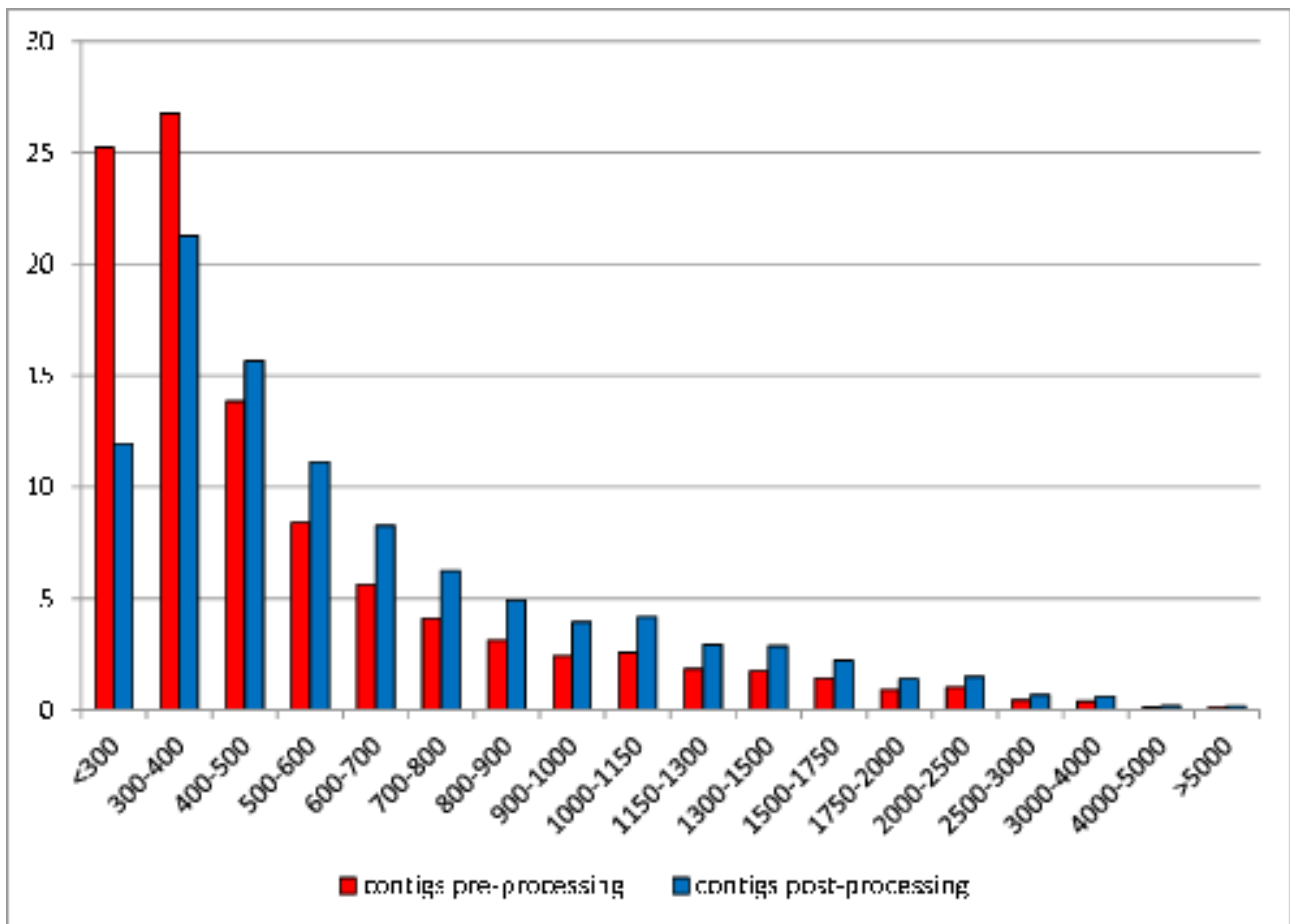


Fig. 1

Contig length distribution, before and after the filtering procedures. The graph highlights a shift towards higher length ranges in the high quality set

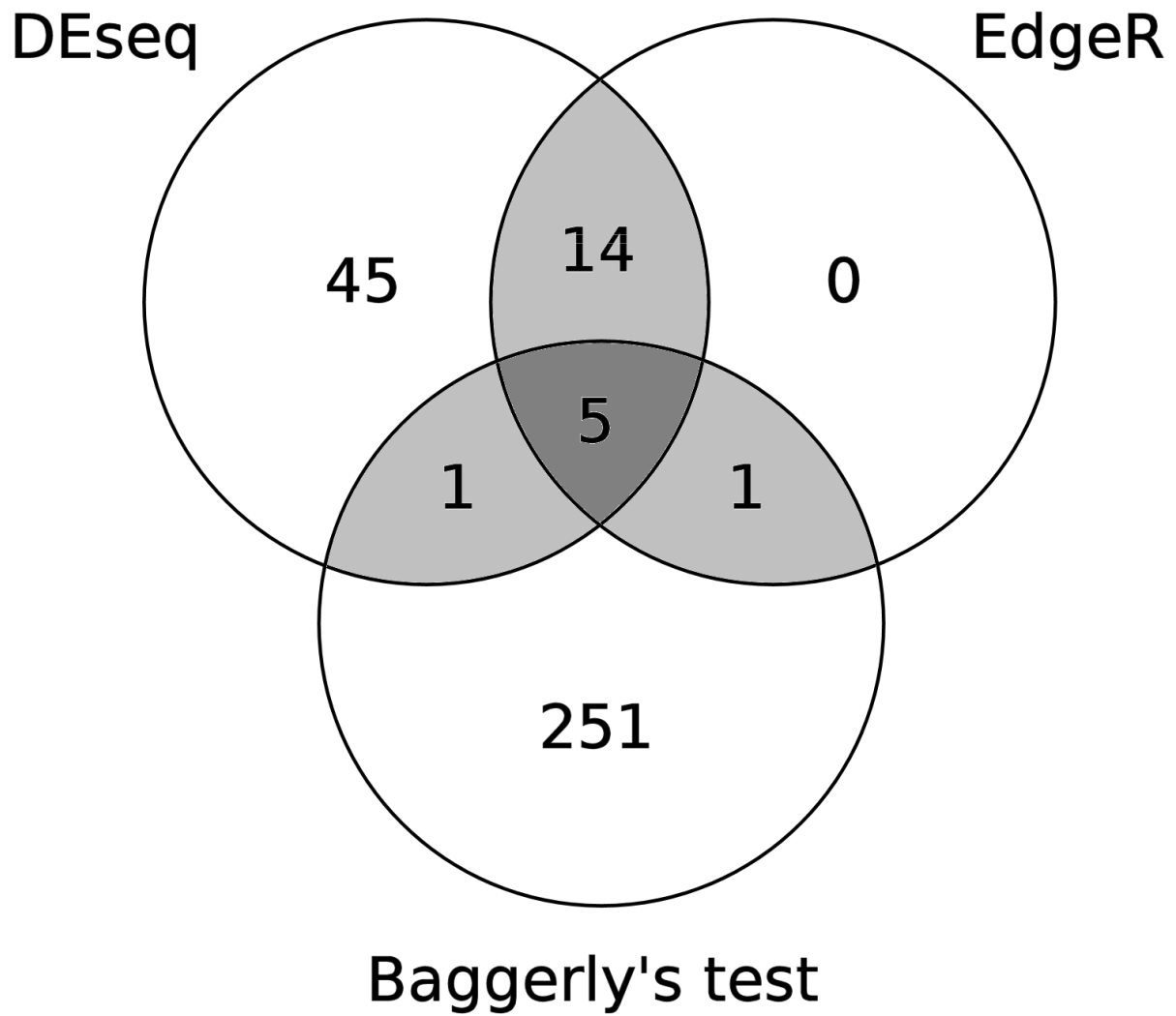


Fig. 2

Venn diagram summarizing the results of expression analysis. The numbers shown in the graph represent the number of differentially expressed genes identified by each of the three statistical tests used (EdgeR, DEseq and the Baggerly's test on proportions) and the overlap between the results

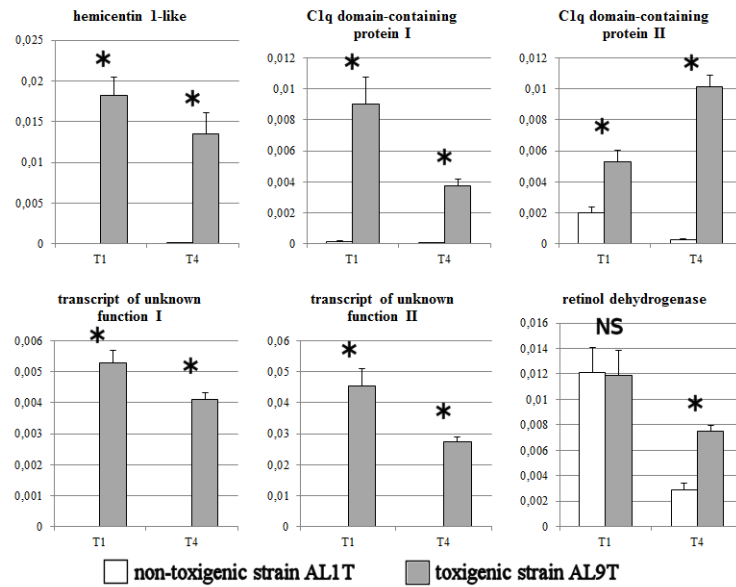


Fig. 3

Expression levels at the experimental time points T1 and T4 of six selected PSP-responsive genes (transcript of unknown function I and II, Clq domain-containing protein I and II, hemicentin 1-like and retinol dehydrogenase). Statistically significant differences between non-toxicogenic (AL1T) and toxicogenic (AL9T) strain-fed mussels are indicated by * ($p < 0,01$). NS = no significant difference. Expression data was collected as described in Methods. Expression values shown on the Y axis are relative to the housekeeping gene EF-1

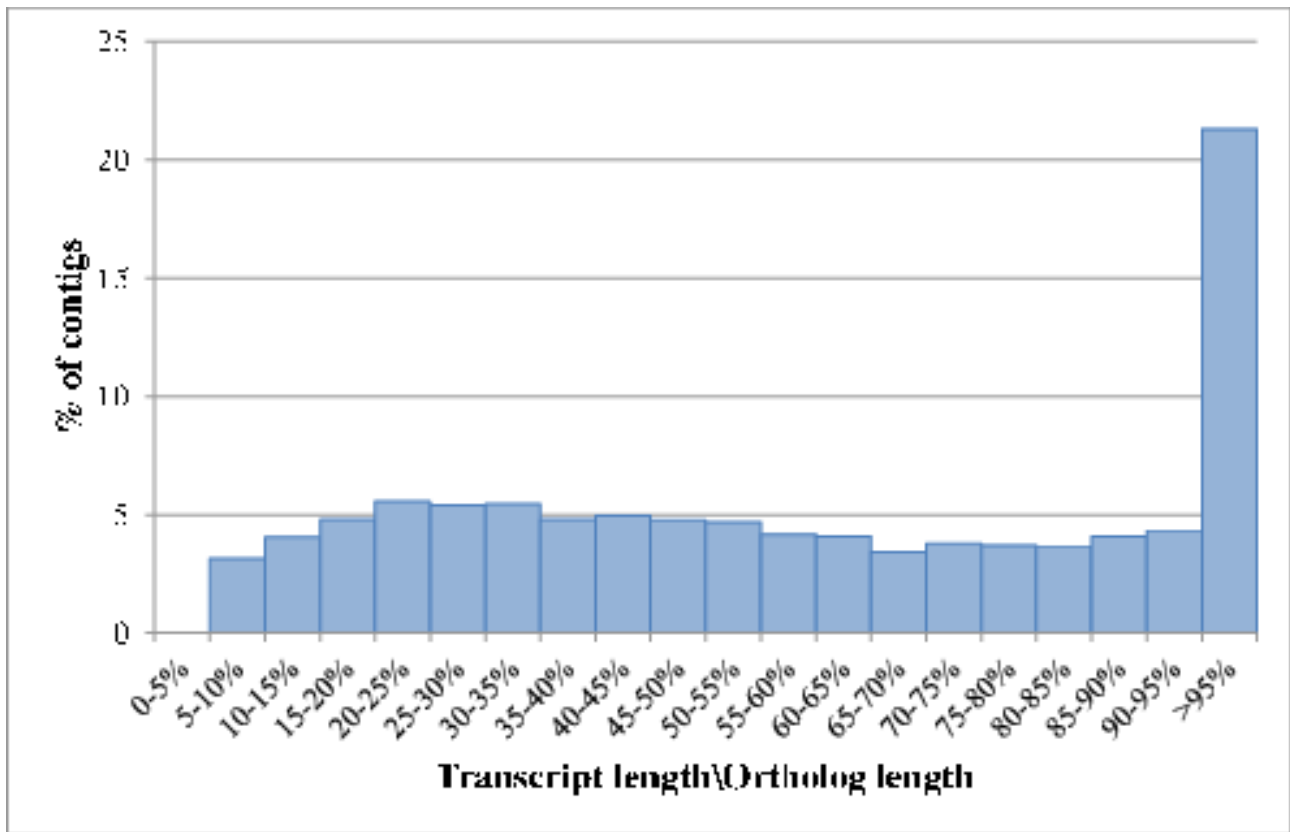


Fig. 4

Distribution of ortholog hit ratios obtained from the BLASTx of the *M. galloprovincialis* high quality contig set vs NCBI nr 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 >95% category

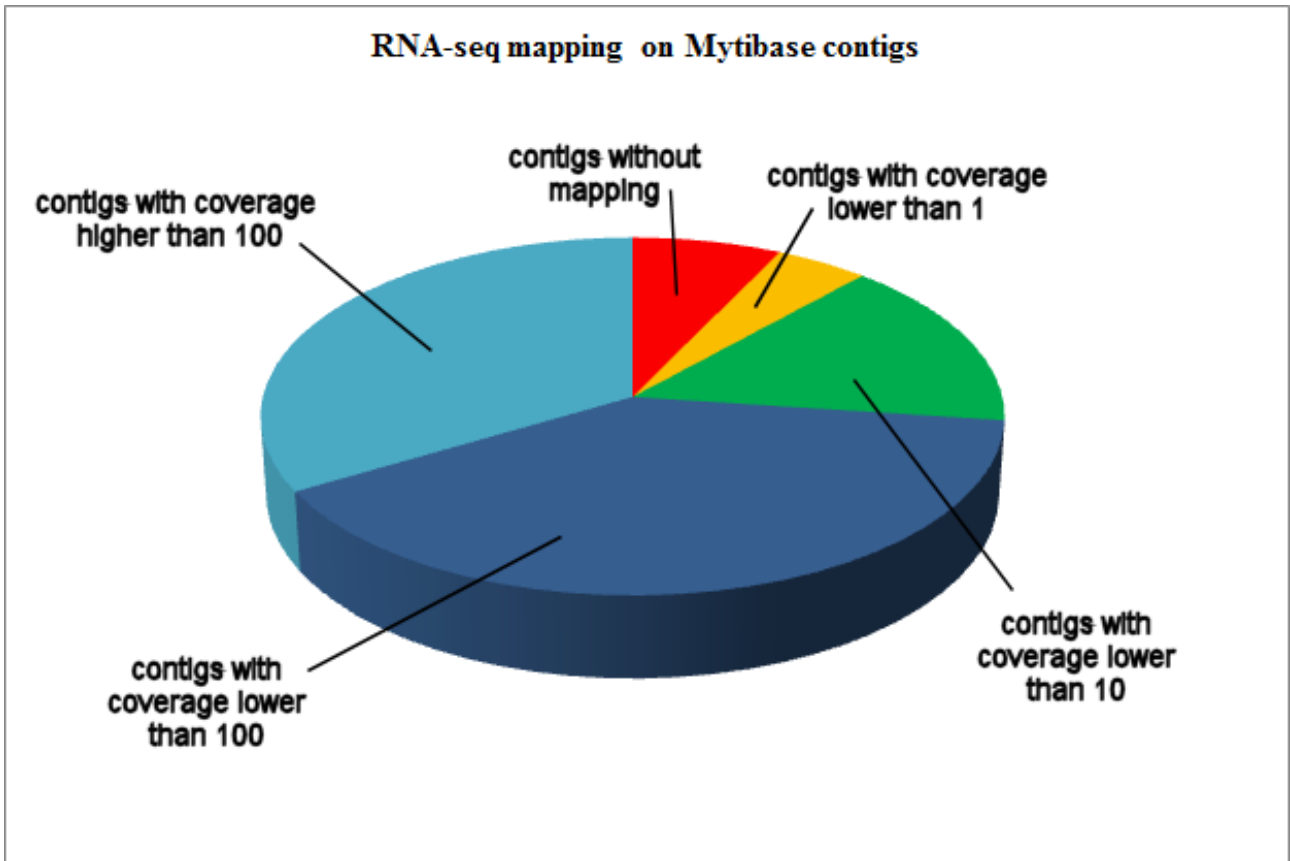


Fig. 5

Mapping of the digestive gland RNA-seq reads on the 7,112 Mytibase contigs. Most of the contigs originally present in Mytibase resulted to be highly covered by RNA-seq reads, whereas just a limited number of contigs was not expressed in the digestive gland

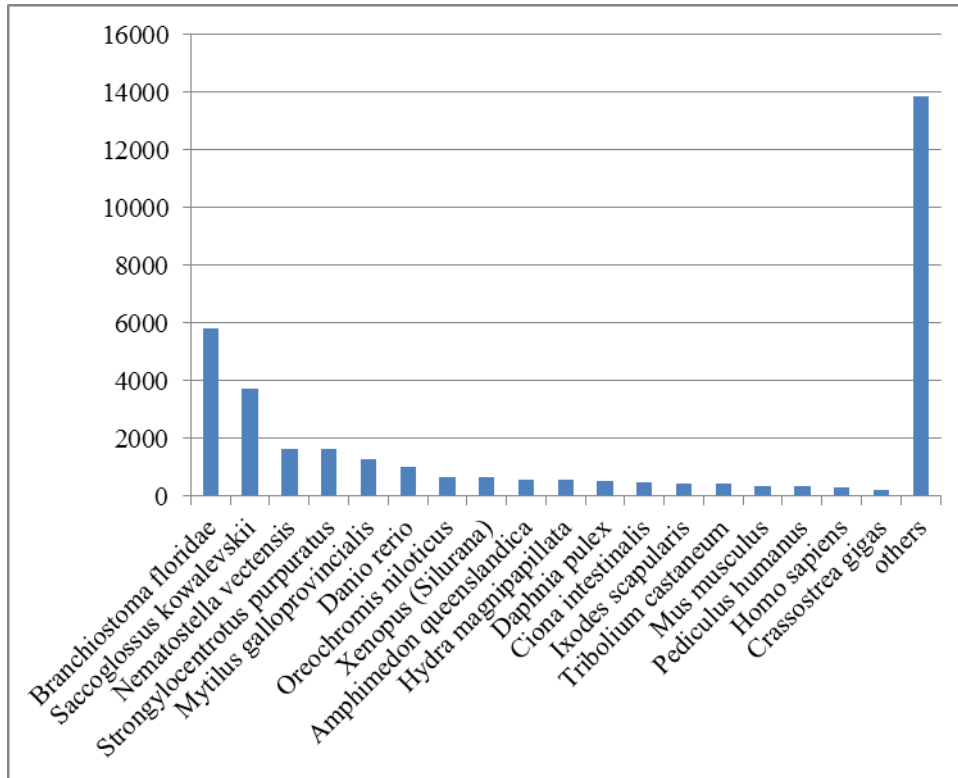


Figure S1: Top hit BLAST species in the NCBI non redundant protein database.

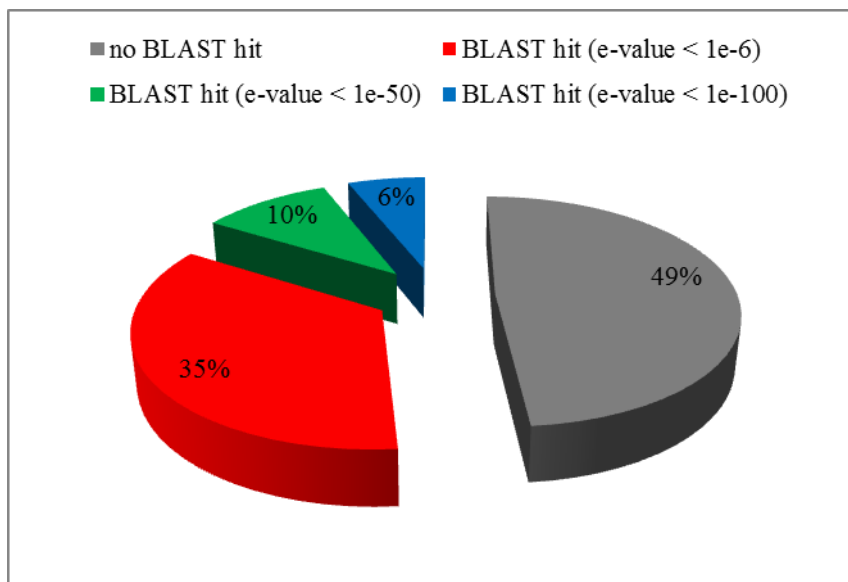


Figure S2: BLAST homologies of the filtered contigs of *M. galloprovincialis* versus the total proteins *L. gigantea* predicted from genome assembly.

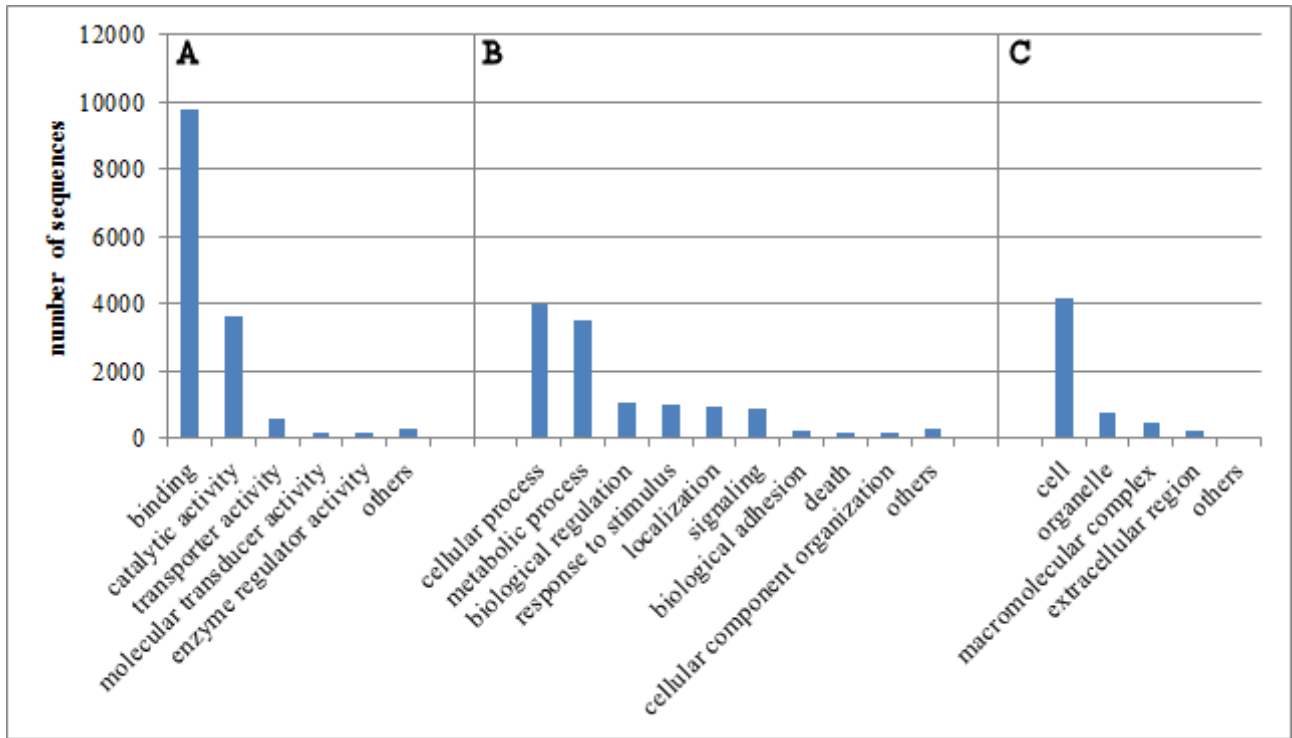


Figure S3: Gene Ontology (GO) terms assignments. A: molecular function. B: biological process. C: cellular component.

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 S4: Top 30 most represented Gene Ontology (GO) at the level 3 of Biological Process.

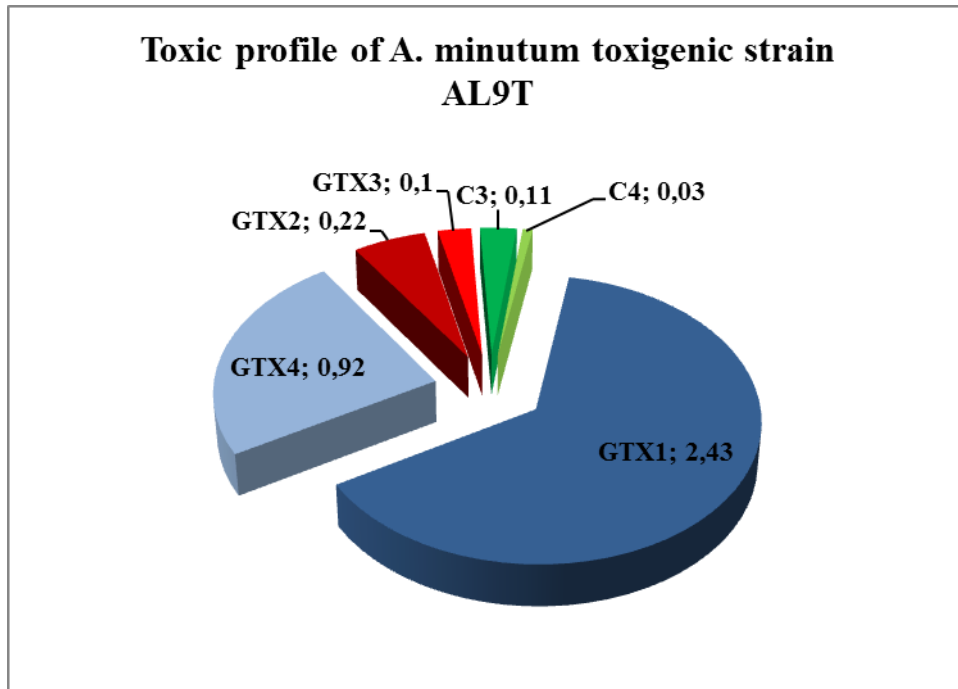


Figure S5: typical toxin profile of the *A. minutum* AL9T strain, as determined by Jaime et al. (personal communication).

References

- Agatha, S., Strüder-Kypke, M.C. & Beran, A. (2004). Morphologic and Genetic Variability in the Marine Planktonic Ciliate *Laboea strobila* Lohmann, 1908 (Ciliophora, Oligotrichia), with Notes on its Ontogenesis. *Journal of Eukaryotic Microbiology*, 51: 267-281.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25: 3389-3402.
- Álvarez, G., Uribe, E., Vidal, A., Ávalos, P., González, F., Mariño, C. & Blanco, J. (2009). Paralytic shellfish toxins in *Argopecten purpuratus* and *Semimytilus algosus* from northern Chile. *Aquatic Living Resources*, 22: 341-347.
- Amzil, Z., Quilliam, M.A., Hu, T. & Wright, J.L.C. (1999). 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*, 7: 271-277.
- Anders, S. & Huber, W. (2010a). Differential expression analysis for sequence count data. *Genome Biology*, 11.
- Anders, S. & Huber, W. (2010b). Differential expression analysis for sequence count data. *Genome Biology*, 11: R106.
- Anderson, D.M., Sullivan, J.J. & Reguera, B. (1989). Paralytic shellfish poisoning in northwest Spain: The toxicity of the dinoflagellate *Gymnodinium catenatum*. *Toxicon*, 27: 665-674.
- Ayres, P.A. (1975). Mussel poisoning in Britain with special reference to paralytic shellfish poisoning. A review of cases reported 1814-1968. *ENVIRONMHLTH*, 83: 261-265.
- Baggerly, K.A., Deng, L., Morris, J.S. & Aldaz, C.M. (2003). Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, 19: 1477-1483.
- Basti, L., Nagai, K., Shimasaki, Y., Oshima, Y., Honjo, T. & Segawa, S. (2009). Effects of the toxic dinoflagellate *Heterocapsa circularisquama* on the valve movement behaviour of the Manila clam *Ruditapes philippinarum*. *Aquaculture*, 291: 41-47.
- Bettencourt, R., Pinheiro, M., Egas, C., Gomes, P., Afonso, M., Shank, T. & Santos, R. (2010). High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*. *BMC Genomics*, 11: 559.
- Blanco, J., Morono, A., Franco, J. & Reyero, M.I. (1997). PSP detoxification kinetics in the mussel *Mytilus galloprovincialis*. One- and two-compartment models and the effect of some environmental variables. *Marine Ecology Progress Series*, 158: 165-175.
- Blanco, J., Reyero, M.I. & Franco, J. (2003). Kinetics of accumulation and transformation of paralytic shellfish toxins in the blue mussel *Mytilus galloprovincialis*. *Toxicon*, 42: 777-784.
- Bravo, I., Reyero, M.I., Cacho, E. & Franco, J.M. (1999). Paralytic shellfish poisoning in *Haliotis tuberculata* from the Galician coast: Geographical distribution, toxicity by lengths and parts of the mollusc. *Aquatic Toxicology*, 46: 79-85.

- Bricelj, V.M., Connell, L., Konoki, K., Macquarrie, S.P., Scheuer, T., Catterall, W.A. & Trainer, V.L. (2005). Sodium channel mutation leading to saxitoxin resistance in clams increases risk of PSP. *Nature*, 434: 763-767.
- Bricelj, V.M., Lee, J.H. & Cembella, A.D. (1991). Influence of dinoflagellate cell toxicity on uptake and loss of paralytic shellfish toxins in the northern quahog *Mercenaria mercenaria*. *Marine Ecology Progress Series*, 74: 33-46.
- Bricelj, V.M., Lee, J.H., Cembella, A.D. & Anderson, D.M. (1990). Uptake kinetics of paralytic shellfish toxins from the dinoflagellate *Alexandrium fundyense* in the mussel *Mytilus edulis*. *Mar Ecol Prog Ser*, 63: 117-188.
- Bricelj, V.M. & Shumway, S.E. (1998). Paralytic Shellfish Toxins in Bivalve Molluscs: Occurrence, Transfer Kinetics, and Biotransformation. *Reviews in Fisheries Science*, 6: 315-383.
- Cembella, A.D., Quilliam, M.A., Lewis, N.I., Bauder, A.G., Dell'aversano, C., Thomas, K., Jellett, J. & Cusack, R.R. (2002). The toxigenic marine dinoflagellate *Alexandrium tamarense* as the probable cause of mortality of caged salmon in Nova Scotia. *Harmful Algae*, 1: 313-325.
- Cembella, A.D., Sullivan, J.J., Boyer, G.L., Taylor, F.J.R. & Andersen, R.J. (1987). Variation in paralytic shellfish toxin composition within the *Protogonyaulax tamaronsis/catenella* species complex; red tide dinoflagellates. *Biochemical Systematics and Ecology*, 15: 171-186.
- Clark, M., Thorne, M., Vieira, F., Cardoso, J., Power, D. & Peck, L. (2010). Insights into shell deposition in the Antarctic bivalve *Laternula elliptica*: gene discovery in the mantle transcriptome using 454 pyrosequencing. *BMC Genomics*, 11: 362.
- Conesa, A. & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, 2008.
- Connell, L.B., Macquarrie, S.P., Twarog, B.M., Iszard, M. & Bricelj, V.M. (2007). Population differences in nerve resistance to paralytic shellfish toxins in softshell clam, *Mya arenaria*, associated with sodium channel mutations. *Marine Biology*, 150: 1227-1236.
- Conte, F.S. (1984). Economic impact of paralytic shellfish poison on the oyster industry in the Pacific United States. *Aquaculture*, 39: 331-343.
- Coulson, J.C., Potts, G.R., Deans, I.R. & Fraser, S.M. (1968). Dinoflagellate crop in the North Sea: Mortality of shags and other Sea Birds caused by paralytic shellfish poison. *Nature*, 220: 23-24.
- Craft, J.A., Gilbert, J.A., Temperton, B., Dempsey, K.E., Ashelford, K., Tiwari, B., Hutchinson, T.H. & Chipman, J.K. (2010). Pyrosequencing of *Mytilus galloprovincialis* cDNAs: Tissue-Specific Expression Patterns. *PLoS ONE*, 5: e8875.
- Delgado, M., Estrada, M., Camp, J., Fernández, J.V., Santmartí, M. & Lletí, C. (1990). Development of a toxic *Alexandrium minutum* Halim (Dinophyceae) bloom in the harbour of Sant Carles de la Ràpita (Ebro Delta, northwestern Mediterranean). *Scientia Marina*, 54: 1-7.
- Desbiens, M. & Cembella, A.D. (1993). Minimization of PSP toxin accumulation in cultured blue mussels (*Mytilus edulis*) by vertical displacement in the water column. IN Smayda, T.J. & Shimizu, Y.T. (Eds.) *Toxic phytoplankton blooms in the sea*. Amsterdam: Elsevier.
- Donovan, C.J., Ku, J.C., Quilliam, M.A. & Gill, T.A. (2008). Bacterial degradation of paralytic shellfish toxins. *Toxicon*, 52: 91-100.

Efsa (2009). 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, 1019: 1-76.

Feldmeyer, B., Wheat, C., Krezdorn, N., Rotter, B. & Pfenninger, M. (2011). 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, 12: 317.

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

Gacutan, R.Q., Tabbu, M.Y., Aujero, E.J. & Icatlo Jr, F. (1985). Paralytic shellfish poisoning due to *Pyrodinium bahamense* var. *compressa* in Mati, Davao oriental, Philippines. Marine Biology, 87: 223-227.

Gainey Jr, L.F. & Shumway, S.E. (1988). Physiological effects of *Protogonyaulax tamarensis* on cardiac activity in bivalve molluscs. Comparative Biochemistry and Physiology Part C, Comparative, 91: 159-164.

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

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

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

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

Gavery, M.R. & Roberts, S.B. Characterizing short read sequencing for gene discovery and RNA-Seq analysis in *Crassostrea gigas*. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics.

Gavery, M.R. & Roberts, S.B. (2012). Characterizing short read sequencing for gene discovery and RNA-Seq analysis in *Crassostrea gigas*. Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics, 7: 94-99.

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

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

Gerdol, M., Manfrin, C., De Moro, G., Figueras, A., Novoa, B., Venier, P. & Pallavicini, A. (2011). The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A

widespread and diverse family of immune-related molecules. *Developmental & Comparative Immunology*, 35: 635-643.

Gessner, B.D. & Middaugh, J.P. (1995). Paralytic shellfish poisoning in Alaska: A 20-year retrospective analysis. *American Journal of Epidemiology*, 141: 766-770.

Ghiselli, F., Milani, L., Chang, P.L., Hedgecock, D., Davis, J.P., Nuzhdin, S.V. & Passamonti, M. (2012). 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*, 29: 771-786.

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., Otilar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D. & Dubchak, I. (2012). The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*, 40: D26-D32.

Haberkorn, H., Tran, D., Massabuau, J.C., Ciret, P., Savar, V. & Soudant, P. (2011). 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*, 62: 1191-1197.

Hallegraeff, G.M., Steffensen, D.A. & Wetherbee, R. (1988). Three estuarine Australian dinoflagellates that can produce paralytic shellfish toxins. *J PLANKTON RES*, 10: 533.

Hansen, P.J., Cembella, A.D. & Moestrup, O. (1992). The marine dinoflagellate *Alexandrium ostenfeldii*: paralytic shellfish toxin concentration, composition, and toxicity to a tintinnid ciliate. *Journal of Phycology*, 28: 597-603.

Hashimoto, T., Matsuoka, S., Yoshimatsu, S.A., Miki, K., Nishibori, N., Nishio, S. & Noguchi, T. (2002). 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*, 43: 1-5.

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

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

Humpage, A.R., Rositano, J., Bretag, A.H., Brown, R., Baker, P.D., Nicholson, B.C. & Steffensen, D.A. (1994). Paralytic shellfish poisons from Australian cyanobacterial blooms. *AUSTJMARFRESHWATER RES*, 45: 761-771.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., Mcanulla, C., Mcdowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J.A., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. & Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37: D211-D215.

- James, K.J., Carey, B., O'halloran, J., Van Pelt, F.N.a.M. & Škrabáková, Z. (2010). Shellfish toxicity: Human health implications of marine algal toxins. *Epidemiology and Infection*, 138: 927-940.
- Jester, R.J., Baugh, K.A. & Lefebvre, K.A. (2009). Presence of *Alexandrium catenella* and paralytic shellfish toxins in finfish, shellfish and rock crabs in Monterey Bay, California, USA. *Marine Biology*, 156: 493-504.
- Joubert, C., Piquemal, D., Marie, B., Manchon, L., Pierrat, F., Zanella-Cleon, I., Cochenec-Laureau, N., Gueguen, Y. & Montagnani, C. (2010). Transcriptome and proteome analysis of *Pinctada margaritifera* calcifying mantle and shell: focus on biomineralization. *BMC Genomics*, 11: 613.
- Kitts, D.D., Smith, D.S., Beitler, M.K. & Liston, J. (1992). Presence of paralytic shellfish poisoning toxins and soluble proteins in toxic butter clams (*Saxidomus giganteus*). *Biochemical and Biophysical Research Communications*, 184: 511-517.
- Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R., Schander, C., Moroz, L.L., Lieb, B. & Halanych, K.M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, 477: 452-456.
- Kodama, M., Ogata, T., Fukuyo, Y., Ishimaru, T., Wisessang, S., Saitanu, K., Panichyakarn, V. & Piyakarnchana, T. (1988). *Protogonyaulax cohorticula*, a toxic dinoflagellate found in the Gulf of Thailand. *Toxicon*, 26: 707-712.
- Kvitek, R.G., Degange, A.R. & Beitler, M.K. (1991). Paralytic shellfish poisoning toxins mediate feeding behavior of sea otters. *Limnology & Oceanography*, 36: 393-404.
- Lawrence, J.F., Niedzwiadek, B. & Menard, C. (2004). Quantitative determination of paralytic shellfish poisoning toxins in shellfish using prechromatographic oxidation and liquid chromatography with fluorescence detection: interlaboratory study. *J AOAC Int*, 87: 83-100.
- Lilly, E.L., Kulis, D.M., Gentien, P. & Anderson, D.M. (2002). 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*, 24: 443-452.
- Livak, K.J. & Schmittgen, T.D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. *Methods*, 25: 402-408.
- Macquarrie, S.P. & Bricelj, V.M. (2008). Behavioral and physiological responses to PSP toxins in *Mya arenaria* populations in relation to previous exposure to red tides. *Marine Ecology Progress Series*, 366: 59-74.
- Maguer, J.F., Wafar, M., Madec, C., Morin, P. & Denn, E.E.L. (2004). Nitrogen and phosphorus requirements of an *Alexandrium minutum* bloom in the Penzé Estuary, France. *Limnology and Oceanography*, 49: 1108-1114.
- Manfrin, C., Dreos, R., Battistella, S., Beran, A., Gerdol, M., Varotto, L., Lanfranchi, G., Venier, P. & Pallavicini, A. (2010). Mediterranean mussel gene expression profile induced by okadaic acid exposure. *Environmental Science and Technology*, 44: 8276-8283.
- Marsden, I.D. & Shumway, S.E. (1993). 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*, 106: 769-773.

- Mcintyre, L., Lopiano, K., Morse, A., Amin, V., Oberg, A., Young, L. & Nuzhdin, S. (2011). RNA-seq: technical variability and sampling. *BMC Genomics*, 12: 293.
- Mee, L.D., Espinosa, M. & Diaz, G. (1986). Paralytic shellfish poisoning with a *Gymnodinium catenatum* red tide on the Pacific coast of Mexico. *Marine Environmental Research*, 19: 77-92.
- Milan, M., Coppe, A., Reinhardt, R., Cancela, L., Leite, R., Saavedra, C., Ciofi, C., Chelazzi, G., Patarnello, T., Bortoluzzi, S. & Bargelloni, L. (2011). Transcriptome sequencing and microarray development for the Manila clam, *Ruditapes philippinarum*: genomic tools for environmental monitoring. *BMC Genomics*, 12: 234.
- Montebruno, D. (1993). Paralytic shellfish poisoning in Chile. *Medicine, Science and the Law*, 33: 243-246.
- Nagai, K., Honjo, T., Go, J., Yamashita, H. & Seok Jin, O. (2006). Detecting the shellfish killer *Heterocapsa circularisquama* (Dinophyceae) by measuring bivalve valve activity with a Hall element sensor. *Aquaculture*, 255: 395-401.
- Narahashi, T., Haas, H.G. & Therrien, E.F. (1967). Saxitoxin and tetrodotoxin: comparison of nerve blocking mechanism. *Science*, 157: 1441-1442.
- Navarro, J. & Contreras, A. (2010). An integrative response by *Mytilus chilensis* to the toxic dinoflagellate *Alexandrium catenella*. *Marine Biology*, 157: 1967-1974.
- O'neil, S., Dzurisin, J., Carmichael, R., Lobo, N., Emrich, S. & Hellmann, J. (2010). Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics*, 11: 310.
- Okumura, M., Yamada, S., Oshima, Y. & Ishikawa, N. (1994). Characteristics of paralytic shellfish poisoning toxins derived from short-necked clams (*Tapes japonica*) in Mikawa Bay. *Natural Toxins*, 2: 141-143.
- Oshima, Y. (1995). Chemical and enzymatic transformation of paralytic shellfish toxins in marine organisms. IN Lassus, P., Arzul, G. & E., E. (Eds.) *Harmful Marine Algal Blooms*. Paris: Lavoisier.
- Oshlack, A., Robinson, M. & Young, M. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, 11: 220.
- Pérez-Enciso, M. & Ferretti, L. (2010). Massive parallel sequencing in animal genetics: Wherefroms and wheretos. *Animal genetics*, 41: 561-569.
- Philipp, E.E.R., Kraemer, L., Melzner, F., Poustka, A.J., Thieme, S., Findeisen, U., Schreiber, S. & Rosenstiel, P. (2012). Massively Parallel RNA Sequencing Identifies a Complex Immune Gene Repertoire in the lophotrochozoan *Mytilus edulis*. *PLoS ONE*, 7: e33091.
- Quilliam, M.A., Janeček, M. & Lawrence, J.F. (1993). Characterization of the oxidation products of paralytic shellfish poisoning toxins by liquid chromatography/mass spectrometry. *Rapid Communications in Mass Spectrometry*, 7: 482-487.
- 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.
- Robinson, M.D., Mccarthy, D.J. & Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26: 139-140.

- Sagou, R., Amanhir, R., Taleb, H., Vale, P., Blaghen, M. & Loutfi, M. (2005). Comparative study on differential accumulation of PSP toxins between cockle (*Acanthocardia tuberculatum*) and sweet clam (*Callista chione*). *Toxicon*, 46: 612-618.
- Schwinghamer, P., Hawryluk, M., Powell, C. & Mackenzie, C.H. (1994). Resuspended hypnozygotes of *Alexandrium fundyense* associated with winter occurrence of PSP in inshore Newfoundland waters. *Aquaculture*, 122: 171-179.
- Shumway, S.E. & Cucci, T.L. (1987). The effects of the toxic dinoflagellate *Protogonyaulax tamarensis* on the feeding and behaviour of bivalve molluscs. *Aquatic Toxicology*, 10: 9-27.
- Shumway, S.E., Pierce, F.C. & Knowlton, K. (1987). 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*, 87: 1021-1023.
- Shumway, S.E., Sherman, S.A., Cembella, A.D. & Selvin, R. (1994). 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*, 2: 236-251.
- Smith, E.A., Grant, F., Ferguson, C.M. & Gallacher, S. (2001). Biotransformations of paralytic shellfish toxins by bacteria isolated from bivalve molluscs. *Appl Environ Microbiol*, 67: 2345-53.
- Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G. & Dunn, C.W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 480: 364-367.
- Sullivan, J.J., Iwaoka, W.T. & Liston, J. (1983). Enzymatic transformation of PSP toxins in the littleneck clam (*Protothaca staminea*). *Biochem Biophys Res Commun*, 114: 465-72.
- Takatani, T., Morita, T., Anami, A., Akaeda, H., Kamijo, Y., Tsutsumi, K. & Noguchi, T. (1998). Appearance of *Gymnodinium catenatum* in association with the toxification of bivalves in Kamae, Oita Prefecture, Japan. *Journal of the Food Hygienic Society of Japan*, 39: 275-280.
- Takati, N., Mountassif, D., Taleb, H., Lee, K. & Blaghen, M. (2007). Purification and partial characterization of paralytic shellfish poison-binding protein from *Acanthocardia tuberculatum*. *Toxicon*, 50: 311-321.
- Terlau, H., Heinemann, S.H., Stuhmer, W., Pusch, M., Conti, F., Imoto, K. & Numa, S. (1991). Mapping the site of block by tetrodotoxin and saxitoxin of sodium channel II. *FEBS Letters*, 293: 93-96.
- Tian, H., Gao, C., Wang, Z., Sun, P., Fan, S. & Zhu, M. (2010). Comparative study on in vitro transformation of paralytic shellfish poisoning (PSP) toxins in different shellfish tissues. *Acta Oceanologica Sinica*, 29: 120-126.
- Toulza, E., Shin, M.-S., Blanc, G., Audic, S., Laabir, M., Collos, Y., Claverie, J.-M. & Grzebyk, D. (2010). Gene Expression in Proliferating Cells of the Dinoflagellate *Alexandrium catenella* (Dinophyceae). *Appl Environ Microbiol*, 76: 4521-4529.
- Tran, D., Haberkorn, H., Soudant, P., Ciret, P. & Massabuau, J.C. (2010). Behavioral responses of *Crassostrea gigas* exposed to the harmful algae *Alexandrium minutum*. *Aquaculture*, 298: 338-345.

- Twarog, B.M. (1974). Immunity to paralytic shellfish toxin in bivalve molluscs. IN Cameron, A.M., Campbell, B.M. & Cribb, A.B. (Eds.) Proc Int Symp Coral Reefs, 2nd. Brisbane, Australia: Great Barrier Reef Comm.
- Twarog, B.M., Hidaka, T. & Yamaguchi, H. (1972). Resistance to tetrodotoxin and saxitoxin in nerves of bivalve molluscs. A possible correlation with paralytic shellfish poisoning. *Toxicon*, 10: 273-278.
- 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: 285-291.
- Van Lenning, K., Vila, M., Masó, M., Garcés, E., Anglès, S., Sampedro, N., Morales-Blake, A. & Camp, J. (2007). Short-term variations in development of a recurrent toxic *Alexandrium minutum*-dominated dinoflagellate bloom induced by meteorological conditions. *Journal of Phycology*, 43: 892-907.
- 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.
- 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.
- Yao, J. & Yu, F. (2011). DEB: A web interface for RNA-seq digital gene expression analysis. *Bioinformatics*, 7: 44-45.



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

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.

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 (Bohlsón 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.

Materials and methods

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.

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 15°C 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 37°C and *V. anguillarum* in TSA supplemented with NaCl 1% at 20°C. 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.

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- $\Delta\Delta$ Ct 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 $P < 0.01$.

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 aegypti* (<http://aaegypti.vectorbase.org/>), *Culex quinquefasciatus* (<http://cquinquefasciatus.vectorbase.org/>) and *Schistosoma mansoni* (<http://www.genedb.org/Homepage/Smansoni>).

Results

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.

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 (MgC1q7 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.

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.

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. aegypti*, *C. quinquefasciatus*, *D. pulex*, *S. mansoni* and *C. elegans* in Table 2).

Discussion

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 hypothesized 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, Rotifera 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 virginica*, 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 aegypti* (<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.

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).

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 expressed 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.

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 Protostome. 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	taaagccggactgtacttgggtgc
MGC1Q5_REV	atctccctctgctgcctgta
MGC1Q6_FOR	ctggtgctgttttcggttgtag
MGC1Q6_REV	ttttcgatttcggtggat
MGC1Q7_FOR	aggtggcgtttatgctgcgttga
MGC1Q7_REV	ggagcagtaaacatgccatttaca
MGC1Q8_FOR	ccaattcgcagtgagttttgt
MgC1q8_REV	gtgtggcctgtaaagatcctgctg
EF-1_FOR	cctcccaccatcaagaccta
EF-1_REV	ggctggagcaaagtaacaa

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

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

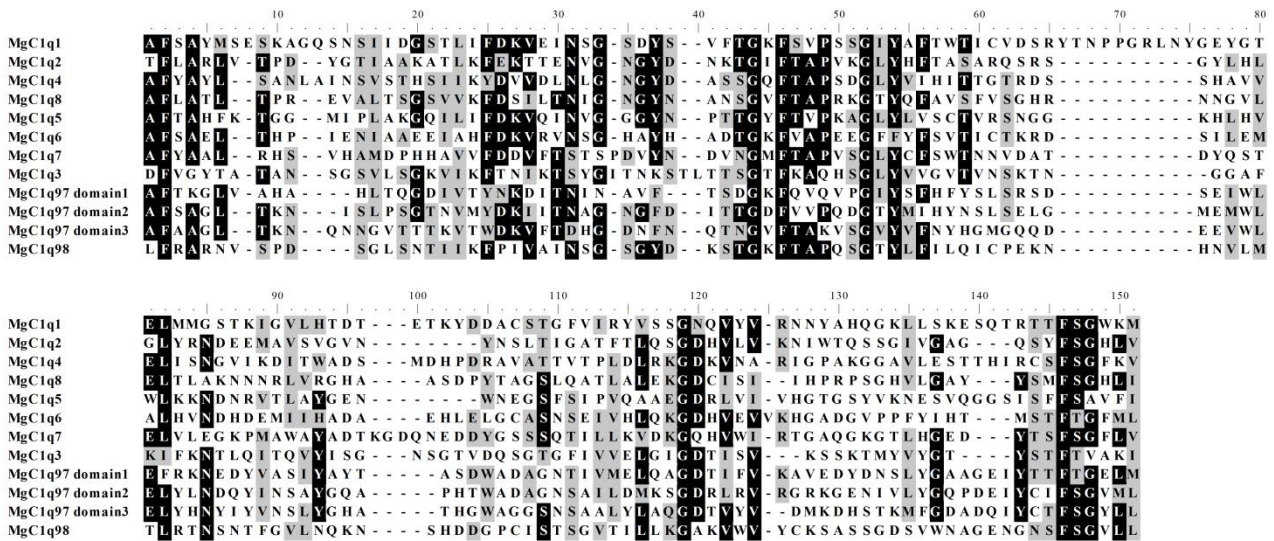


Figure 1: Multiple alignment of c1q domains from the 10 selected *Mytilus galloprovincialis* C1qDC proteins MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7, MgC1q8, MgC1q97 and MgC1q98; all the 3 different MgC1q97 C1q domains are represented in the alignment.

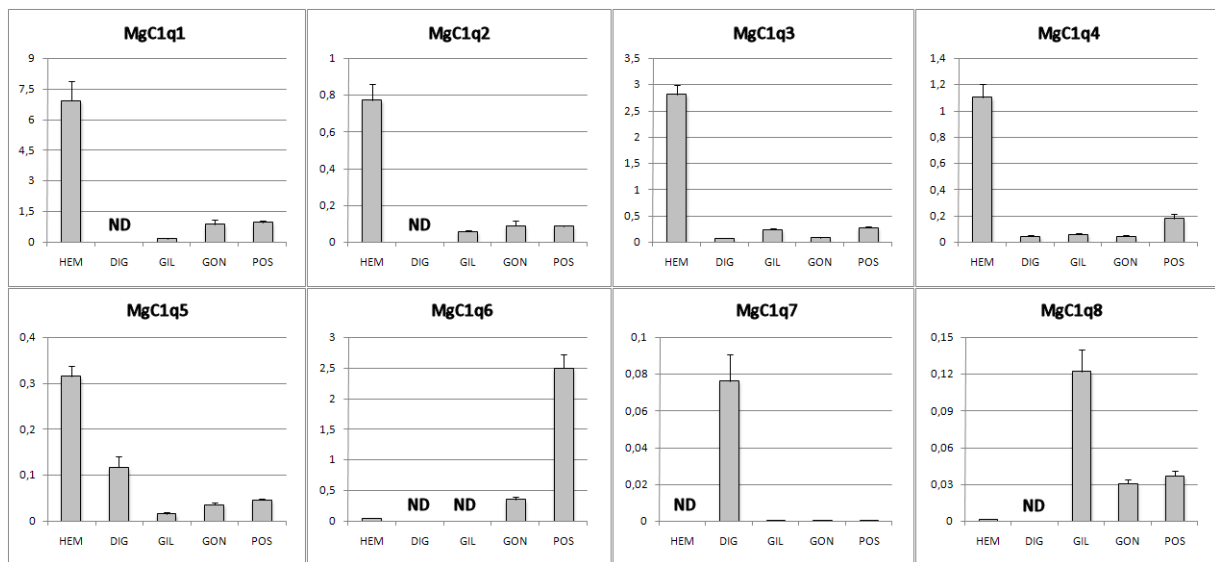


Figure 2: Tissue-specific expression of the mussel C1qDC transcripts MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7 and MgC1q8. Bars depict the transcript expression relative to the elongation factor EF-1. Results are mean \pm SD of 3 technical replicates. Y axis of each graph is scaled based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

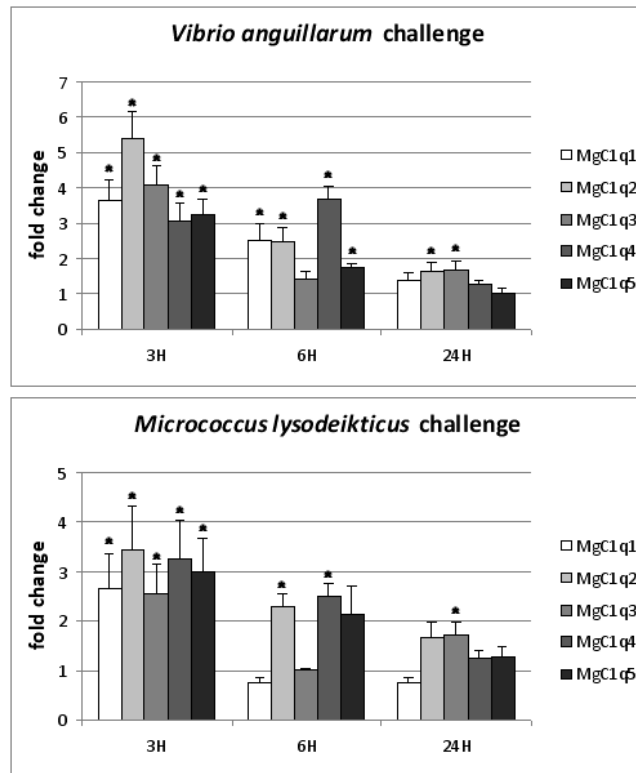


Figure 3: Expression changes of the transcripts MgC1q1, MgC1q2, MgC1q3, MgC1q4 and MgC1q5 in hemocytes sampled at 3, 6 and 24 hours post-injection from mussels challenged with Gram- (*V. anguillarum*, black bars) and Gram+ (*M. lysodeikticus*, white bars) bacteria; error bars represent fold change \pm 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 ($P < 0.01$).

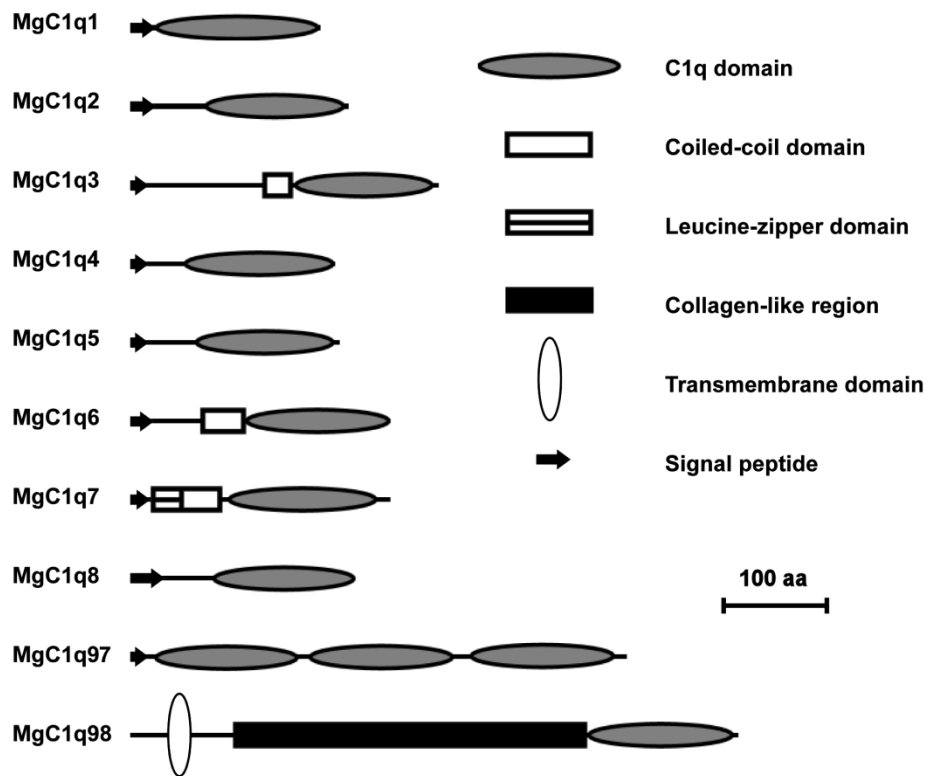


Figure 4: Structural organization of the selected *M. galloprovincialis* C1qDC proteins MgC1q1, MgC1q2, MgC1q3, MgC1q4, MgC1q5, MgC1q6, MgC1q7, MgC1q8, MgC1q97 and MgC1q98.

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/ves1-1L) 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.



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,*}

^aLaboratorio di Genetica, Department of Life Sciences, University of Trieste, Via Licio Giorgeri 5, 34126 Trieste, Italy

^bDepartment of Biology, CRIBI Biotechnology Center, University of Padova, Padova, Italy

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.

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 α -helix and β -sheet ($CS\alpha\beta$) 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.

Materials and methods

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.

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.

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.

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.

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.

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 μ L reaction mix included 7.5 μ L of 2X IQTM SYBR Green[®] Supermix (Biorad), 0.3 μ L of each 10 μ M primer and 2 μ L 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- $\Delta\Delta$ Ct 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.

Results and discussion

Big defensins

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, comprises 110259 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 length (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 cysteines 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 positive net charge at neutral pH whereas the isoelectric 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*.

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).

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 Lophotrochozoans. 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, horseshoe crabs and lancelets could be merely the result of either a 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).

Mytimacins

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 hydramacin-1, the 5 *M. galloprovincialis* mytimacins are likely to act as AMPs.

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. Mytimacin-3 was almost exclusively detected in the mantle, 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.

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 orthologues 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 acid 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 orthologous 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 Asterozoa), 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 organisms 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); c) long macins with 5 disulfide bridges (10-Cys macins); d) 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 positions (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	ggctggagcaaagtaacaa
Mytimacin-1 FOR	ctcctgcaaattcccacatc
Mytimacin-1 REV	atcttttgtccgccagaga
Mytimacin-2 FOR	gtggtggtggaagtggaagt
Mytimacin-2 REV	tccaagctcttgcacatctg
Mytimacin-3 FOR	acaatcaccaatgggaccac
Mytimacin-3 REV	ttgggcagcaaattctctc
MgBD1 FOR	gcgtagattccatcgagca
MgBD1 REV	tgttgatactcctgctcag
MgBD3 FOR	ccgattctaggacgagttgtggca
MgBD3 REV	ggcaactttccaagcgccatagc
MgBD6 FOR	agcatcatagcaggattgtc
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

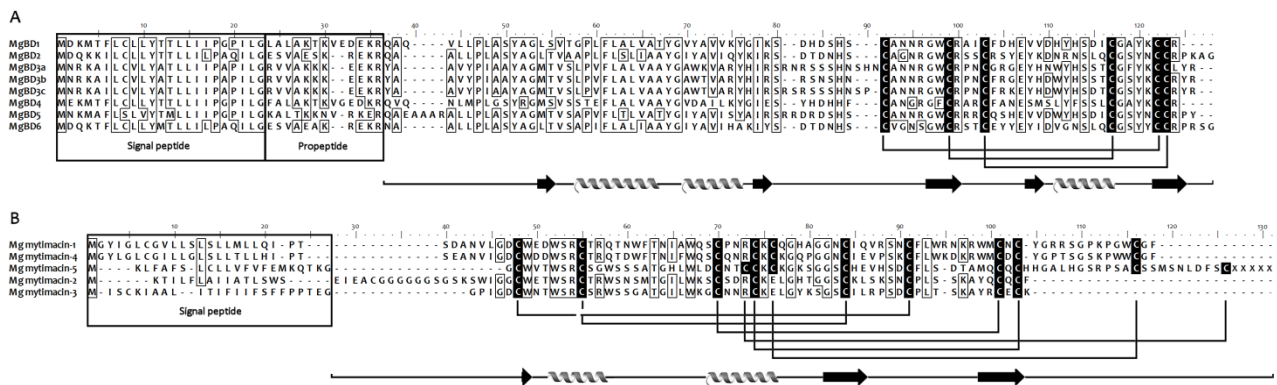


Figure 1: Panel A: multiple alignment of deduced amino acid sequences of the 8 mussel big defensins. Panel B: multiple alignment of deduced amino acid sequences from the five mytimacins. Conserved residues are outlined, cysteine residues engaged in disulfide bridges are *black boxed* and the organization 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).

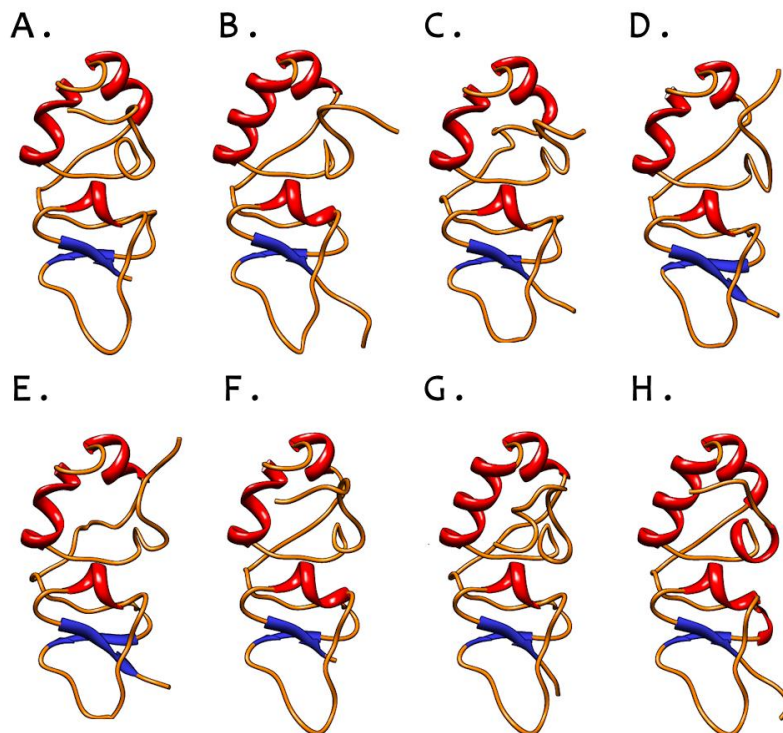


Figure 2: Predicted ribbon structures of mussel big defensins, obtained by Phyre2 modeling. A: MgBD1; B: MgBD2; C: MgBD3a; D: MgBD3b; E: MgBD3c; F: MgBD4; G: MgBD5; H: MgBD6.

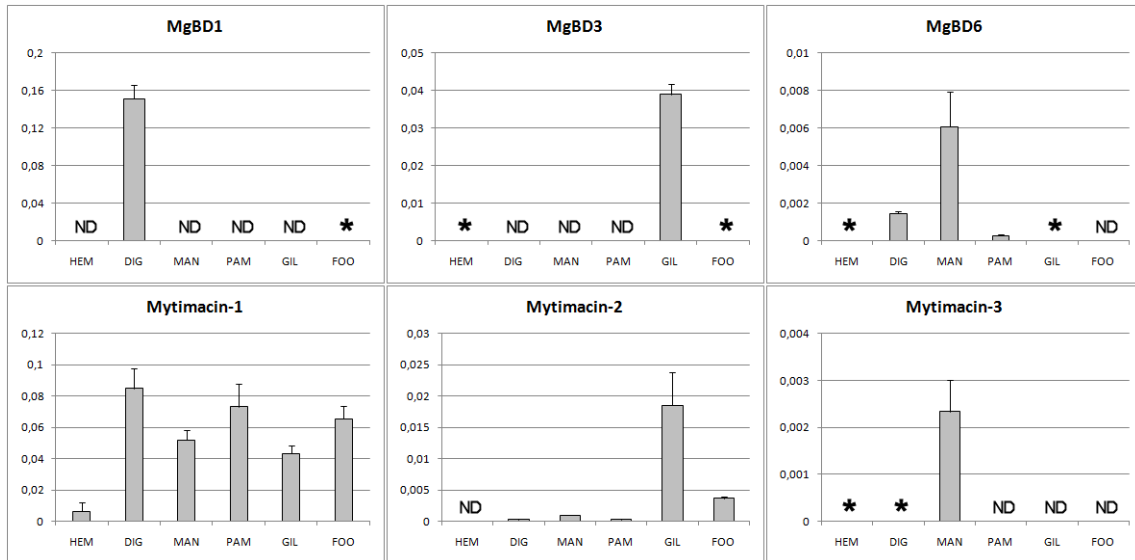


Figure 3: Tissue-specific expression of the transcripts mytimacin-1, mytimacin-2, mytimacin-3, BD1, BD3 and BD6. The expression values (bars) are relative to the elongation factor EF-1. Results are mean \pm SD of 3 technical replicates. The Y axis of each graph is scaled based on the highest level of 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).

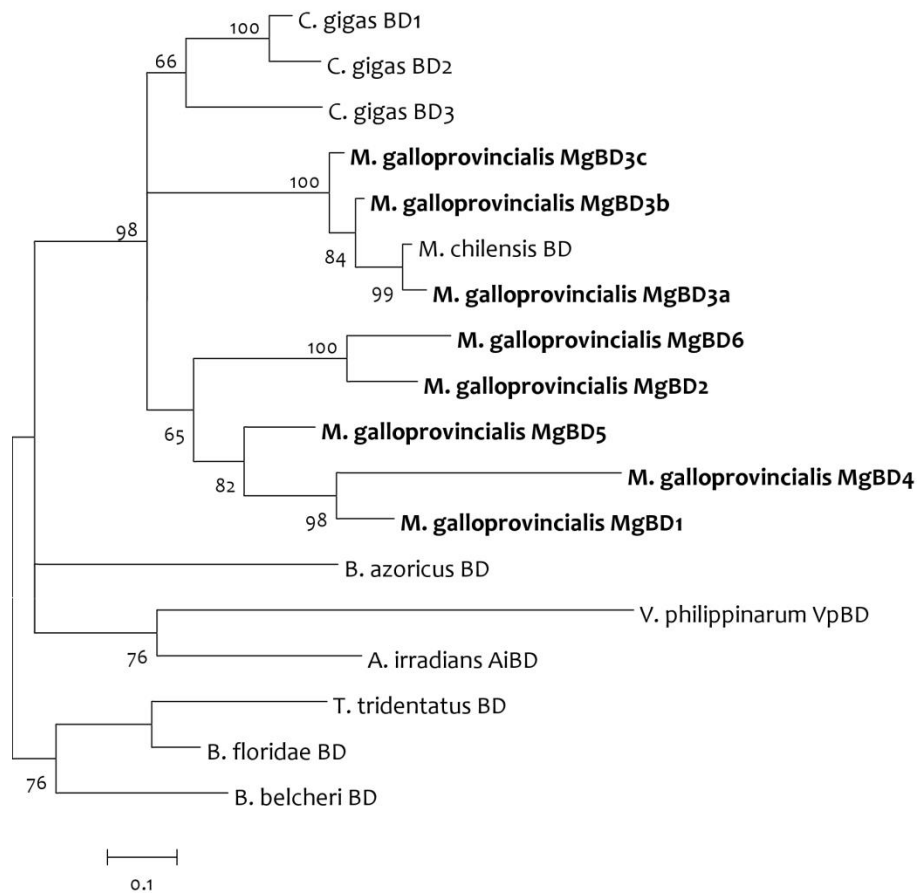


Figure 4: Bayesian phylogeny of big defensins inferred from the alignment of the predicted mature peptides. Posterior probabilities are shown for each branch. Entry IDs: *T. tridentatus* BD: P80957.2; *B. belcheri* BD: Q86QN6.1; *B. floridae* BD: ADH03419.1; *A. irradians* AiBD: Q0H293.1; *V. philippinarum* 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.

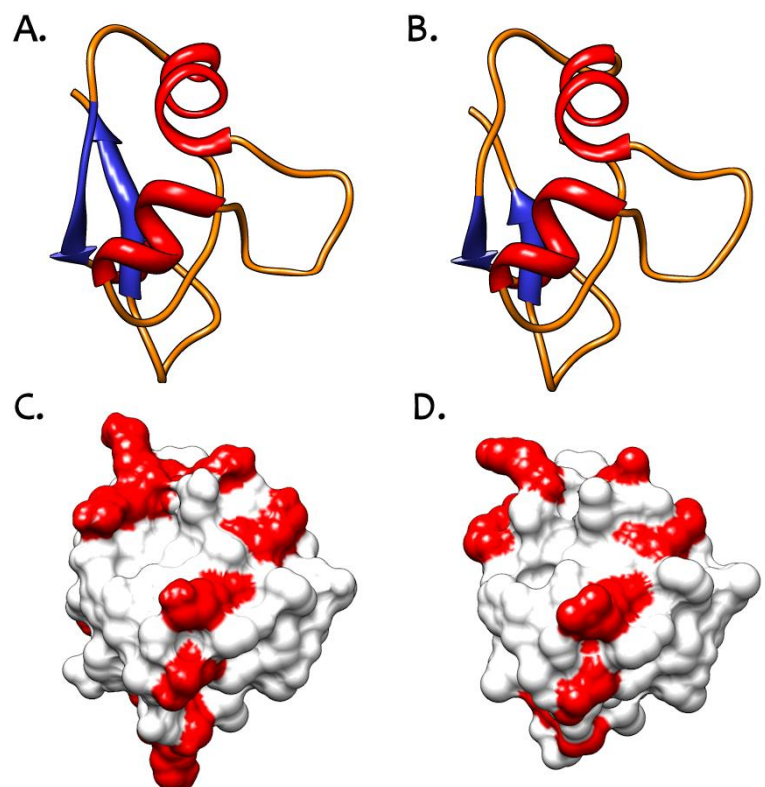


Figure 6: Ribbon structures and molecular surfaces of hydramacin-1 (panels A and C, PDB accession: 2K35) and mytimacin-3 (panels B and D, obtained by Phyre 2 modeling). Positive charges of lysine and arginine are colored, highlighting the high conservation of the positively charged residues distribution.

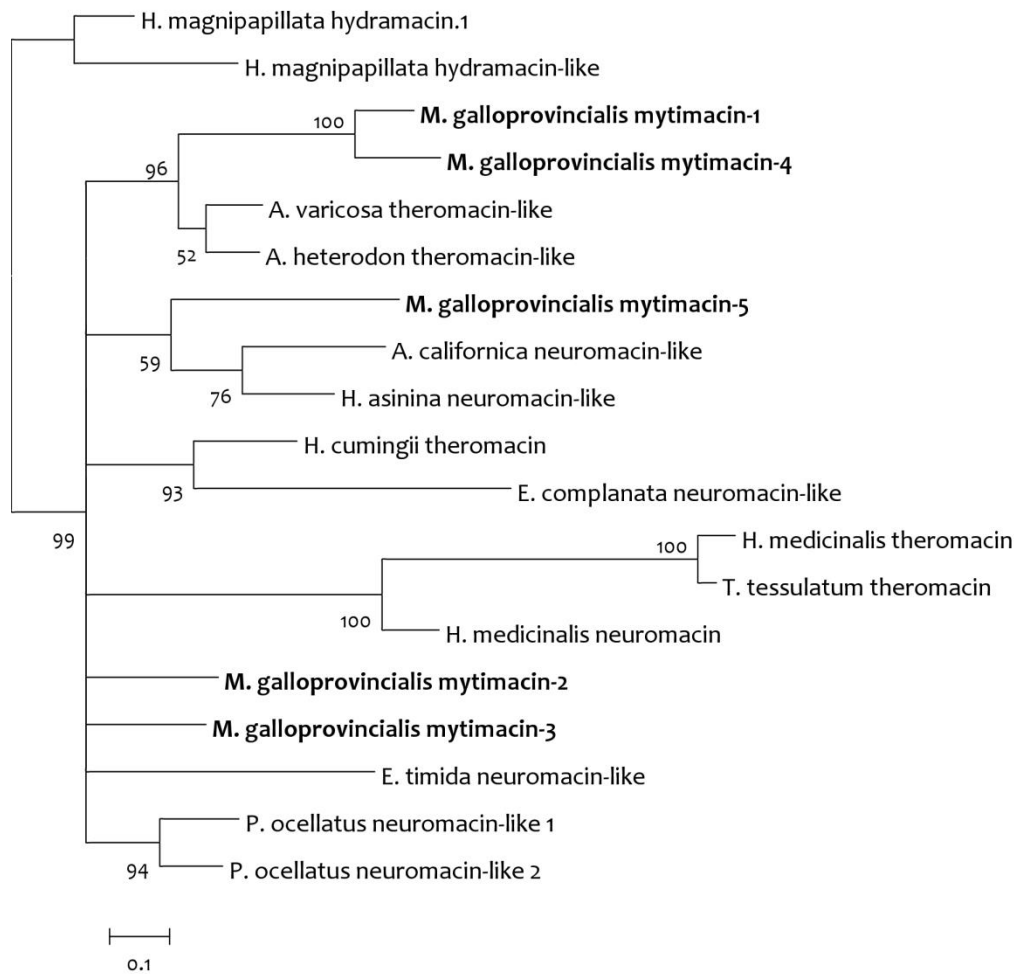


Figure 7: Bayesian phylogeny of macins inferred from the alignment of the predicted mature peptides. Posterior probabilities are shown for each branch.

(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		■		
	Bilateria	Protostomia	Ecdysozoa	Nematoda	■			
				Tardigrada				
			Lophotrochozoa	Arthropoda	■			
				<i>Mytilus galloprovincialis</i>	■	■	■	■
				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., Gutschmann, 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 $2^{-\Delta\Delta CT}$ Method. *Methods* 25, 402-408.
- Mitta, G., Hubert, F., Dyrzynda, 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., Salzert, 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., Salzert, 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., Salzert, 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.

REVIEW

How gene expression profiles disclose vital processes and immune responses in *Mytilus* spp.**S Domeneghetti¹, C Manfrin², L Varotto¹, U Rosani¹, M Gerdol², G De Moro², A Pallavicini², P Venier¹**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

Abstract

Gene expression studies largely support the understanding of gene-environment interactions in humans and other living organisms but the lack of genomic and genetic information often complicates the analysis of functional responses in non-traditional model species. Nevertheless, the fast advancement of DNA microarray and sequencing technologies now makes global gene expression analysis possible in virtually any species of interest. As regards the *Mytilus* genus, tens of thousands Expressed Sequence Tags (ESTs) are currently available for *M. californianus* and *M. galloprovincialis*, and DNA microarrays have been developed. Among them, Immuchip 1.0 specifically includes 1,820 probes of genes centrally involved or modulated in the innate immune responses of the Mediterranean mussel. This review recalls peculiarities and applications of the existing mussel DNA microarrays and finally summarizes facts concerning a variety of transcript sequences likely involved in the mussel immunity. Beside DNA microarrays, Next Generation Sequencing (NGS) technologies now offer new and broader research perspectives, from the whole transcriptome coverage to the *Mytilus* genome sequencing.

Key Words: *Mytilus*; DNA microarray; innate immunity; ESTs; antimicrobial peptides; C1q

Introduction

Global gene expression analyses in organisms selected to represent a given ecosystem currently support ecotoxicological investigations and create a conceptual bridge between the early organism responses and late population changes (Steinberg et al., 2008). The animal response to a variety of detrimental conditions usually starts with alarm signals followed by adjustment reactions aimed to neutralize the physiological unbalance, and may end up in a general decline of vital processes ultimately marked by disease and death. Depending on the stress type and exposure intensity, the expression of definite sets of genes makes available specific proteins and other molecules in cells and tissues.

Appeared in the 1990s, the DNA microarray technology enables the simultaneous expression measure of thousands of genes represented in the microarray platform by unambiguous polynucleotide probes (Schena et al., 1995; Lockhart et al., 1996). The gene expression profiles emerging from suitable sampled cells or tissues can provide a dynamic view of biological processes and allow the correct sorting of different functional states. Based on the availability of sequence data, DNA microarrays can be used to solve a variety of biological questions: from the identification of molecular markers pathognomonic of disease and transcriptional signatures of various stress factors to the understanding of complex phenomena such as the epigenome in normality and disease (Martín-Subero and Esteller, 2011).

Specific microarray platforms and advanced deep sequencing technologies now support studies on the cellular functions of microRNAs and their role in human diseases (Thomas et al., 2010). Leading research institutions are currently using both the mRNA and miRNA expression profiling to examine the genomic responses to environmental stresses (NCT). Central to the toxicogenomics studies is the concept of ‘phenotypic anchoring’ which recalls the importance to correlate the observed gene expression changes to adverse effects defined by conventional parameters of toxicity and pathology.

In the controlled vocabulary of the Natl. Library of Medicine, the term ‘DNA microarray’ is indexed under the following category which indicates the large application range of such innovative technology (MESH): Oligonucleotide Array Sequence Analysis- the hybridization of a nucleic acid sample to a very large set of oligonucleotide probes, which are attached to a solid support, to determine sequence or to detect variations in a gene sequence or expression or for gene mapping.

Relevant to the gene expression profiling research area is Gene Expression Omnibus, a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the scientific community (GEO). To fulfil the current standards (Minimum Information About a Microarray Experiment) the contents

submitted to GEO should include the following: raw hybridization data; normalized data from which the main experimental findings can be outlined; description of the tested samples and whole experimental design, with details on the biological and technical replicates; identity and location of all probes and controls of the microarray platform, with external reference in the case of commercial all arrays; concise but precise description of laboratory and data processing protocols related to the experiment under submission. According to the aims of the Microarray Gene Expression Data Society, dating back to the late '90s, the compliance to the MIAME standards should assure the data comparability among different platforms and testing protocols while supporting common work criteria and the reduction of random data variation (Rogers and Cambrosio, 2007).

Based on the comparative data analysis, the guidelines for standardization and reporting have been further refined (Chen et al., 2007; Shi et al., 2008). At present, GEO contains as much as 9,000 platform records which can be accessed and browsed in full detail.

Figure 1 illustrates the annual increase of PubMed records including the term "DNA-microarray" or "Mytilus" (subject heading or title/abstract) and suggests that pioneering technologies open the way to new ideas more than an unconventional model organisms. In fact, the gene expression profiling field has substantially diversified: specialized equipments and various related software make today the DNA microarrays powerful tools for the study of gene sequence, structure and expression, particularly for the best known model organisms. Nonetheless, one must remember that transcription is just one step in gene expression, and post transcriptional events referred to maturation of the primary transcript, RNA editing and RNA silencing as well as various modifications of the translation products overall influence the final amounts and activity of cellular proteins.

Mytilus DNA microarrays: preparation strategy and applications

Six GEO records refer to mussel DNA microarrays at July 2011.

MytArray 1.0 (GEO platform GPL1799, Oct 2006) is composed by 1,712 cDNA probes, univocally tagging the 3'-end region of transcripts from the main tissues of adult mussels (*Mytilus galloprovincialis*) and 46 unrelated cDNA control probes, all printed in duplicate and twice per slide (1.7 k mussel probes per array, 7.0 k total probes per slide). The probes were designed in the 3'-UTR, one among the least conserved gene regions, so that competition of different mRNAs from genes with similar coding sequence and cross-hybridization to the same microarray probe should be minimal.

Also, the probe size of 400 - 800 bp is expected to ensure comparable efficiency in the amplification and spotting of the cDNA inserts as well as uniform hybridization kinetics (Venier et al., 2006).

MytArray 1.0 was first used to investigate the specificity of gene transcription in mussel tissues with different functional role and the transcriptional profiles of mussels treated with chemical mixtures or living wild in different sites of the Venice lagoon (Venier et al., 2006; GEO series GSE2176, GSE2183 and GSE2184). Sample pairs combined according to dye-swap labelling (reference and test samples labelled with Cy3/Cy5 cyanine dyes in alternate combinations) were competitively hybridized on the two equal arrays of cDNAs spotted on the same slide (Fig. 2). Gills, digestive gland, tissues involved in contraction/motility (foot, adductor muscles, ligaments) and reproduction (gonads and mantle) displayed specific transcriptional footprints, as expected. The results obtained in mussels treated with mixtures of inorganic metal salts or persistent organic chemicals guided the interpretation of the gene expression profiles of mussels living in the inner industrial canals or at the lagoon border open to the sea (this exercise yielded a provisional list of contamination marker probes). In this study, the evident transcriptional down-regulation detected in the reproductive tissues was consistent with the depleted status of the mussel gonads whereas the greatest variety and abundance of transcripts was found in the digestive gland. Additional analysis of these expression data is reported elsewhere (Pantartzzi et al., 2010).

The same platform was then used to evaluate in a time-course study the gene expression changes in the digestive gland of mussels exposed to okadaic acid (OA) via food contamination for five weeks (Manfrin et al., 2010; GEO series GSE14885). One relevant purpose of the study was the identification of molecular biomarkers which could enable an easy and rapid detection of the Diarrhoeic Shellfish Poisoning biotoxins in marketable mussel stocks, i.e., novel reliable assays complementing the existing diagnostic methods. An unsaturated loop design, combining control and treated samples with different dye-labelling for the competitive hybridization on Mytarray 1.0, was adopted to take into account all the time points and the biological replicates, with some combinations only inferred (Kerr and Churchill, 2001). A considerable number of transcriptional changes was detected in the OA-exposed mussels, with a prevalence of up-regulated probes at 3 days and a subsequent progressive increase of down-regulated probes (from 58 % over-expressed to 76 % under-expressed genes, respectively detected at day 3 and day 35). The biphasic time-related trend of response observed in this study recalls the changes occurring in the mussel digestive gland along different phases of the mussel reaction to the experimental stimulus, from the early acute response to the late overall unbalance of the functional processes. Many candidate markers are now under study to evaluate their predictive value in the diagnosis of biotoxin-contaminated mussels.

MytArray 1.1 (GPL102699, March 2010) contains the same cDNA probes of MytArray 1.0 in a slightly modified platform geometry. It has been used to study the gene expression profiles of *M. galloprovincialis* with monthly samplings for one year, hence taking into account seasonal differences which are known to influence metabolism rates and gonad development among other vital functions (Banni et al., 2011; GEO series GSE22915, GSE23049- GSE23051). Mussels were collected from an anthropized and industrialized lagoon of the Southern Mediterranean Sea (Ben Said et al., 2009) and competitive hybridizations were performed with dye-swap-labelled samples (dual colour analysis).

Following a loop design with 3-4 biological replicates and parallel histological evaluation of the gonad status, the authors could analyze the transcriptional profiles of digestive gland tissue of female mussels collected during 12 months, and those of digestive gland and mantle tissues from male and female individuals representing all four gonad maturation stages. In the examined annual period, the transcriptional profiles globally highlighted the higher expression of genes associated to mussel nutrition and digestion in May- August compared to the other months, and trends for gonad transcripts consistent with the reproductive mussel status.

The same cDNA platform contributed to the toxicological evaluation of a neonicotinoid insecticide mixture (Dondero et al., 2010), an organophosphate compound (Canesi et al., 2011) and to the integrated measure of the functional mussel responses in the estuarine Tamar region in UK (Shaw et al., 2011).

The Hofmann_UCSB_Mytilus_2.5K_v1.0 record (GPL5795, Mar 2008) describes a platform of nearly 2500 spotted cDNAs of *Mytilus californianus* consisting of both unsequenced and sequenced clones referring to gill and muscle of environmentally challenged mussels. The related GEO series GSE8935 include data on latitudinal gene expression changes. Five biological replicates from four populations of Californian mussels were compared to a common reference sample in dual colour analysis (dye-swap labelling).

The HMS/SomeroLab-Mytilus-105K array-v1.0 (GPL9676, Jun 2010) and HMS/Somero-Mytilus-105K Agilent-v1.0 salinity stress (GPL11156, Jan 2011) are two successive versions of a platform composed by oligomer probes in-situ synthesized by Agilent Technologies (Santa Clara, CA, USA). These microarrays include probes of both *M. californianus* and *M. galloprovincialis*, and are intended for homologous and heterologous gene expression profiling. The processing and assembling of about 26,000 ESTs from *M. californianus* (Gracey et al., 2008) and 3,984 ESTs from *M. galloprovincialis* (Venier et al., 2003) resulted in a total of 12,961 and 1,688 transcript clusters or singletons,

respectively. Long (60-mer) oligoprobes were designed against the *M. californianus* series and the resulting 43,969 total unique probes (2.6 probes per transcript sequence) were analyzed through BLAST searches against the *M. galloprovincialis* series to support selection and design of related probes (556 probe pairs matching transcripts of both species, with a mean number of 4.6 divergent nucleotide bases per probe). A total of 44,524 unique probes were duplicated or triplicated randomly to fill a microarray of 105,000 elements (105 k probes).

These two platforms have been used to investigate the transcriptional responses to thermal and osmotic stresses in *M. californianus*, *M. trossulus* and *M. galloprovincialis* (Evans and Somero, 2010; Lockwood et al., 2010; Lockwood and Somero, 2011). To control the effects of sequence mismatches in the case of *M. galloprovincialis* probes included in the GPL9676 platform, only probes experimentally confirmed in the hybridization of 84 samples of both *M. galloprovincialis* and *M. trossulus* were used in the related data analysis. Following a large set of hybridization experiments and stringent quality control, misleading probes were removed from the dataset and the second platform version (GPL11156/Agilent 019153) was generated.

In the central and southern coasts of California, *M. galloprovincialis* has largely displaced the native congener, *M. trossulus*, and such evidence could be explained by species differences in physiological traits related to the adaptation to warm habitats. To investigate the hypothesis, gene expression profiling was performed on gill RNA from mussels subjected to acute heat-stress (GEO series GSE19031). A total of 1,531 probes, out of 4,488 different genes represented on the microarray and recognizing mRNAs of both species, showed temperature- dependent expression changes highly similar in the two congeners whereas 96 probes denoting oxidative stress, proteolysis, energy metabolism, ion transport, cell signalling, and cytoskeleton reorganization outlined species-specific responses to the heat-stress. Among them, the one encoding the small heat shock protein 24 was highly induced in the Mediterranean mussel and showed only a small change in *M. trossulus*. Six biological replicates per mussel group were included in this study which exemplifies the use of a cross-species microarray as well as heterologous and homologous hybridization. According to the authors and published literature, *M. trossulus* and *M. galloprovincialis* are approximately 7.6 million years divergent from *M. californianus*, and only 3.5 million years divergent from each other: in other words, the heterologous hybridization of target sequences from *M. trossulus* should occur on microarray probes from *M. galloprovincialis* without inherent sequence bias and should provide a reliable comparison of their transcriptional responses. Though debated, prudent evaluations of the sequence divergence by *in silico* approaches and phylogenetic data could expand the use of cross-

species hybridization as a compromise solution for investigating gene expression in species with unsequenced genomes (Costa et al., 2010; Nazar et al., 2010; Ptitsyn et al., 2010).

Gene expression profiling was also performed on gill RNA from mussels subjected to salinity stress (GEO series GSE25111). A total of 117 probes, out of 6,777 genes represented on the microarray, showed significant changes similar between *M. californianus* and *M. galloprovincialis* whereas 12 probes, denoting mRNA splicing, polyamine synthesis, exocytosis, translation, cell adhesion, and cell signaling, outlined species-specific responses. The study was based on AlexaFluor-labelling (555 and 647 fluorescence dyes) of amplified RNA, pooled reference samples, six biological replicates, and competitive hybridization in agreement to the recommended Agilent protocols. In addition to the overall stringent processing of the fluorescence signals, the heterologous hybridization design suggested the elimination of data from probes with low signal intensity (signal intensity < 150 % of the local background and hybridized spot diameter < 30 % of the nominal spot diameter).

The work performed at the A. Gracey's and G.N. Somero's laboratories (University of Southern California -Los Angeles, CA, U.S.A. and Stanford University -Palo Alto, CA, U.S.A., respectively) on *Mytilus* (GEO series GSE19031 and GSE25111) and other species is facing the fundamental aspects of the organism adaptation to fluctuating environments and global climate changes, and gene expression profiling has been essential to their findings. For instance, the study of gene-expression changes in the Californian mussels at different phases in the tidal cycle revealed at least four distinct physiological states, corresponding to metabolism and respiration phase, cell-division phase, and two stress-response signatures linked to moderate and severe heat-stress events. The metabolism and cell-division phases appeared to be functionally linked and anti-correlated in time whereas magnitude and timing of the above states resulted to be influenced by the microhabitat conditions according to the vertical position on the shore (Gracey et al., 2008). Based on comparative physiology, a recent paper offers an overview on the expected consequences of global climate changes (Somero, 2011).

Finally, the Mussel ImmunoChip 1.0 (GPL10758, April 2011) is a spotted oligonucleotide platform consisting of four-replicated 1820 oligomer probes plus unrelated controls prepared at CRIBI for the purposes of a recent European project (IMAGUQUANIM). Oligomers of 57 bases average length were designed at short distance from the 3' end of transcript sequences selected previously in Mytibase, the interactive knowledgebase of *M. galloprovincialis* which includes most of the ESTs publicly available for this species (Venier et al., 2009). Based on multiple criteria, the subset of transcripts selected from Mytibase as putatively immune-related molecules should denote central "players" of the mussel innate immunity or genes whose expression is modulated during the mussel responses to immunostimulation (Venier et al., 2011). In the platform description, the probe ID is hyperlinked to

the relative Mytibase record: for instance the probe MGO_07346 relates to MGC07346, a mussel transcript featured by the protein domain IPR000098-Interleukin 10 and yet functionally unknown. The performance of ImmunoChip 1.0 was tested with hemolymph samples collected at 3 and 48 h from *Vibrio*-challenged mussels (GEO series GSE23535) according to competitive hybridization of dye-swap labelled amplified RNA samples.

In agreement with the above descriptions, Figure 3 provides an updated summary of the nucleotide and protein sequences publicly available at July 2011 and highlights the importance of EST sequencing for the preparation of new DNA microarrays. More about the molecular “players” of the innate immunity and the immune responses of *M. galloprovincialis* is reported in the following paragraph.

How much can simple sequences tell us about the mussel immune responses?

Taking advantage of the continuous increase of the nucleotide and amino acid sequences in the public databases, the current methods of bioinformatics can extract instructive data from simple sequences: from the analysis of various gene/transcript regions to the evaluation of protein/peptide structure and to the comparative analysis of evolutionary differences across the tree of life. This procedural approach complements and integrates the data derived from long-standing disciplines such as measures of structural changes and protein amounts/activity, among others.

The overall analysis of 18,788 high-quality ESTs rationally organized in 7,112 independent clusters or singletons (Mytibase transcript collection) highlighted some particularly abundant transcript groups: namely, transcripts featured by a complement component C1q-like domain, antimicrobial peptide (AMP) precursors of all four families known in the Mediterranean mussel and many heterogeneous lectins including fibrinogen-related molecules (Venier et al., 2011). To explain the abundance of immune-related molecules in Mytibase it is important to remember that such collection has been prepared by 16 primary (5 from hemocytes) and 1 normalized cDNA libraries from mussels subjected to various challenges, for instance mussels immune stimulated with preparations of Gram positive and Gram negative cells and viral-like molecules.

Searches by protein domain revealed a total of 168 different Mytibase transcripts containing the C1q signature IPR001073, almost invariably associated with the overlapping TNF-like IPR008983 motif. Curiously, the C1q domain-containing proteins predicted from the transcript sequences, display a

short N-terminal signal peptide and a C-terminal globular domain but no central collagen-like repeats which are instead typical of vertebrate C1q domain-containing proteins. According to the current literature, these mussel proteins could represent secreted globular receptors, components of ancient complement pathways expected to mediate pathogen recognition and lysis (Dodds and Matsushita, 2007). The modularity and versatility of binding mediated by the globular C1q domain explain the variety of roles currently attributed to this still expanding family of proteins, and also supports their involvement in pathogen pattern recognition (Carland and Gerwick, 2010). The abundance and variety of mussel C1q domain-containing transcripts are consistent with this view.

One among these transcripts, named MgC1q, resulted to be expressed at detectable levels in the main tissues of naïve adult mussels, with the hemocytes showing the highest expression levels, and from 2 h post-fertilization up to 3 months later. The MgC1q expression was significantly modulated after mussel infection with Gram positive or Gram negative bacteria, data which confirm MgC1q as an immune-related gene. The striking molecular diversity of MgC1q was confirmed at both the DNA and cDNA levels, hence posing mechanistic questions on the origin of such variation (Gestal et al., 2010). Experimental findings and sequence analyses support the hypothesis of gene duplication, functional diversification and positive selection of many C1qDC variants in selected taxa, including the mussel lineage (Gerdol et al., 2011). Defensins, mytilins, myticins and mytimycins are cationic antimicrobial peptides stabilized by 4 intrachain disulphide bonds (6 in mytimycin) in a typical 3-D motif (Yeaman and Yount, 2007). A remarkable diversity of a new group of myticins, with specific variant profiles detectable in single mussels, was reported in *M. galloprovincialis* (Pallavicini et al., 2008; Costa et al., 2009). Following the discovery of the myticin-C variants, their molecular diversity and evolution has been further discussed (Padhi and Verghese, 2008) and the most recent findings indicate myticin C as a chemotactic molecule with antiviral activity and immunoregulatory properties (Balseiro et al., 2011). Just one singleton and other four similar sequences denote the antifungal AMP mytimycin in Mytibase (rare transcript). Mytimycin is composed by 54 aminoacids (6.2 - 6.3 kDa, 12 cysteines) and two main precursor variants, both featured by a signal peptide and a C-terminal extension, are expressed in mussels from different European regions (Sonthi et al., 2011). The presence of a calcium binding (EF hand) motif in the C-terminal extension suggests further characterization of such unusual AMP.

The "effector" role of the mussel antimicrobial peptides (AMPs) is confirmed in many experimental studies and a comprehensive review have been recently provided (Li et al., 2011). Whether these effectors can modulate the mussel immune responses with mechanisms other than membrane disruption, as reported for mammalian AMPs, it is not clear. Based on deep amplicon sequencing, the

sequence diversity of mussel AMPs is now under study in natural mussel populations from different geographical regions and in mussels challenged with bacterial cells.

Lectins are a rather heterogeneous protein family comprising 8 to 15 subgroups, depending on the scientist's view (Dodd and Drickamer, 2001).

Lectins typically possess carbohydrate binding domains and participate in many cell processes. Similarly to the mammalian C1q, the C-terminal fibrinogen-like domain IPR002181 of ficolins forms a tulip-like structure able to bind the carbohydrate residues of foreign and apoptotic cells (with consequent opsonization, phagocytosis and cell clearance) or triggering the proteolytic complement cascade and pathogen lysis. Fibrinogen-related lectin proteins (FREPs) are expressed also in mussels (Venier et al., 2011) and are codified by at least 2 (*M. edulis*) 4 (*M. californianus*) and 7 genes (*M. galloprovincialis*) (Gorbushin and Iakovleva, 2011). These molecules can be regarded as immune pattern-recognition receptors and their involvement in the native immunity is supported by the evidence of species-specific expansion of FREPs in the snail *Biomphalaria glabrata* and the mosquito *Anopheles gambiae* (Waterhouse et al., 2007; Zhang et al., 2008). In mussel, FREPs are significantly up-regulated after bacterial infection or PAMP treatment, and display opsonizing activity similar to that of mammalian ficolins; moreover, the different sets of FREP sequences detected among and within individuals further emphasize the great complexity of the invertebrate immune systems (Romero et al., 2011). Other lectin-like sequences expressed in mussels are commented in Venier et al. (2011).

The cases reported above are a few examples of the many classes of transcripts specifically expressed or modulated during the mussel response to potential pathogens. Considering in a dynamic view the behaviour of one cell population only, the versatile mussel hemocytes, one can imagine that almost all cellular processes could be influenced by the contact with pathogen-associated molecular patterns: from the cytoskeleton remodelling supportive of chemotaxis, migration and phagocytosis to the intracellular signalling possibly shaping the inflammatory response and finely tuned expression of many regulatory and effector genes. Cross-talking signalling pathways have been traced in mussel and the Mytibase collection includes transcripts denoting the regulatory cytokine MIF (migration inhibiting factor) and cytokine-related molecules, consistent with the idea of an invertebrate cytokine network (Malagoli, 2010). The recent definition of a species-specific ImmunoChip aims to the experimental validation of a selected subset of transcripts: a synopsis of the main gene expression changes detected in mussels at 3 and 48 h after challenge with live bacterial cells is reported in Fig. 4. The general AMP down-regulation observed in this particular laboratory treatment was confirmed by quantitative PCR data and is discussed also in Li et al. (2010).

Concluding remarks

EST sequencing and DNA microarrays have substantially improved the identification of genes expressed in the *Mytilus* species. Compared to the first EST collection and the related cDNA microarray, Mytibase includes an interesting variety of immune-related molecules which can be further characterized with traditional and innovative approaches as exemplified by Romero et al. (2011). Nonetheless, in the Mytibase collection about half of the mussel transcripts are still unknown, devoid of functional annotation. Hence, much work remains to be done both *in silico* and in laboratory to provide a comprehensive view of the global gene transcription in mussels, particularly the part of the transcriptome mediating the response to potential invaders (immunome).

Undoubtedly, the application of the available mussel DNA microarray platforms can further reveal expression trends of different gene categories and identify useful markers of functional state, if not global molecular signatures useful to disentangle the complex mussel physiology. Depending on the study design and on the type of microarray platform, independent validation of the expression data can be accomplished by quantitative PCR or with other experimental measures. All the steps of the DNA microarray testing could be used to strengthen the final data interpretation, from the microarray preparation strategy to the stringency of the hybridization reaction to the algorithms applied to data processing.

The maintenance of the physical collection of the cDNAs, i.e., recombinant bacterial clones, is a prerequisite for the use of spotted cDNA microarrays (for instance, the current use of Mytarray 1.0 slides, printed at the CRIBI facility depends on long work performed at the Department of Biology, University of Padua). Such work is not more affordable as long as the clustered ESTs increase in number, and external commercial services or deep sequencing become an attractive alternative.

As a matter of fact, next-generation sequencing (NGS) technologies are now complementing and challenging the DNA microarrays as alternative tools for genome analysis and transcriptome sequencing (Hurd and Nelson, 2009; Morozova et al., 2009). For instance, the so called 454 pyrosequencing has been already applied to the study of tissue-specific expression patterns in *M. galloprovincialis* (Craft et al., 2010) and many laboratories in the world are now investing in this kind of work.

Acknowledgments

This work has been sequentially supported by FOOD-CT-2005-007103 (IMAGUQUANIM), FP7-KBBE-2010-4-266157 (BIVALIFE) and, partially, 20084BEJ9F (PRIN08).

Figures

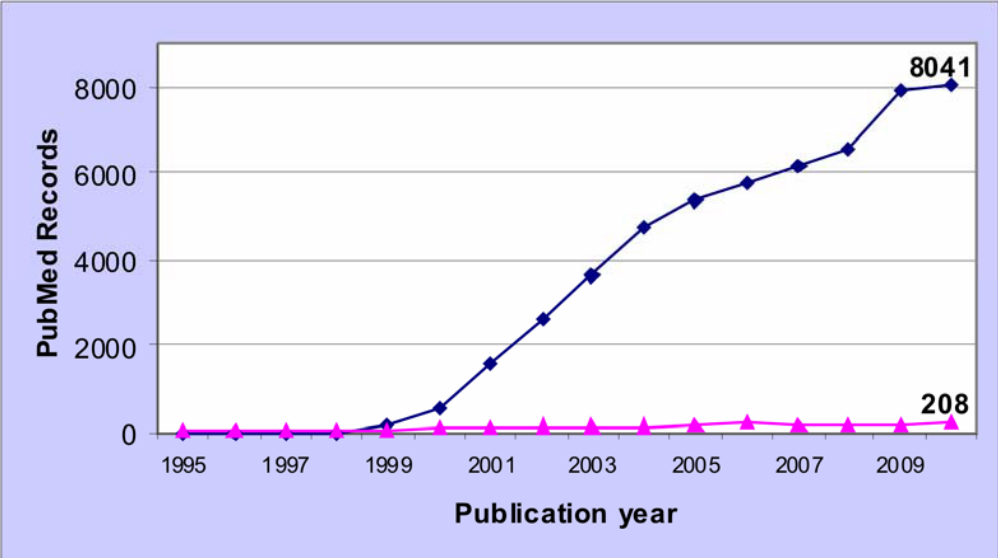


Fig. 1 Number of PubMed publications including the terms "DNA-microarray" (blue line) or "Mytilus" (purple line) from 1995 to 2010. Has the DNA microarray revolution reached its peak?

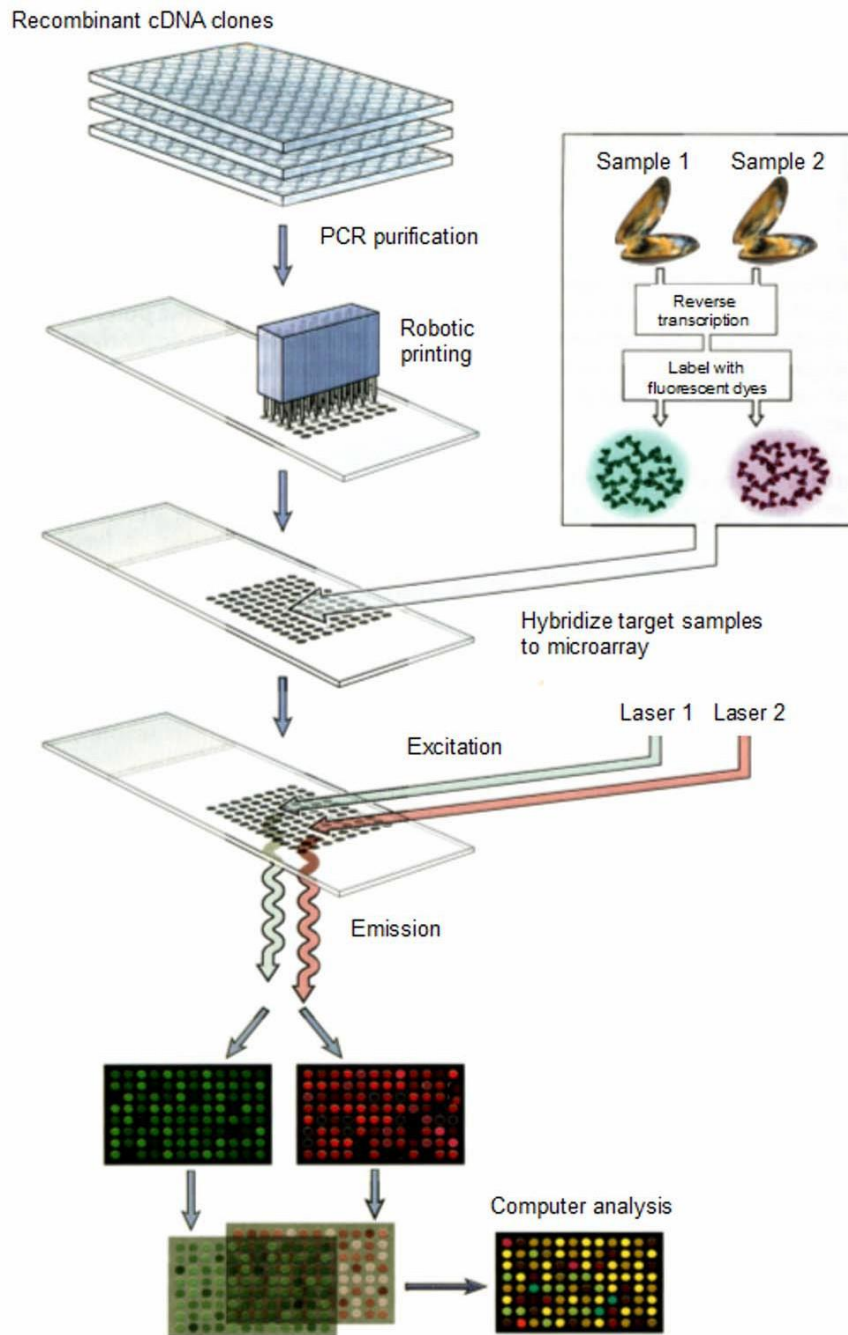


Fig. 2 Work diagram referred to the competitive hybridization of two dye-swap-labelled samples on a cDNA microarray with two-channel detection of the fluorescence signals (modified from Gibson and Muse, 2004).

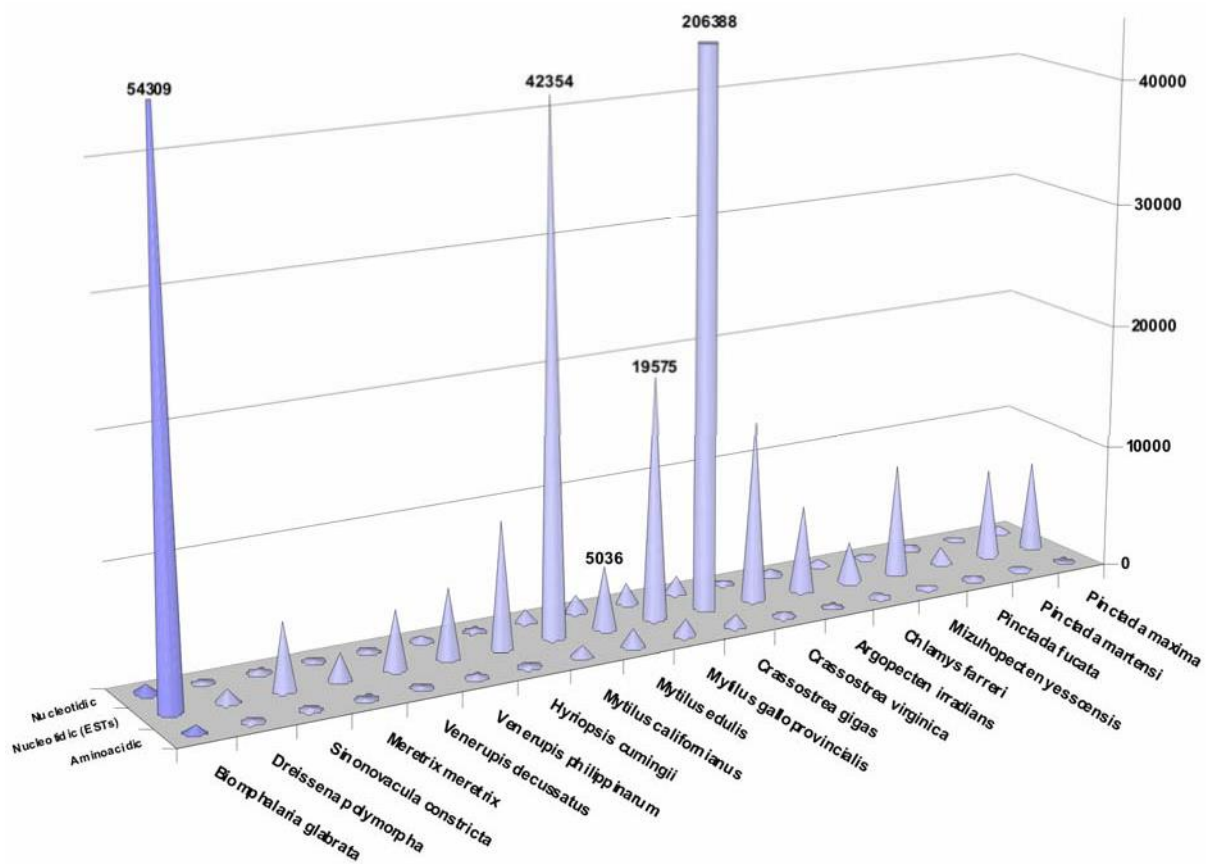


Fig. 3 Number of sequence records available for selected mollusc species at the Natl Center for Biotechnology Information at July 2011. DNA, RNA and protein sequences refer to *Biomphalaria glabrata* (Gastropoda) and bivalves belonging to the Veneroidea, Unionoidea, Mytiloidea, Ostreoidea, Pectinoidea and Pterioidea orders.

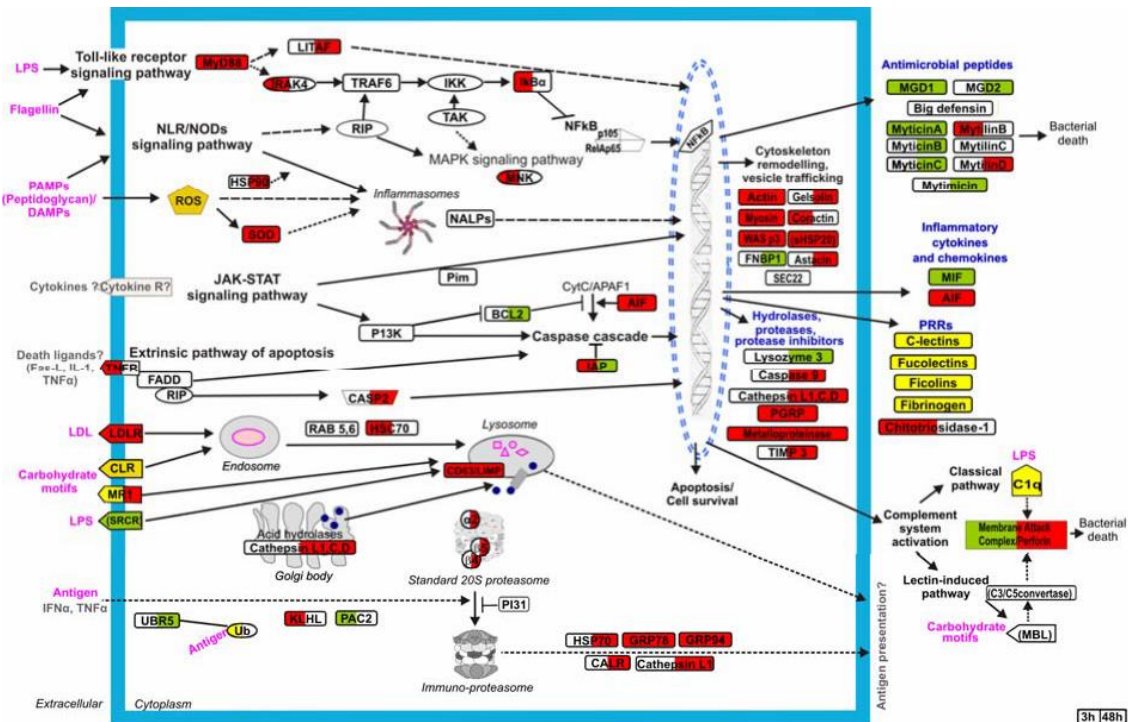


Fig. 4 Main transcriptional changes detected in mussels at 3 h and 48 h from the injection of live *Vibrio* cells (modified from Venier et al., 2011). Only relevant molecular "players" represented in the ImmunoChip of *M. galloprovincialis* are reported (framed). In each frame, the detected expression trends are indicated in red, green and yellow (up- and down-regulation and not homogeneous trends, respectively). Annotations based only on protein domains are reported in brackets. Overall, the figure draws the attention to a number of mussel genes, still not characterized, whose expression is modulated in response to immune stimulation.

Abbreviation list (Fig. 4):

AIF: Allograft Inflammatory Factor

APAF1: Apoptotic Peptidase Activating Factor 1 BCL2: Baculoviral apoptosis regulator 2

C1-C5: Complement component 1-5 CALR: Calreticulin

CASP: Caspase

CD63/LIMP: Tetraspanin-7 (lysosome membrane protein) CLR: C-type Lectin Receptor

CLR: C-type Lectin Receptor

DAMPs: damage-associated molecular patterns FADD: FAS (TNFRSF)-Associated via Death Domain

FNBP1: Formin-Binding Protein 1

GRP 78/94: Glucose-Regulated Protein 78/94 HSC70: Heat Shock Cognate 70

HSP70/90: Heat Shock Protein 70/90 IAP: Inhibitor of Apoptosis Proteins;

IKB α : Inhibitor of nuclear factor Kappa-B kinase alpha

IKK: Inhibitor of nuclear factor Kappa-B Kinase complex IL: InterLeukin

IRAK4: Interleukin Receptor-Associated Kinase 4

JAK: Janus kinase KLHL: Kelch-like protein

LDLR: Low-Density Lipoprotein Receptor

LITAF: LPS-Induced TNFAlpha Factor LPS: LipoPolySaccharide

MAPKs: Mitogen-Activated Protein Kinases MBL: Mannose Binding Lectin

MGD1/2: Mytilus galloprovincialis Defensin 1 /2 MIF: Migration Inhibitory Factor

MNK: MAP kinase-interacting serine/threonine-protein kinase

MR1: Mannose Receptor 1

MyD88: Myeloid Differentiation primary response gene 88 NALPs: NATCH, LRR, and PYR containing proteins

NFkB: Nuclear Factor of kappa light polypeptide gene enhancer in B-cells

NLR: NOD-Like Receptor

NOD: Nucleotide Binding Oligomerization Domain P13K: Phosphatidylinositol-4,5-bisphosphate 3-Kinase

PAC2: Proteasome Assembly Chaperone 2 PAMPs: Pathogen Associated Molecular Patterns PGRP: Peptidoglycan Recognition Protein

PI31: Proteasome Inhibitor PI31 subunit

Pim: proto-oncogene serine/threonine-protein kinase Pim PRR: Pathogen Recognition Receptors

RAB: Ras-related gtp-Binding protein

RIP: Receptor-Interacting serine-threonine kinase ROS: Reactive Oxygen Species

SEC22: vesicle transport protein SEC22

SOD: SuperOxide Dismutase

STAT: Signal Transducer and Activator of Transcription protein

SRCR: Scavenger Receptor Cysteine-Rich protein precursor

TAK: mitogen activated protein kinase kinase

TIMP3: Tissue Inhibitors of MetalloProteinase 3 TNF: Tumour Necrosis Factor

TNFR: Tumour Necrosis Factor Receptor

TRAF6: TNF receptor-associated factor 6 Ub: Ubiquitin

UBR5: Ubiquitin protein Ligase E3 (component n-recognin 5)

α 2: proteasome subunit alpha type 2

β 4, β 5: Proteasome subunit beta type 4/5

References

- Balseiro P, Falcó A, Romero A, Dios S, Martínez- López A, Figueras A, et al. Mytilus galloprovincialis Myticin C: a chemotactic molecule with antiviral activity and immunoregulatory properties. PLoS One. 2011;6(8):e23140.
- Banni M, Negri A, Mignone F, Boussetta H, Viarengo A, Dondero F. Gene Expression Rhythms in the Mussel Mytilus galloprovincialis (Lam.) across an Annual Cycle. PLoS One 6(5): e18904, 2011.
- Ben Said O, Goñi-Urriza M, El Bour M, Aissa P, Duran R. Bacterial community structure of sediments of the Bizerte lagoon (Tunisia), a southern Mediterranean coastal anthropized lagoon. Microb. Ecol. 59: 445-456, 2010.
- Carland TM, Gerwick L. The C1q domain containing proteins: Where do they come from and what do they do? Dev. Comp. Immunol. 34: 785-790, 2010.
- Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC Bioinformatics 8: 412, 2007.
- Costa F, Alba R, Schouten H, Soglio V, Gianfranceschi L, Serra S, et al. Use of homologous and heterologous gene expression profiling tools to characterize transcription dynamics during apple fruit maturation and ripening. BMC Plant Biol. 10: 229, 2010.
- Costa MM, Dios S, Alonso-Gutierrez J, Romero A, Novoa B, Figueras A. Evidence of high individual diversity on myticin C in mussel (Mytilus galloprovincialis). Dev. Comp. Immunol. 33: 162-170, 2009.
- Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B, et al. Pyrosequencing of Mytilus galloprovincialis cDNAs: Tissue-Specific Expression Patterns. Plos One 5(1): e8875, 2010.
- Dodd RB, Drickamer K. Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. Glycobiology. 11(5):71R-79R, 2001. Dodds AW, Matsushita M: The phylogeny of the complement system and the origins of the classical pathway. Immunobiology 212: 233-243, 2007.
- Evans TG, Somero GN. Phosphorylation events catalyzed by major cell signaling proteins differ in response to thermal and osmotic stress among native (Mytilus californianus and Mytilus trossulus) and invasive (Mytilus galloprovincialis) species of mussels. Physiol. Biochem. Zool. 83: 984-996, 2010.
- GEO. The Gene Expression Omnibus can be accessed at <http://www.ncbi.nlm.nih.gov/geo>.

- Gerdol M, Manfrin C, De Moro G, Figueras A, Novoa B, Venier P, et al. The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: a widespread and diverse family of immune-related molecules. *Dev. Comp. Immunol.* 35: 635-643, 2011.
- Gestal C, Pallavicini A, Venier P, Novoa B, Figueras. MgC1q, a novel C1q-domain-containing protein involved in the immune response of *Mytilus galloprovincialis*. *Dev. Comp. Immunol.* 34: 926-34, 2010.
- Gibson G, Muse SV. *A Primer of Genome Science*. Sinauer Associates Inc., Sunderland, MA USA, 2004.
- Gibson G. The environmental contribution to gene expression profiles. *Nat. Rev. Genet.* 9: 575-581 2008.
- Gorbushin AM, Iakovleva NV. A new gene family of single fibrinogen domain lectins in *Mytilus*. *Fish Shellfish Immunol.* 30: 434-438, 2011.
- Gracey AY, Chaney ML, Boomhower JP, Tyburczy WR, Connor K, Somero GN. Rhythms of gene expression in a fluctuating intertidal environment. *Curr. Biol.* 18: 1501-1507, 2008.
- Hook SE. Promise and progress in environmental genomics: a status report on the applications of gene expression-based microarray studies in ecologically relevant fish species. *J. Fish Biol.* 77: 1999-2022, 2010.
- Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief. Funct. Genomic. Proteomic.* 8:174-183, 2009.
- IMAQUANIM. Facts on this FP6 Integrated Project are at <http://imaquanim.dfvf.dk/info>.
- Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostatistics* 2:183-201, 2001.
- Larsen PF, Schulte PM, Nielsen EE. Gene expression analysis for the identification of selection and local adaptation in fishes. *J. Fish Biol.* 78: 1-22, 2011.
- Lenoir T, Giannella E. The emergence and diffusion of DNA microarray technology. *J. Biomed. Discov. Collab.* 1: 11, 2006.
- Li H, Parisi MG, Parrinello N, Cammarata M, Roch P. Molluscan antimicrobial peptides, a review from activity-based evidences to computer- assisted sequences. *Inv. Surv. J.* 8: 85-97, 2011.

Li H, Venier P, Prado-Alvarez M, Gestal C, Toubiana M, Quartesan R, et al. Expression of *Mytilus* immune genes in response to experimental challenges varied according to the site of collection. *Fish Shellfish Immunol.* 28: 640-648, 2010.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnology* 14: 1675-1680, 1996.

Lockwood BL, Sanders JG, Somero GN. Transcriptomic responses to heat stress in invasive and native blue mussels (genus *Mytilus*): molecular correlates of invasive success. *J. Exp. Biol.* 213(Pt 20): 3548-3558, 2010.

Lockwood BL, Somero GN. Transcriptomic responses to salinity stress in invasive and native blue mussels (genus *Mytilus*). *Mol. Ecol.* 20: 517-29, 2011.

Malagoli D. Cytokine network in invertebrates: the very next phase of comparative immunology. *Inv. Sur. J.* 7: 146-148, 2010.

Manfrin C, Dreos R, Battistella S, Beran A, Gerdol M, Varotto L, et al. Mediterranean mussel gene expression profile induced by okadaic acid exposure. *Environ. Sci. Technol.* 44: 8276-8283, 2010.

Martın-Subero JI, Esteller M. Profiling epigenetic alterations in disease. *Adv. Exp. Med. Biol.* 711: 162-177, 2011.

MESH. Current MEDical Subject Headings can be accessed at <http://www.ncbi.nlm.nih.gov/mesh>.

Morozova O, Hirst M, Marra MA. Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet.* 10: 135-151, 2009.

Nazar RN, Chen P, Dean D, Robb J. DNA chip analysis in diverse organisms with unsequenced genomes. *Mol. Biotechnol.* 44: 8-13, 2010.

NCT. The National Center for Toxicogenomics is at the National Institute of Environmental Health Sciences (<http://www.niehs.nih.gov/research/atniehs/nct.cfm>).

Padhi A, Verghese B. Molecular diversity and evolution of myticin-C antimicrobial peptide variants in the Mediterranean mussel, *Mytilus galloprovincialis*. *Peptides* 229: 1094-10101, 2008.

Pallavicini A, Costa MM, Gestal C, Dreos R, Figueras A, et al. High sequence variability of myticin transcripts in hemocytes of immune-stimulated mussels suggests ancient host- pathogen interactions. *Dev. Comp. Immunol.* 32: 213-226, 2008.

- Pantartzzi C, Drosopoulou E, Yiangou M, Drozdov I, Tsoka S, Ouzounis CA, et al. Promoter complexity and tissue-specific expression of stress response components in *Mytilus galloprovincialis*, a sessile marine invertebrate species. *PLoS Comput. Biol.* 6: e1000847, 2010.
- Ptitsyn A, Schlater A, Kanatous S. Transformation of metabolism with age and lifestyle in Antarctic seals: a case study of systems biology approach to cross-species microarray experiment. *BMC Syst. Biol.* 4: 133, 2010.
- Rogers S, Cambrosio A. Making a new technology work: the standardization and regulation of microarrays. *Yale J. Biol. Med.* 80: 165-178, 2007.
- Romero A, Dios S, Poisa-Beiro L, Costa MM, Posada D, Figueras A, et al. Individual sequence variability and functional activities of fibrinogen-related proteins (FREPs) in the Mediterranean mussel (*Mytilus galloprovincialis*) suggest ancient and complex immune recognition models in invertebrates. *Dev. Comp. Immunol.* 35: 334-344, 2011.
- Romero A, Estévez-Calvar N, Dios S, Figueras A, Novoa B. New insights into the apoptotic process in mollusks: characterization of caspase genes in *Mytilus galloprovincialis*. *PLoS One* 6(2):e17003, 2011.
- Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470, 1995.
- Shaw JP, Dondero F, Moore MN, Negri A, Dagnino A, Readman JW, et al. Integration of biochemical, histochemical and toxicogenomic indices for the assessment of health status of mussels from the Tamar Estuary, UK *Mar. Environ. Res.* [in press] 2011.
- Shi L, Perkins RG, Fang H, Tong W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr. Opin. Biotechnol.* 19: 10-18, 2008.
- Somero GN. Comparative physiology: a "crystal ball" for predicting consequences of global change. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 301: R1-R14, 2011
- Sonthi M, Toubiana M, Pallavicini A, Venier P, Roch P. Diversity of coding sequences and gene structures of the antifungal peptide mytimycin (MytM) from the Mediterranean mussel, *Mytilus galloprovincialis*. *Mar. Biotechnol.* (NY) [in press] 2011.

- Steinberg CE, Stürzenbaum SR, Menzel R. Genes and environment - striking the fine balance between sophisticated biomonitoring and true functional environmental genomics. *Sci. Total Environ.* 400: 142-161, 2008.
- Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. *Nat. Struct. Mol. Biol.* 17: 1169-1174, 2010.
- Venier P, De Pittà C, Bernante F, Varotto L, De Nardi B, Bovo G, et al. MytiBase: a knowledge base of mussel (*M. galloprovincialis*) transcribed sequences. *BMC Genomics* 10: 72, 2009.
- Venier P, De Pittà C, Pallavicini A, Marsano F, Varotto L, Romualdi C, et al. Development of mussel mRNA profiling: Can gene expression trends reveal coastal water pollution? *Mutat. Res.* 602: 121-134, 2006.
- Venier P, Pallavicini A, De Nardi B, Lanfranchi G. Towards a catalogue of genes transcribed in multiple tissues of *Mytilus galloprovincialis*. *Gene* 314: 29-40, 2003.
- Venier P, Varotto L, Rosani U, Millino C, Celegato B, Bernante F, et al. Insights into the innate immunity of the Mediterranean mussel *Mytilus galloprovincialis*. *BMC Genomics* 12: 69, 2011.
- Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, et al. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316: 1738-1743, 2007.
- Yeaman MR, Yount NY. Unifying themes in host defence effector polypeptides. *Nat. Rev. Microbiol.* 5: 727-740, 2007.
- Zhang SM, Zeng Y, Loker ES. Expression profiling and binding properties of fibrinogen-related proteins (FREPs), plasma proteins from the schistosome snail host *Biomphalaria glabrata*. *Innate Immun.* 14: 175-189, 2008.

REVIEW

Physiological and molecular responses of bivalves to toxic dinoflagellates**C Manfrin¹, G De Moro¹, V Torboli¹, P Venier², A Pallavicini¹, M Gerdol¹**¹*Department of Life Sciences, University of Trieste, Trieste, Italy*²*Department of Biology, University of Padua, Padua, Italy*

*Accepted October 31, 2012***Abstract**

Dinoflagellates and other microalgae can produce a wide spectrum of heterogeneous toxic molecules, which are the main responsible of shellfish poisoning syndromes. During seasonal harmful algal blooms (HABs), many filter-feeding marine invertebrates, including bivalve mollusks, can accumulate phycotoxins at extremely high levels, thus representing a serious threat to human health. Furthermore, HABs also have a severe impact on the aquaculture sector due to the forced prolonged closure of large harvesting areas.

Although the targets and mechanism of action of many phycotoxins have been extensively studied on vertebrate model organisms, so far just a little attention has been focused on their effects on marine invertebrates. Here we provide an overview about the molecular response of marine bivalves to phycotoxins, with a particular focus on toxins produced by dinoflagellates. Even though large-scale genomic and proteomic approaches on mollusks are still hindered by the limited molecular knowledge of these organisms, a few studies exploiting the most recent technological advances provide promising perspectives for a better comprehension of the mechanisms involved in shellfish toxicity and for the identification of molecular markers of contamination.

Abbreviations: ASP: Amnesic Shellfish Poisoning; AZA/AZT: Azaspiracid/ Azaspiracid Shellfish Poisoning; pBTx: Brevetoxin; DA: Domoic Acid; DSP/DST: Diarrhetic Shellfish Poisoning/Toxin; DTX: Dinophysistoxin; GYM: Gymnodimine; HAB: Harmful Algal Bloom; NSP: Neurotoxic Shellfish Poisoning; OA: Okadaic Acid; PLTX: Palytoxin; PSP/PST: Paralytic Shellfish Poisoning/Toxin, PTX: Pectenotoxin; SPX: Spirolides; STX: Saxitoxin; YTX: Yessotoxin.

Introduction

The largest component of the large universe of algae, estimated to comprise between one and ten million different species, is represented by unicellular microalgae (Barsanti and Gualtieri, 2006). Some of them can produce, through complex and not completely understood biochemical processes, toxic compounds of various chemical composition and mode of action, collectively named phycotoxins. In response to favorable environmental conditions, toxic microalgae can proliferate and/or aggregate to form dense concentrations, called “Harmful Algal Blooms” (HABs). Phycotoxins are responsible of a number of human illnesses associated with the consumption of contaminated seafood and, in some cases, with respiratory exposure to aerosolized toxins. In fact, filter-feeding shellfish, zooplankton, and herbivorous fishes can ingest these algae and act as vectors to humans either directly (e.g. shellfish) or through further food web transfer to higher trophic levels (Van Dolah, 2000).

Although seasonal micro-algal blooms are considered as a natural phenomenon, their frequency of occurrence appears to have increased in the recent years. Certainly, the worsening of the hygienic characteristic of the aquaculture areas, the transportation of ship ballast water (Carlton and Geller, 1993), water eutrophication (Reigman, 1998) and global climate changes (Richardson, 1997) are factors which altogether provide favorable conditions for the spreading and the growth of toxic algae, thus contributing to the increase of threats for human by the consumption of contaminated shellfish. Unfortunately, algal toxins are not detectable by sight or smell and contaminated seafood appears normal and in most cases they are heat stable and thereby largely unaffected by cooking (Sobel and Painter, 2005).

The consumption of contaminated seafood often results in shellfish poisoning syndromes, which are classified according to symptoms and the chemical nature of the toxins involved. **Fig. 1** summarizes the geographical occurrence of the six main poisoning syndromes; although some are endemic of specific areas, their altogether distribution clearly points out shellfish toxicity as a global problem for human health, which consequently have an important economic impact on aquaculture worldwide. International laws promoted by environmental monitoring agencies and food safety associations, impose the routinely control of the toxicity of seawater and market shellfish stocks. While these monitoring strategies prevent episodes of massive intoxication, the closure of aquaculture hatcheries, sometimes even for a very prolonged time, is responsible of severe economic losses in this sector. In the following section, we provide a brief overview of the main human syndromes associated with contaminated marine bivalves consumption. The causative algae, the toxin structure and the most characterizing symptoms are reported in **Table 1**.

Main classes of shellfish poisoning syndromes

Amnesic Shellfish Poisoning (ASP) is caused by the consumption of domoic acid (DA) contaminated food. This phenomenon was first documented in 1987 in Canada, with 105 cases of acute human poisoning, including 3 casualties, related to the consumption of contaminated mussels (Hallegraeff, 1993). DA, produced by marine diatoms belonging to the genus *Pseudo-nitzschia*, is a water-soluble tricarboxylic acid that acts as an analog of the neurotransmitter glutamate and is a potent glutamate receptor agonist. In mammals DA intoxication cause both gastrointestinal and neurological disorders as headache, disorientation, short-term memory loss, brain damage, and in severe cases it can also be fatal (Todd, 1993).

Diarrhetic Shellfish Poisoning (DSP), a syndrome characterized by diarrhea, nausea, vomiting and abdominal cramps, is one of the most common pathologies associated to HABs worldwide. Diarrhetic toxins (DSTs) are lipophilic molecules such as okadaic acid (OA) and the structurally related dinophysistoxins (DTXs), produced by *Dinophysis* and *Prorocentrum* spp. (Yasumoto *et al.*, 1985). OA selectively inhibits protein phosphatases and modifies the phosphorylation state of many regulators of cellular processes involved in metabolism and various cell activities, causing diarrhea because of the impairment of the sodium secretion control by intestinal cells (Van Dolah, 2000).

The marine biotoxins called azaspiracids (AZA) cause a syndrome similar to DSP (AZP), although they chemically differ from any previously known toxin found in shellfish (Satake *et al.*, 1998). AZA accumulates in bivalve mollusks that feed on toxic microalgae of the genus *Protoperdinium*, previously considered to be toxicologically harmless.

Another widespread syndrome caused by contaminated bivalve mollusks is the Paralytic Shellfish Poisoning (PSP), caused by saxitoxin (STX) and its analogues (GTXs), globally indicated as Paralytic Shellfish Toxins (PSTs). Dinoflagellates of the genus *Alexandrium*, in particular *A. minutum*, *A. catenella*, *A. tamarense* and *A. fundyense*, are the most numerous PST producers and are responsible for PSP blooms all around the world (**Fig. 1**). In fact, almost 2000 PSP cases are reported per year in human, with occasional fatal consequences (Hallegraeff, 1993). PSTs inhibit the voltage-dependent sodium channel conductance causing the blockade of neuronal activity. Symptoms include nausea, diarrhea, abdominal ache, shortness of breath, dry mouth, confused speech, tingling or burning sensations (Clark *et al.*, 1999).

Neurotoxic shellfish poisoning (NSP) is caused by brevetoxins (PbTx); the main symptoms are gastrointestinal and neurological, including nausea, loss of motor control and muscular ache. The formation of toxic aerosol by wave action produces respiratory asthma-like symptoms, even though no fatalities have been reported so far. PbTx is produced by *Karenia brevis* and binds with high

affinity to the voltage-dependent sodium channel, altering its sensitivity and inhibiting its inactivation (Wang, 2008).

The sixth and last group of shellfish poisoning is the Ciguatera Fish Poisoning (CFP), an intoxication caused by a heterogeneous group of toxins, including ciguatoxin, maitotoxin, palytoxin (PLTX) and others. While the main concerns for human health derive from the consumption of contaminated fish (Lewis and Holmes, 1993), PLTX and its analogues, produced by species of the *Ostreopsis* genus, can also be accumulated in bivalves. Likewise ciguatoxin, PLTX also lowers the threshold for the opening of voltage-gated sodium channels in synapses of the nervous system, determining severe neurological disorders, including lethargy, muscle spasms, myalgia, cyanosis, respiratory distress, rhabdomyolysis and even death in severe cases (Ramos and Vasconcelos, 2010).

Beside the toxins classifiable among the six large classes listed above, microalgae can produce an extremely broad spectrum of additional toxic compounds. Some species of dinoflagellates, such as *Lingulodinium polyedrum*, *Gonyaulax spinifera* and *Protoceratium reticulatum* produce yessotoxins (YTXs), structurally related to PbTx and ciguatoxins. YTXs cause diarrhea and therefore they were initially wrongly classified as DSTs, only to be later assigned to a novel independent group after the discovery of their different mechanism of action (Tubaro *et al.*, 2010). Pectenotoxins (PTXs) belong to a group of polyether-lactone toxins that, like YTXs cause symptoms similar to DSP. They are exclusively produced by *Dinophysis* spp., which have a large distribution worldwide (Draisci *et al.*, 1996). Gymnodimines (GYMs) and spirolides (SPXs) produced by *Karenia selliformis* and *Alexandrium ostefeldii*, respectively (Miles *et al.*, 2003; Roach *et al.*, 2009) are emerging lipophilic marine toxins that belong to a heterogeneous group of macrocyclic compounds called cyclic imines. Since their discovery in the early 1990s, GYMs and SPXs are well known due their fast acting toxicity in mouse bioassay, by blocking the nicotinic receptors and causing neurological symptoms (Munday *et al.*, 2004).

Moreover, dinoflagellates are certainly capable of producing several harmful compounds which have not been fully characterized yet. As a matter of fact, the biochemical potential of microalgae is underestimated and many algal natural products, including toxins, remain yet to be discovered (Sasso *et al.*, 2012). Therefore it is not surprising that the number of phycotoxins isolated from marine microalgae continues to increase exponentially.

Nevertheless, dinoflagellates are not the only source of marine toxins, as certain diatoms (such as the DA producing *Pseudo-nitzschia* spp.) and several species of seawater Cyanobacteria produce compounds hazardous to human health. While these toxins are extremely diverse both by chemical composition and by physiological effects, these aspects will be not discussed in the present review,

which is mainly focused on the effects of toxic dinoflagellates. For a more comprehensive overview suggesting the reading of more specific literature on the topic (Ferrão-Filho and Kozlowsky-Suzuki, 2011; Funari and Testai, 2008).

Toxicological studies on human and other vertebrate model organisms

The study of marine toxins has been historically connected with the hazard they represent to human health. Therefore, the large majority of toxicological studies performed so far have been focused on model vertebrates or human cell lines. Moreover, the most used method for marine algal toxins detection is the mouse bioassay developed by the Japanese Ministry of Health and Welfare (Yasumoto *et al.*, 1978). Although this outworn method is currently being replaced by other more reliable tests, it has been used worldwide as the main tool for shellfish toxicity biomonitoring for several decades.

In parallel to this test, routinely used for the detection of contamination, many research studies investigated more specific aspects of the toxic effects of marine drugs, unraveling their molecular targets, their mode of action and the kinetics of accumulation and detoxification in model organisms. Mice and other vertebrates have been repeatedly used as model organisms for *in vivo* studies to better understand the effects of several classes of phycotoxins, including AZA (Vilariño, 2008), OA (Le Hégarat *et al.*, 2006), STX (Andrinolo *et al.*, 1999), and YTX (Franchini *et al.*, 2010), on human. Also cell lines provide useful models: a number of *in vitro* studies clarified the mode of action of diverse phycotoxins, including YTX (Bianchi *et al.*, 2004; Franchini *et al.*, 2010), PLTX (Sala *et al.*, 2009), OA (Sala *et al.*, 2009; Valdiglesias *et al.*, 2010), STX (Perez *et al.*, 2011) and AZA (Twiner *et al.*, 2012). Moreover, genomics and proteomics methods applied to vertebrate models contributed to elucidate the molecular pathways acting in response to shellfish poisoning (Ryan *et al.*, 2010; Wang *et al.*, 2012).

Phycotoxin effects on bivalves

A direct comparison between the widely documented molecular effects of phycotoxins on vertebrates and bivalves is definitely hindered by several obstacles. One of the key factors is the specificity of interaction of many toxins with their molecular targets, which are often membrane channels. The sequence divergence among species explains why bivalves are often completely insensible to many toxins having a lethal effect on human and other vertebrates and vice-versa. In fact, biotransformation

processes which likely reduce toxicity in bivalves, can sometimes produce analogues which are even more dangerous to human than the original compound (O'Driscoll *et al.*, 2011). The specificity of interaction between phycotoxins and their molecular targets is refined to the point that polymorphisms to single ion channels can even lead to dramatic differences in sensitivity within bivalve populations (Bricelj *et al.*, 2010).

Nevertheless, over the past decades numerous studies have been carried out to explore the specific effects of phycotoxins on marine invertebrates, even though their focus has been mainly addressed on physiological adaptations and behavioral modifications. This bias on non-molecular studies is caused by the still limited genetic knowledge of these organisms. In fact, before the next generation sequencing era, the main pool of bivalve genetic knowledge came from large EST collections (Venier *et al.*, 2009), but the recent technological advances have quickly led to the availability of many transcriptomes and even to the complete sequencing of the oyster genome (Zhang *et al.*, 2012).

Studies of physiological responses

The rich literature documenting the behavioral, physiological and histo-pathological alterations in contaminated filter-feeding mollusks, despite not providing outright molecular data, represent an important base of knowledge both for a better planning of molecular biology experiments and for the interpretation of the results in their biological context.

Hemocyte parameters

Internal defense in bivalve mollusks, is characterized by a non-adaptive, non-specific, innate immune system (Loker *et al.*, 2004), based on specialized circulating cells, the hemocytes, involved in pathogen recognition and elimination through the production of a broad spectrum of defense molecules (Venier *et al.*, 2011). Therefore, the study of the effects of phycotoxins on bivalve immunity necessarily takes into account the monitoring of hemocyte parameters.

The feeding on PSP dinoflagellate species determines an increase of hemocyte counts, accompanied by changes in the hemocyte subpopulations and alteration of the phagocytic activity in some species (Hégaret and Wikfors, 2005a; Hégaret and Wikfors, 2005b). On the contrary, some studies report non-significant effects on hemocyte number, morphology, or functions (Hégaret *et al.*, 2007), while others evidenced a strong individual variability in the response (Bricelj *et al.*, 2011). Histo-

pathological studies on mussels exposed to *A. fundyense* and *P. minimum* detected an inflammatory response consisting of degranulation and diapedesis of hemocytes into the alimentary canal and, as the exposure continued, hemocyte migration into the connective tissue surrounding the gonadal follicles (Galimany *et al.*, 2008a; Galimany *et al.*, 2008b).

Also other classes of toxins can trigger immune responses: experiments by Mello and collaborators (2010) showed the variation of various hemocyte parameters in different bivalve species affected by natural DSP blooms. Also in this case, a species-specific response was observed, demonstrating the presence of vulnerable (e.g. *Perna perna*) and unaffected (e.g. *Crassostrea gigas*) species. Moreover, also GYMs can induce the alteration of hemocyte parameters in *Ruditapes philippinarum* (da Silva *et al.*, 2008), while YTX, PLTX and OA can increase the phagocytic activity of *Mytilus galloprovincialis* hemocytes (Malagoli *et al.*, 2008; Malagoli *et al.*, 2007).

Valve activity and filtration rate

The closure of the shells or the reduction of the filtration rate are well known mechanisms used by bivalves for isolating themselves from the external environment, in presence of negative conditions, such as pollutants or toxins (Gainey and Shumway, 1988). This mechanism has been extensively used to study the effects of phycotoxins on mussel physiology, even though sometimes passive valve closure can occur in response to PSTs, which notoriously cause adverse effects on some species, such as burrowing incapacitation (Bricelj *et al.*, 2010). Different species display different valve activity modulation in response to paralytic dinoflagellates (Lassus *et al.*, 1999; Shumway and Cucci, 1987), so the reduction of the filtration rate is a key parameter used for the classification of bivalves as susceptible or resistant to PSP (Bricelj and Shumway, 1998).

Other negative effects on valve closure and filtration activity have been reported in various bivalve species in response to microalgae producing different toxin classes, such as *Gymnodinium mikimotoi*, *Heterocapsa circularisquama*, *Pseudo-nitzschia* and *Azadinium spinosum* (Basti *et al.*, 2009; Jauffrais *et al.*, 2012; Matsuyama, 1999; Thessen *et al.*, 2010).

Feeding and excretion

Pseudofaeces production is a particularly important pre-ingestive mechanism preventing the animal's ingestive capacity from being exceeded, but it also facilitates the process of particle selection, whereby less nutritious particles can be rejected and the quality of the ingested material can be improved proportionally (Newell and Jordan, 1983). The most complete comparative study concerning the feeding behavior of bivalves on toxic dinoflagellates is provided by Hégaret and colleagues (2007). Clearance rates of five species of bivalve mollusks were assessed, following the exposure for one hour to three harmful-algal strains: *Prorocentrum minimum* (PSP and DSP), *A. fundyense* (PSP), and *Heterosigma akashiwo* (NSP). The analysis of faeces and pseudofaeces revealed species-specific responses to the different harmful algae, indicating in most cases a preferential retention of harmful cells. The production of faeces and pseudofaeces varied appreciably between the different bivalve/algae pairs.

Effects on juveniles

The understanding of the physiological effects played by toxic algae on bivalve juveniles is very important, especially for farmed species, in order to better understand how aquaculture activity is threatened by the HABs. Li and colleagues (2002) exposed juveniles of *R. philippinarum* and *Perna viridis* to *A. tamarensis* (PSP), measuring the scope for growth (SFG), and the growth rate. SFG was significantly reduced in both clams and mussels while *R. philippinarum* resulted to be the most sensitive to PSP while considering the growth rate only.

Leverone and collaborators (2007) reported the effects on the clearance rate of juvenile bivalves of 4 different species in relation with a *Karenia brevis* exposure (PbTx). Both in a short and long-term exposure *Argopecten irradians* resulted to be the most sensitive species, *C. virginica* the least responsive and *P. viridis* and *Mercenaria mercenaria* displayed intermediated responses.

On the contrary, DA did not have a significant effects on feeding rate and shell valve clatter on the juvenile king scallops *Pecten maximus* that, but registered a negative impact on their growth rate and survival (Liu *et al.*, 2007).

Other effects

Considering the different nature, composition and targets of marine toxins, it is likely that many additional effects, not described above, could be detected in different bivalve species. As an example, beside paralysis, PSTs can also produce increased mantel melanization and abnormal vitellogenesis in *Nodipecten subnodosus* and *Argopecten ventricosus* (Escobedo-Lozano *et al.*, 2012; Estrada *et al.*, 2007; Estrada *et al.*, 2010), production of white mucus and inhibition of byssus production in *M. edulis* and *G. demissa* (Shumway *et al.*, 1987).

Moreover, Landsberg (1996) hypothesized a possible connection between neoplasia in bivalves with the presence of micro-algal blooms, even though specific surveys on the topic are still lacking. Some phycotoxins such as OA (Florez-Barros *et al.*, 2011) and DA (Dizer *et al.*, 2001), certainly have a genotoxic effect on bivalves. Furthermore several classes of toxins (PTX, DSTs, PLTX) are known to be potent dysregulators of cytoskeleton dynamics in vertebrates and are likely to exert a similar action also in bivalves (Silvestre and Tosti, 2010)

Studies at molecular level

A summary of the main studies focused on the molecular responses of bivalves to phycotoxins is reported in **Table 2** and discussed in detail below.

Toxin metabolism and biotransformation

Many species of bivalve mollusks are capable of biochemical transformation of the toxins accumulated by filtration, thus generating novel metabolites not found in the causative algal species, suggesting that extremely complex mechanisms of selective accumulation and chemical or enzymatic conversion might be involved in the development of shellfish toxicity (Asakawa *et al.*, 2006). While, in some cases, this could be interpreted as a strategy specifically developed to decrease toxin potency, in other cases these toxin derivatives are likely just the by-products of normal metabolic pathways.

The metabolism of many classes of phycotoxins has been documented in a large number of bivalve species, even though comparative studies demonstrated significant differences in the ability of biotransformation between species (Choi *et al.*, 2003; Li *et al.*, 2012; Sullivan *et al.*, 1983), pointing out that different organisms might have adopted specific biochemical mechanisms as a defensive strategy to recurrent HAB events.

Among the many bioconverting bivalves, the scallop *Patinopecten yessoensis* is definitely one of the most active, since it is able to metabolize YTX, DTX, PTX and OA (Suzuki *et al.*, 2005; Suzuki *et al.*, 1998; Suzuki *et al.*, 1999). The mussels *M. galloprovincialis* and *Mytilus edulis* can metabolize

AZA (McCarron *et al.*, 2009; O'Driscoll *et al.*, 2011), OA (Rossignoli *et al.*, 2011; Torgersen *et al.*, 2008) and, to some extent, also PSTs (Dell'Aversano *et al.*, 2008). Episodes of PbTx conversion to less toxic analogues have been documented in many different bivalves, including the oyster *Crassostrea virginica* (Plakas *et al.*, 2002; Plakas *et al.*, 2004).

Although toxin biotransformations in bivalves are well documented, their biological significance is often unclear, since the modifications do not always result in a decrease of toxin potency. One of the few exceptions is represented by the esterification of OA and DTX1, which has been studied in a number of different bivalves (Rossignoli *et al.*, 2011; Torgersen *et al.*, 2008; Vale and De M. Sampayo, 2002), which is thought to play a role in the sequestration in lipid rich tissues (Svensson and Förlin, 2004) and the conjugation to lipoproteins (Rossignoli and Blanco, 2010).

While some of the toxin biotransformations are likely the effect of passive processes or enzymatic activities provided by symbiotic bacteria (Donovan *et al.*, 2008; Smith *et al.*, 2001), in many cases they have been shown to be active processes catalyzed by bivalves themselves. Nevertheless, despite the overwhelming evidence of phycotoxin transformation, just a few enzymes specifically involved in these processes have been isolated and described so far.

An enzyme capable of hydrolyzing PTX and OA has been recently discovered in the digestive gland of the mussel *Perna canaliculus* (MacKenzie *et al.*, 2012): the protein identified, an acidic serine esterase, did not show any similarity with known sequences and was active on a rather broad range of PTXs and OA esters. Furthermore, two enzymes involved in different PST modifications have been isolated in *Peronidia venulosa* and *Macra chinensis* (Cho *et al.*, 2008; Lin *et al.*, 2004), even though only partial amino-acidic sequences have been characterized: carbamoylase I can hydrolyze the carbamoyl moiety in both carbamoyl and N-carbamoyl PSTs, whereas sulfocarbamoylase catalyzes the hydrolysis of the N-sulfocarbamoyl moiety of the weakly toxic C-PSTs.

The digestive gland is the main site of accumulation of different toxin species, so this tissue is assumed to be the most active in the bioconversion processes (Blanco *et al.*, 2007; Jauffrais *et al.*, 2012; Mafra *et al.*, 2010; Medhioub *et al.*, 2012). Therefore, not surprisingly, all the enzymes involved in toxin transformation so far identified have been isolated from this tissue and novel, yet unknown, enzymes are expected to be highly expressed and active in this tissue in response to HABs.

Toxin binding and accumulation

Despite the important role of digestive gland, in some cases other tissues may play a role in the accumulation of phycotoxins. In fact, some bivalve species can retain toxins for an extremely prolonged time in specific non-visceral tissues, exploiting toxin accumulation as a chemical defense towards natural predators.

The better known phenomenon of chemical defense by phycotoxin accumulation is certainly the butter clam *Saxidomus giganteus* (Smolowitz and Doucette, 1995), which selectively stores PSTs in its siphon epithelium (Kvitek and Beiter, 1991), thus discouraging predation by siphon-nipping fishes (Kvitek, 1991). As a matter of fact, the occurrence of HABs significantly influences the feeding behavior of many vertebrate predators, such as sea otters and shorebirds (Kvitek and Bretz, 2005; Kvitek *et al.*, 1991). Although proteins involved in toxin-binding, transport and accumulation have been described in many organisms, so far the molecular mechanisms leading to this selective retention in bivalves are almost completely unknown. When tested for saxiphilin-like activity, both *M. edulis* and *S. giganteus*, capable of accumulating very high levels of PSTs in their tissues, resulted to be negative, likewise *Spisula solidissima*, *Donax deltoides* and *Vepricardium multispinosum* (Llewellyn *et al.*, 1997). Nevertheless, a case of a PSP-binding protein has been reported in the Moroccan cockle *Acanthocardia tuberculata*; in this species, a 181 KDa protein named PSPBP contributes to the prolonged retention of PSTs in the foot (Takati *et al.*, 2007).

Beside PSTs, other classes of toxins can be conjugated to specific proteins produced by bivalves. Rossignoli and Blanco provided the first lines of evidence of a soluble cytoplasmatic component binding OA in mussel, which was identified as a high density lipoprotein (Rossignoli and Blanco, 2010). Nzoughet and colleagues were able to identify a 45 KDa protein weakly binding AZA in the hepatopancreas of the blue mussel *M. edulis* and another unknown 22 KDa protein which was apparently highly expressed only in contaminated mussels (Nzoughet *et al.*, 2008).

Even though only a little is known about bivalves phycotoxin-binding proteins, these few studies suggest that they may be involved in selective retention or detoxification, depending on whether the toxins will be used in the frame of a chemical defense strategy or simply metabolized.

Metabolic activities

Although relatively few bivalve proteins related to HABs have been identified so far, the alteration of a number of enzymatic activities, possibly related to both xenobiotic metabolism and stress response has been reported. Some classes of phycotoxins can affect the overall metabolic rates of bivalve organs in a time- and doses-dependent way (Haberkorn *et al.*, 2010; Louzao *et al.*, 2010).

A multi biomarker approach by Gorbi and colleagues clearly showed that the activity of the Na⁺/K⁺-ATPase, the target of PLTX, was strongly inhibited by *Ostreopsis ovata* blooms in *M. galloprovincialis*, which can therefore be considered as susceptible organisms. Consequently to this reduction, a significant alteration of other enzymatic activities was also observed (acetylcholinesterase in particular), while on the contrary no enzymatic activities typical of ROS (Reactive Oxygen Species) production were significantly altered, indicating that oxidative pathways are not involved in *O. ovata* toxicity (Gorbi *et al.*, 2012). The cholinesterase activity was also monitored in *M. edulis* after intramuscular injection of DA, highlighting just a moderate reduction after 48h (Dizer *et al.*, 2001).

Other studies have been focused on the analysis of the effects of different toxins on the oxidative pathways; the monitoring of GST, GPx and SOD activities and of lipid peroxidation in the clam *R. philippinarum* contaminated with PSTs highlighted only a modest involvement of these enzymes (Choi *et al.*, 2006). In a different species (the giant lions-paw scallop *Nodipecten subnodosus*) the same toxin class produced a significant alteration of both GPx (up-regulated) and SOD (down-regulated) in gills, whereas no changes were observed in other tissues. Moreover, CAT activity and lipid peroxidation were not markedly altered in any tissue (Estrada *et al.*, 2007; Estrada *et al.*, 2010). On the other hand, other studies observed a stronger correlation between phycotoxins and oxidative damage: in fact, cyanobacteria were able to produce a significant alteration of the activity of the two antioxidant enzymes GST and GPx in *M. galloprovincialis* (Puerto *et al.*, 2011) and of CAT in *M. edulis* (Kankaanpää *et al.*, 2007). Moreover, Haberkorn and colleagues monitored ROS production in PSP-contaminated oysters, observing a linear correlation between PST accumulation and ROS production in haemocytes (Haberkorn *et al.*, 2010).

As we have reported in the previous sections, toxic microalgae can significantly alter the feeding of bivalves, with consequent effects on enzymatic activities linked to digestion. Fernandez-Reiriz and colleagues carefully examined the effects of a diet based on the toxic dinoflagellate *A. catenella* on *Mytilus chilensis*, showing that mussels are able to adapt mechanisms which allow the feeding with toxic algae. The authors were in fact able to observe a logarithmic relationship for the assimilatory balance and the carbohydrate metabolism of the digestive gland, by monitoring the enzymatic activity

of amylase, laminarinase and cellulase (Fernández-Reiriz *et al.*, 2008). The amylase activity of oyster digestive gland is also affected by the exposure to *A. minutum*, even though the changes observed are also largely dependent on the physiological status of bivalves (Haberhorn *et al.*, 2010).

Finally, the monitoring of several hydrolytic enzymes revealed marked alterations in different tissues of *N. subnodosus* exposed to PSP. These enzymatic activities could be potentially used as molecular markers of PSP contamination in scallops; in particular lipase, α -galactosidase, α -glucosidase and α -mannosidase were altered in the digestive gland and trypsin and α -chymotrypsin in the gill, but other modifications were also observed in other tissues, including mantle and the adductor muscle (Estrada *et al.*, 2007).

Biomarkers of contamination: proteomic approaches

Although there is no doubt that large-scale proteomic approaches can provide a tool of the utmost importance for the identification of molecular markers of aquatic pollution (Campos *et al.*, 2012), to date just a very few studies have exploited this potential to explore the effects of phycotoxins on bivalves.

So far proteomic approaches have been mainly used to study the expression of proteins regulated by known molecular targets of toxins, such as in the case of protein phosphatases, the upstream effectors of the p38 mitogen-activated protein kinase, selectively inhibited by OA. An increase in the phosphorylation/activation of p38 MAPK has been in fact observed in the heart of OA treated oysters (Talarmin *et al.*, 2008) and contributes to the increased phagocytotic activity in the immunocytes of PLTX-treated mussels (Malagoli *et al.*, 2008).

One of the very few large-scale approaches concerned the identification of biomarkers of AZA contamination in *Mytilus edulis*; four proteins highly expressed in the digestive gland of toxic mussels were identified, namely cathepsin D, superoxide dismutase, glutathione S-transferase Pi and a flagellar protein of bacterial origin. This data seem to suggest the activation of xenobiotic defense response in bivalves following AZP blooms (Nzoughet *et al.*, 2009).

The negative effects of a toxic cyanobacteria, *Cylindrospermopsis raciborskii*, have been investigated using a similar approach on *M. galloprovincialis* (Puerto *et al.*, 2011). The expression of several structural proteins was remarkably altered, indicating a situation of stress and cytoskeletal destabilization. At the same time, other important proteins such as the mitochondrial HSP60, the major extrapallial fluid protein and a triosephosphate isomerase were significantly down-regulated in the toxic strain-fed mussels, highlighting a complex response of mussel to cyanotoxins.

Although the efforts put so far in large-scale proteomic studies have been very scarce, the possibility to identify potential biomarkers of contamination demonstrates that this represents a valid approach for better understanding the molecular mechanisms involved in shellfish toxicity.

Biomarkers of contamination: target genes focused approaches

The number of proteomic studies exploring the effects of phycotoxins on marine bivalves is quite narrow, but even less efforts have been put in genomic researches. The few studies focused on gene expression have been mostly aimed at the monitoring of a limited number of target sequences. Mello and colleagues (Mello *et al.*, 2012) selected a few genes and assessed their expression in *C. gigas* hemocytes subject to *in vitro* PbTx exposure. The early activation of HSP70, CYP365A1 and FABP, genes related to stress and detoxification, suggest that oyster hemocytes activate a defense response which protects them from cytotoxic damage, which does not involve immune and antioxidant processes, as the expression of BPI, IL-17, EcSOD, Prx6, GPx and SOD was not altered.

Other studies based on gene expression, exploiting the knowledge gathered from previous experiments, have only been planned but not performed yet: an interesting OA biomonitoring approach based on the evaluation of the expression of genes critically important in OA-induced genotoxic damage has been in fact proposed (González-Romero *et al.*, 2012). In particular, the authors claim that the expression of several histone variants, such as the histone H2A.Z, is strongly down-regulated in response to harmful levels of OA, basing their hypothesis on preliminary data which will be published in an upcoming manuscript.

Biomarkers of contamination: whole-transcriptome scale approaches

To the best of our knowledge, only two studies have so far tried to tackle the issue of marine toxin effect on bivalves from a genomic perspective. The two studies, both performed on the Mediterranean mussel *M. galloprovincialis*, investigated the effects of toxins on the gene expression of the main tissue of accumulation, the digestive gland, by using two different techniques.

In the first study, focused on DSP, Manfrin *et al.* (2010) assessed the effects of OA accumulation over 35 days of exposure by cDNA microarray, identifying several transcripts candidates as OA-stress markers. Although most of the sequences could not be linked to known metabolic pathways correlated to biotransformation, the up-regulation of several stress-related proteins, mainly linked to

apoptosis, proteolysis and cytoskeleton destabilization, denoted a possible sufferance of OA exposed mussels.

The preliminary observations collected from this experiment were further validated on the digestive gland of mussels subject to naturally occurring HABs (unpublished data). The expression of 14 selected up-regulated genes was monitored by real-time PCR in samples collected during 2 DSP events occurred in the Gulf of Trieste. The analysis permitted to confirm 11 out of 14 genes as OA-responsive, highlighting that the results of experimental contaminations can be applied also on naturally occurring DSP events (**Fig. 2**).

The second and more recent study exploited the recent advances offered by the application of next generation sequencing technologies to RNA-sequencing. The study was aimed at the investigation of the possible molecular mechanisms activated or repressed in the digestive gland in response to the accumulation of PSTs (Gerdol *et al.*, 2012) produced by the toxigenic dinoflagellate *A. minutum* strain AL9T over a time course of 5 days. The contamination with these toxins apparently only led to negligible effects on gene expression in mussel, which is an organism insensible to the paralytic effects of STX-like toxins, due to the resistance of its nerve voltage-gated sodium channel (Twarog *et al.*, 1972). Therefore, the RNA-seq experiment results seem to disprove the sporadic reports of negative effects of paralytic HABs in mussel (Galimany *et al.*, 2008a; Shumway *et al.*, 1987) and support their classification as organisms refractory to PSP (Bricelj and Shumway, 1998).

Although preliminary, the two above mentioned studies were certainly able to give a better overview about the possible molecular effects of marine phycotoxins on bivalve mollusks and represent the first steps in moving the focus of dinoflagellate toxicity from a human-centered to a bivalve-centered point of view.

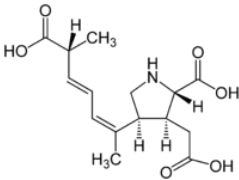
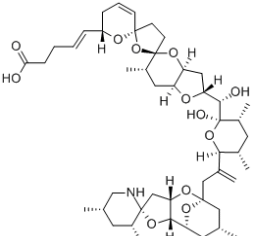
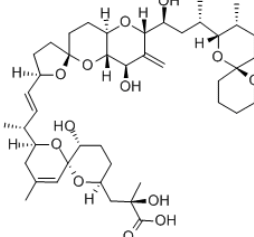
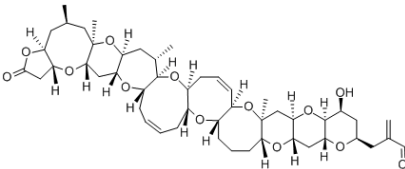
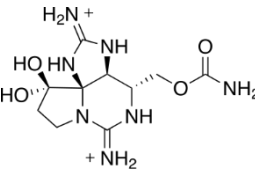
There is no doubt that, thanks to the technological advancements and the increasing availability of bivalve sequence data, including fully sequence genomes (Zhang *et al.*, 2012), the design of similar studies performed on large panels of toxins will be much easier in the near future.

Conclusion

HABs represent a serious and emerging issue for human health, so considerable research efforts have been put in the study of their toxicological effects on vertebrate model organisms, useful for a better comprehension of the dynamics of shellfish poisoning in human. Comparatively, only limited data are available about the effects on bivalves, and the large majority of them concerns physiological aspects. While the lack of molecular studies is mainly caused by the still limited genetic knowledge of these organisms, there is an urge for further research in this field, as highlighted by the promising findings provided by the handful of molecular studies documented in this review.

Certainly, there is the need for combining large scale approaches (both on a proteomic and on a genomic level) for the identification of trustworthy biomarkers of contamination for a more effective biomonitoring of HABs, also contributing to a better comprehension of the molecular mechanisms adapted by bivalve mollusks to deal with phycotoxin toxicity.

Tables

Toxic syndrome	Toxins	Molecular structure	Toxic Algae	Symptoms in human
ASP – Amnesic Shellfish Poisoning	Domoic acid (DA) and analogues		<i>Pseudo-nitzschia</i> spp.	gastrointestinal disorders, headache, disorientation, short-term memory loss, brain damage, death in severe cases (Todd et al., 1993)
AZP – Azaspiracid Shellfish Poisoning	Azaspiracids (AZA) and analogues		<i>Proto-peridinium crassipes</i> , <i>Azadinium spinosum</i> .	diarrhea, nausea, vomiting, abdominal cramps (Satake et al., 1998).
DSP – Diarrhetic Shellfish Poisoning	Okadaic acid (OA) and Dinophysistoxins (DTXs)		<i>Dinophysis</i> spp., <i>Prorocentrum</i> spp.	diarrhea, nausea, vomiting, abdominal cramps (Hallegraeff, 1995).
NSP – Neurotoxic Shellfish Poisoning	Brevetoxins (PbTx)		<i>Karenia brevis</i> (formerly <i>Gymnodinium breve</i> and <i>Ptychodiscus brevis</i>)	nausea, loss of motor control, muscular ache, asthma (Wang, 2008).
PSP – Paralytic Shellfish Poisoning	Saxitoxin (STX) and analogues (GTXs)		<i>Alexandrium</i> spp., <i>Gymnodinium</i> spp.	nausea, diarrhea, abdominal ache, shortness of breath, dry mouth, confused speech, tingling lips, tongue, neck, face, arms, legs and toes (Clark et al., 1999).

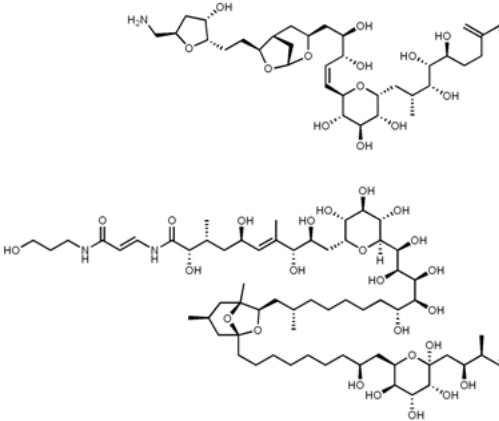
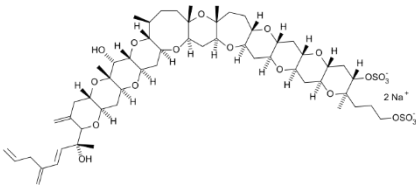
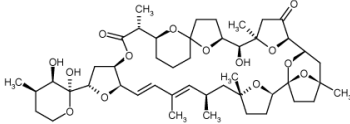
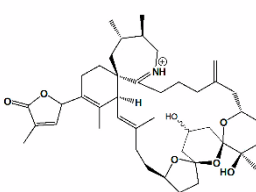
CFP – Ciguatera Fish Poisoning	Only palytoxin (PLTX) and analogues in bivalves, other toxins are responsible of CFP in fishes (ciguatoxin, maitotoxin, etc.)		<i>Ostreopsis</i> spp. (PLTX, ovatoxin and analogues)	lethargy, muscle spasms, tremor myalgia, cyanosis, respiratory distress, rhabdomyolysis, death in severe cases (Ramos and Vasconcelos, 2010)
Others	Yessotoxins (YTXs)		<i>Lingulodinium polyedrum</i> , <i>Gonyaulax spinifera</i> , <i>Prorocentrum reticulatum</i> .	diarrhea (Tubaro et al., 2010)
Others	Pectenotoxins (PTXs)		<i>Dinophysis</i> spp.	diarrhea (Draisci et al., 1996)
Others	Spirolides (SPXs) and Gymnodimine (GYMs)		<i>Karenia selliformis</i> (GYMs) <i>Alexandrium ostenfeldii</i> (SPXs)	neurological symptoms (Munday et al., 2004)

Table 1: Main shellfish poisoning syndromes: summary of the toxins involved, causative algal species and symptoms in human.

Study	Toxin class	Bivalve species	Strategy
Llewellyn et al., 1997	PSTs (STX)	<i>Mytilus edulis</i> , <i>Saximodus giganteus</i> , <i>Spisula solidissima</i> , <i>Donax deltoides</i> and <i>Vepricardium multispinosum</i>	screening for saxiphin-like activity
Dizer et al., 2001	DA	<i>Mytilus edulis</i>	Monitoring of the cholinesterase activity following intramuscular DA injection
Lin et al., 2004	PSTs	<i>Macra chinensis</i>	enzyme purification from the digestive gland (sulfocarbamoylase I)
Choi et al., 2006	PSTs	<i>Ruditapes philippinarum</i>	Monitoring of oxydative stress related enzymatic activities
Estrada et al., 2007	PSTs	<i>Nodipecten subnodosus</i>	Monitoring of antioxidant and hydrolitic enzymes in different tissues
Takati et al., 2007	PSTs	<i>Acanthocardia tuberculata</i>	protein purification from the foot (PSPBP)
Nzoughet et al., 2007	AZA	<i>Mytilus edulis</i>	isoelectric focusing and SEC
Fernandez-Reiriz et al., 2008	PSTs	<i>Mytilus chilensis</i>	Monitoring of the activity of amylase, laminarinase and cellulase
Talarmin et al., 2008	DSTs (OA)	<i>Crassostrea gigas</i>	Western blot
Malagoli et al., 2008	PLT	<i>Mytilus galloprovincialis</i>	Immunoblot
Cho et al., 2008	PSTs	<i>Peronidia venulosa</i>	enzyme purification from the digestive gland (sulfocarbamoylase I)
Nzoughet et al., 2009	AZA	<i>Mytilus edulis</i>	SDS-page and de novo sequencing; comparative analysis between control and contaminated bivalves
Haberkorn et al., 2010	PSTs	<i>Crassostrea gigas</i>	Monitoring of prophenoxidase, amylase activity and ROS production in contaminated oyster hemocytes
Estrada et al., 2010	PSTs	<i>Nodipecten subnodosus</i>	Monitoring of antioxidant and hydrolitic enzymes in different tissues, with a particular emphasis on hemocytes
Manfrin et al., 2010	DSP (OA)	<i>Mytilus galloprovincialis</i>	microarray and real-time PCR on digestive glands of experimentally and naturally contaminated mussels
Rossignoli and Blanco, 2010	DSTs (OA)	<i>Mytilus galloprovincialis</i>	fractionation and enzymatic digestion of cellular fractions
MacKenzie et al., 2012	PTX and OA	<i>Perna canaliculus</i>	enzyme purification from the digestive gland
Gorbi et al., 2012	PLT and ovatoxin	<i>Mytilus galloprovincialis</i>	Multibiomarker approach, assessment of several physiological parameters, including enzymatic activities
Mello et al., 2012	brevetoxin	<i>Crassostrea gigas</i>	real-time PCR on selected genes following heamocytes exposure to brevetoxin
Gerdol et al., 2012	PSP	<i>Mytilus galloprovincialis</i>	RNA-seq on digestive glands of experimentally contaminated mussels
Gonzales et al., in preparation	DSTs (OA)	<i>Mytilus galloprovincialis</i>	monitoring of histones gene expression levels

Table 2: Main molecular studies related to shellfish poisoning in bivalve mollusks.

Figures

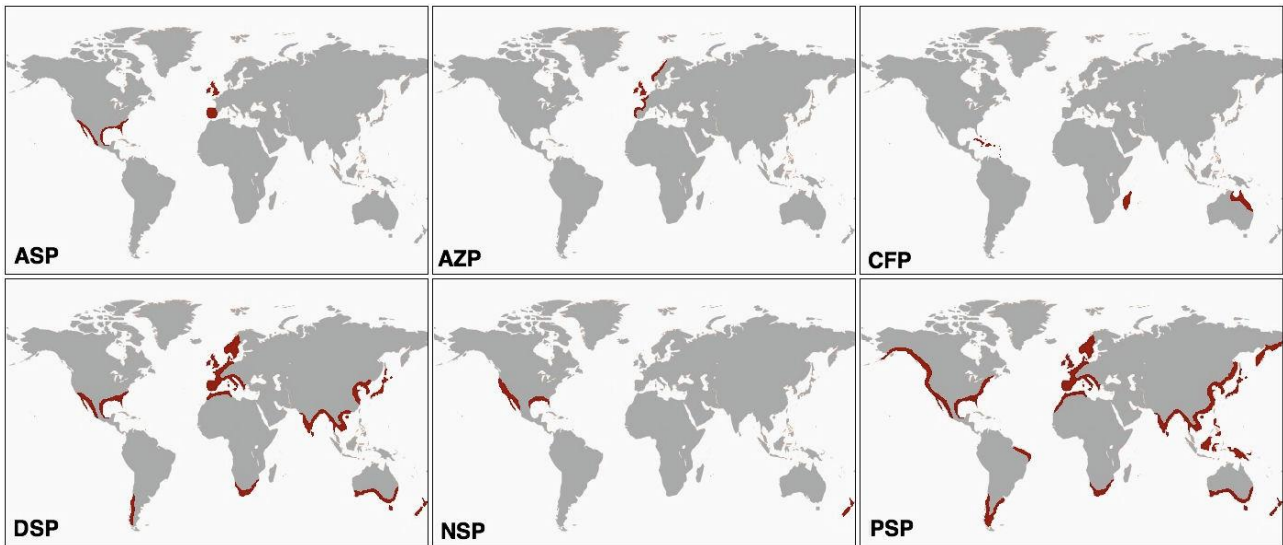


Fig. 1: Geographical occurrence of shellfish poisoning syndromes. ASP: Amnesic Shellfish Poisoning; AZP: Azaspiracid Shellfish Poisoning; CFP: Ciguatera Shellfish Poisoning; DSP: Diarrhetic Shellfish Poisoning; NSP: Neurotoxic Shellfish Poisoning; PSP: Paralytic Shellfish Poisoning

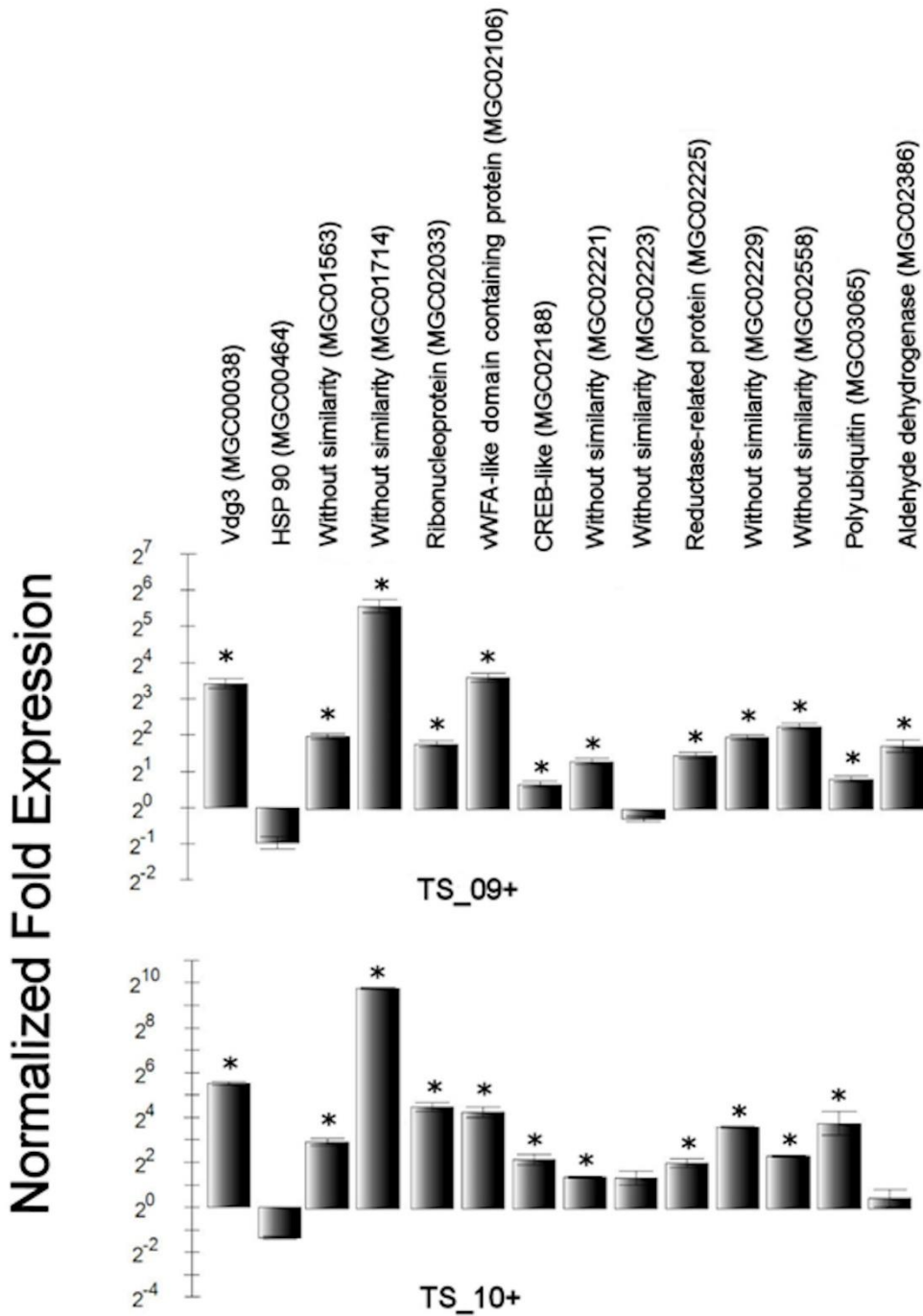


Fig. 2: Validation of potential DSP-biomarkers in mussels digestive gland by quantitative RT-PCR. The expression of fourteen genes identified as differentially expressed in Manfrin et al., 2010 were monitored in two samples collected during naturally occurring DSP-HAB in the Gulf of Trieste (TS_09+ and TS_10+). The normalized fold expression values are shown in a log₂ scale, significant upregulation is indicated by *.

References

- Andrinolo, D., Michea, L. and Lagos, N. Toxic effects, pharmacokinetics and clearance of saxitoxin, a component of paralytic shellfish poison (PSP), in cats. *Toxicon*. 37: 447-464, 1999.
- Asakawa, M., Beppu, R., Ito, K., Tsubota, M., Takayama, H. and Miyazawa, K. Accumulation of paralytic shellfish poison (PSP) and biotransformation of its components in oysters *Crassostrea gigas* fed with the toxic dinoflagellate *Alexandrium tamarense*. *Journal of the Food Hygienic Society of Japan*. 47: 28-32, 2006.
- Barsanti, L. and Gualtieri, P. (2006) *Algae: Anatomy, Biochemistry and Biotechnology*. Taylor and Francis, New York.
- Basti, L., Nagai, K., Shimasaki, Y., Oshima, Y., Honjo, T. and Segawa, S. Effects of the toxic dinoflagellate *Heterocapsa circularisquama* on the valve movement behaviour of the Manila clam *Ruditapes philippinarum*. *Aquaculture*. 291: 41-47, 2009.
- Bianchi, C., Fato, R., Angelin, A., Trombetti, F., Ventrella, V., Borgatti, A.R., *et al.* Yessotoxin, a shellfish biotoxin, is a potent inducer of the permeability transition in isolated mitochondria and intact cells. *Biochim. Biophys. Acta*. 139–147, 2004.
- Blanco, J., Mariño, C., Martín, H. and Acosta, C.P. Anatomical distribution of diarrhetic shellfish poisoning (DSP) toxins in the mussel *Mytilus galloprovincialis*. *Toxicon*. 50: 1011-1018, 2007.
- Bricelj, V.M., Ford, S.E., Lambert, C., Barbou, A. and Paillard, C. Effects of toxic *Alexandrium tamarense* on behavior, hemocyte responses and development of brown ring disease in Manila clams. *Marine Ecology Progress Series*. 430: 35-48, 2011.
- Bricelj, V.M., MacQuarrie, S.P., Doane, J.A.E. and Connell, L.B. Evidence of selection for resistance to paralytic shellfish toxins during the early life history of soft-shell clam (*Mya arenaria*) populations. *Limnology and Oceanography*. 55: 2463-2475, 2010.
- Bricelj, V.M. and Shumway, S.E. Paralytic shellfish toxins in bivalve molluscs: Occurrence, transfer kinetics, and biotransformation. *Reviews in Fisheries Science*. 6: 315-383, 1998.
- Campos, A., Tedesco, S., Vasconcelos, V. and Cristobal, S. Proteomic research in bivalves. Towards the identification of molecular markers of aquatic pollution. *Journal of Proteomics*. 75: 4346-4359, 2012.
- Carlton, J.T. and Geller, J.B. Ecological Roulette: The global transport of nonindigenous marine organisms. *Science*. 261: 78-82, 1993.
- Cho, Y., Ogawa, N., Takahashi, M., Lin, H.-P. and Oshima, Y. Purification and characterization of paralytic shellfish toxin-transforming enzyme, sulfocarbamoylase I, from the Japanese bivalve

Peronidia venulosa. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. 1784: 1277-1285, 2008.

Choi, M.C., Hsieh, D.P.H., Lam, P.K.S. and Wang, W.X. Field depuration and biotransformation of paralytic shellfish toxins in scallop *Chlamys nobilis* and green-lipped mussel *Perna viridis*. *Marine Biology*. 143: 927-934, 2003.

Choi, N.M.C., Yeung, L.W.Y., Siu, W.H.L., So, I.M.K., Jack, R.W., Hsieh, D.P.H., *et al.* Relationships between tissue concentrations of paralytic shellfish toxins and antioxidative responses of clams, *Ruditapes philippinarum*. *Marine Pollution Bulletin*. 52: 572-578, 2006.

Clark, R.F., Williams, S.R., Nordt, S.P. and Manoguerra, A.S. A review of selected seafood poisonings. *Undersea and Hyperbaric Medicine*. 26: 175-184, 1999.

da Silva, P.M., Hégarret, H., Lambert, C., Wikfors, G.H., Le Goïc, N., Shumway, S.E., *et al.* Immunological responses of the Manila clam (*Ruditapes philippinarum*) with varying parasite (*Perkinsus olseni*) burden, during a long-term exposure to the harmful alga, *Karenia selliformis*, and possible interactions. *Toxicon*. 51: 563-573, 2008.

Dell'Aversano, C., Walter, J.A., Burton, I.W., Stirling, D.J., Fattorusso, E. and Quilliam, M.A. Isolation and Structure Elucidation of New and Unusual Saxitoxin Analogues from Mussels. *Journal of Natural Products*. 71: 1518-1523, 2008.

Dizer, H., Fischer, B., Harabawy, A.S.A., Hennion, M.C. and Hansen, P.D. Toxicity of domoic acid in the marine mussel *Mytilus edulis*. *Aquatic Toxicology*. 55: 149-156, 2001.

Donovan, C.J., Ku, J.C., Quilliam, M.A. and Gill, T.A. Bacterial degradation of paralytic shellfish toxins. *Toxicon*. 52: 91-100, 2008.

Draisci, R., Lucentini, L., Giannetti, L., Boria, P. and Poletti, R. First report of pectenotoxin-2 (PTX-2) in algae (*Dinophysis fortii*) related to seafood poisoning in Europe. *Toxicon*. 34: 923-935, 1996.

Escobedo-Lozano, A.Y., Estrada, N., Ascencio, F., Gerardo Contreras, G. and Alonso-Rodriguez, R. Accumulation, Biotransformation, Histopathology and Paralysis in the Pacific Calico Scallop *Argopecten ventricosus* by the Paralyzing Toxins of the Dinoflagellate *Gymnodinium catenatum*. *Mar. Drugs*. 1044-1065, 2012.

Estrada, N., de Jesús Romero, M., Campa-Córdova, A., Luna, A. and Ascencio, F. Effects of the toxic dinoflagellate, *Gymnodinium catenatum* on hydrolytic and antioxidant enzymes, in tissues of the giant lions-paw scallop *Nodipecten subnodosus*. *Comparative Biochemistry and Physiology - C Toxicology and Pharmacology*. 146: 502-510, 2007.

Estrada, N., Rodríguez-Jaramillo, C., Contreras, G. and Ascencio, F. Effects of induced paralysis on hemocytes and tissues of the giant lions-paw scallop by paralyzing shellfish poison. *Marine Biology*. 157: 1401-1415, 2010.

Fernández-Reiriz, M.J., Navarro, J.M., Contreras, A.M. and Labarta, U. Trophic interactions between the toxic dinoflagellate *Alexandrium catenella* and *Mytilus chilensis*: Feeding and digestive behaviour to long-term exposure. *Aquatic Toxicology*. 87: 245-251, 2008.

Ferrão-Filho, A.D.S. and Kozłowsky-Suzuki, B. Cyanotoxins: Bioaccumulation and effects on aquatic animals. *Marine drugs*. 9: 2729-2772, 2011.

Florez-Barros, F., Prado-Alvarez, M., Mendez, J. and Fernandez-Tajes, J. Evaluation of genotoxicity in gills and hemolymph of clam *ruditapes decussatus* fed with the toxic dinoflagellate *prorocentrum lima*. *Journal of Toxicology and Environmental Health - Part A: Current Issues*. 74: 971-979, 2011.

Franchini, A., Malagoli, D. and Ottaviani, E. Targets and Effects of Yessotoxin, Okadaic Acid and Palytoxin: A Differential Review. *Marine drugs*. 658-677, 2010.

Funari, E. and Testai, E. Human health risk assessment related to cyanotoxins exposure. *Critical Reviews in Toxicology*. 38: 97-125, 2008.

Gainey, L.F. and Shumway, S.E. A compendium of the responses of bivalve molluscs to toxic dinoflagellates. *Journal of Shellfish Research*. 7: 623-628, 1988.

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

Galimany, E., Sunila, I., Hégaret, H., Ramón, M. and Wikfors, G.H. Pathology and immune response of the blue mussel (*Mytilus edulis* L.) after an exposure to the harmful dinoflagellate *Prorocentrum minimum*. *Harmful Algae*. 7: 630-638, 2008b.

Gerdol, M., De Moro, G., Manfrin, C., Milandri, A., Riccardi, E., Beran, A., *et al.* RNA sequencing and de novo assembly of the digestive gland transcriptome in *Mytilus galloprovincialis* fed with toxinogenic and non-toxic strains of *Alexandrium minutum*. *Marine biotechnology*. Submitted, 2012.

González-Romero, R., Rivera-Casas, C., Fernández-Tajes, J., Ausió, J., Méndez, J. and Eirín-López, J.M. Chromatin specialization in bivalve molluscs: A leap forward for the evaluation of Okadaic Acid genotoxicity in the marine environment. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*. 155: 175-181, 2012.

Gorbi, S., Bocchetti, R., Binelli, A., Bacchiocchi, S., Orletti, R., Nanetti, L., *et al.* Biological effects of palytoxin-like compounds from *Ostreopsis cf. ovata*: A multibiomarkers approach with mussels *Mytilus galloprovincialis*. *Chemosphere*. 89: 623-632, 2012.

Haberkorn, H., Lambert, C., Le Goïc, N., Guéguen, M., Moal, J., Palacios, E., *et al.* Effects of *Alexandrium minutum* exposure upon physiological and hematological variables of diploid and triploid oysters, *Crassostrea gigas*. *Aquatic Toxicology*. 97: 96-108, 2010.

- Hallegraeff, G.M. A review of harmful algal blooms and their apparent global increase. *Phycologia*. 32: 79-99, 1993.
- Hégaret, H. and Wikfors, G.H. Effects of natural and field-simulated blooms of the dinoflagellate *Prorocentrum minimum* upon hemocytes of eastern oysters, *Crassostrea virginica*, from two different populations. *Harmful Algae*. 4: 201-209, 2005a.
- Hégaret, H. and Wikfors, G.H. Time-dependent changes in hemocytes of eastern oysters, *Crassostrea virginica*, and northern bay scallops, *Argopecten irradians irradians*, exposed to a cultured strain of *Prorocentrum minimum*. *Harmful Algae*. 4: 187-199, 2005b.
- Hégaret, H., Wikfors, G.H. and Shumway, S.E. Diverse feeding responses of five species of bivalve mollusc when exposed to three species of harmful algae. *J. Shellfish Res.* 549–559, 2007.
- Hégaret, H., Wikfors, G.H., Soudant, P., Lambert, C., Shumway, S.E., Bérard, J.B., *et al.* Toxic dinoflagellates (*Alexandrium fundyense* and *A. catenella*) have minimal apparent effects on oyster hemocytes. *Mar. Biol.* . 441–447, 2007.
- Jauffrais, T., Contreras, A., Herrenknecht, C., Truquet, P., Séchet, V., Tillmann, U., *et al.* Effect of *Azadinium spinosum* on the feeding behaviour and azaspiracid accumulation of *Mytilus edulis*. *Aquatic toxicology*. 124: 179-187, 2012.
- Jauffrais, T., Marcaillou, C., Herrenknecht, C., Truquet, P., Séchet, V., Nicolau, E., *et al.* Azaspiracid accumulation, detoxification and biotransformation in blue mussels (*Mytilus edulis*) experimentally fed *Azadinium spinosum*. *Toxicon*. 60: 582-595, 2012.
- Kankaanpää, H., Leiniö, S., Olin, M., Sjövall, O., Meriluoto, J. and Lehtonen, K.K. Accumulation and depuration of cyanobacterial toxin nodularin and biomarker responses in the mussel *Mytilus edulis*. *Chemosphere*. 68: 1210-1217, 2007.
- Kvitek, R. and Bretz, C. Shorebird foraging behavior, diet, and abundance vary with harmful algal bloom toxin concentrations in invertebrate prey. *Marine Ecology Progress Series*. 293: 303-309, 2005.
- Kvitek, R.G. Paralytic shellfish toxins sequestered by bivalves as a defense against siphon-nipping fish. *Marine Biology*. 111: 369-374, 1991.
- Kvitek, R.G. and Beiter, M.K. Relative insensitivity of butter clam neurons to saxitoxin: a pre-adaptation for sequestering paralytic shellfish poisoning toxins as a chemical defense. *Marine Ecology Progress Series*. 69: 47-54, 1991.
- Kvitek, R.G., Degange, A.R. and Beiter, M.K. Paralytic shellfish poisoning toxins mediate feeding behavior of sea otters. *Limnology & Oceanography*. 36: 393-404, 1991.
- Landsberg, J.H. Neoplasia and biotoxins in bivalves: is there a connection? . *Journal of Shellfish Research*. 15: 203-230, 1996.

Lassus, P., Bardouil, M., Beliaeff, B., Masselin, P., Naviner, M. and Truquet, P. Effect of a continuous supply of the toxic dinoflagellate *Alexandrium minutum* Halim on the feeding behavior of the Pacific oyster (*Crassostrea gigas* Thunberg). *Journal of Shellfish Research*. 18: 211-216, 1999.

Le Hégarat, L., Jacquin, A., Bazin, E. and Fessard, V. Genotoxicity of the marine toxin okadaic acid, in human Caco-2 cells and in mice gut cells. *Environmental Toxicology*. 21: 55-64, 2006.

Leverone, J., Shumway, S. and Blake, N. Comparative effects of the toxic dinoflagellate *Karenia brevis* on clearance rates in juveniles of four bivalve molluscs from Florida, USA. *Toxicon*. 49: 634-645, 2007.

Lewis, R.J. and Holmes, M.J. Origin and transfer of toxins involved in ciguatera. *Comparative Biochemistry and Physiology - C Pharmacology Toxicology and Endocrinology*. 106: 615-628, 1993.

Li, A., Ma, J., Cao, J., Wang, Q., Yu, R., Thomas, K., *et al.* Analysis of paralytic shellfish toxins and their metabolites in shellfish from the North Yellow Sea of China. *Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment*. 29: 1455-1464, 2012.

Li, S., Wang, W. and Hsieh, D. Effects of toxic dinoflagellate *Alexandrium tamarense* on the energy budgets and growth of two marine bivalves. *Mar Environ Res*. 53: 145-160, 2002.

Lin, H.-P., Cho, Y., Yashiro, H., Yamada, T. and Oshima, Y. Purification and characterization of paralytic shellfish toxin transforming enzyme from *Mactra chinensis*. *Toxicon*. 44: 657-668, 2004.

Liu, H., Kelly, M., Campbell, D., Dong, S., Zhu, J. and Wang, S. Exposure to domoic acid affects larval development of king scallop *Pecten maximus* (Linnaeus, 1758). *Aquatic Toxicology*. 81: 152-158, 2007.

Llewellyn, L.E., Bell, P.M. and Moczydlowski, E.G. Phylogenetic survey of soluble saxitoxin-binding activity in pursuit of the function and molecular evolution of saxiphilin, a relative of transferrin. *Proceedings of the Royal Society B: Biological Sciences*. 264: 891-902, 1997.

Loker, E.S., Adema, C.M., Zhang, S.M. and Kepler, T.B. Invertebrate immune systems - Not homogeneous, not simple, not well understood. *Immunological Reviews*. 198: 10-24, 2004.

Louzao, M.C., Espiña, B., Cagide, E., Ares, I.R., Alfonso, A., Vieytes, M.R., *et al.* Cytotoxic effect of palytoxin on mussel. *Toxicon*. 56: 842-847, 2010.

MacKenzie, L.A., Selwood, A.I. and Marshall, C. Isolation and characterization of an enzyme from the Greenshell™ mussel *Perna canaliculus* that hydrolyses pectenotoxins and esters of okadaic acid. *Toxicon*. 60: 406-419, 2012.

Mafra, L.L., Bricelj, V.M. and Fennel, K. Domoic acid uptake and elimination kinetics in oysters and mussels in relation to body size and anatomical distribution of toxin. *Aquatic Toxicology*. 100: 17-29, 2010.

Malagoli, D., Casarini, L. and Ottaviani, E. Effects of the marine toxins okadaic acid and palytoxin on mussel phagocytosis. *Fish and Shellfish Immunology*. 24: 180-186, 2008.

Malagoli, D., Casarini, L., Sacchi, S. and Ottaviani, E. Stress and immune response in the mussel *Mytilus galloprovincialis*. *Fish and Shellfish Immunology*. 23: 171-177, 2007.

Manfrin, C., Dreos, R., Battistella, S., Beran, A., Gerdol, M., Varotto, L., *et al.* Mediterranean mussel gene expression profile induced by okadaic acid exposure. *Environmental Science and Technology*. 44: 8276-8283, 2010.

Matsuyama, Y. Harmful effect of dinoflagellate *Heterocapsa circularisquama* on shellfish aquaculture in Japan. *Japan Agricultural Research Quarterly*. 33: 283-293, 1999.

McCarron, P., Kilcoyne, J., Miles, C.O. and Hess, P. Formation of azaspiracids-3, -4, -6, and -9 via decarboxylation of carboxyazaspiracid metabolites from shellfish. *Journal of Agricultural and Food Chemistry*. 57: 160-169, 2009.

Medhioub, W., Lassus, P., Truquet, P., Bardouil, M., Amzil, Z., Sechet, V., *et al.* Spirolide uptake and detoxification by *Crassostrea gigas* exposed to the toxic dinoflagellate *Alexandrium ostenfeldii*. *Aquaculture*. 358-359: 108-115, 2012.

Mello, D.F., De Oliveira, E.S., Vieira, R.C., Simoes, E., Trevisan, R., Dafre, A.L., *et al.* Cellular and transcriptional responses of *Crassostrea gigas* hemocytes exposed in vitro to brevetoxin (PbTx-2). *Marine Drugs*. 10: 583-597, 2012.

Mello, D.F., Proença, L.A.O. and Barracco, M.A. Comparative study of various immune parameters in three bivalve species during a natural bloom of *Dinophysis acuminata* in Santa Catarina Island, Brazil. *Toxins*. 2: 1166-1178, 2010.

Miles, C.O., Wilkins, A.L., Stirling, D.J. and Mackenzie, A.L. Gymnodimine C, an isomer of gymnodimine B, from *Karenia selliformis*. *Journal of Agricultural and Food Chemistry*. 51: 4838-4840, 2003.

Munday, R., Towers, N.R., Mackenzie, L., Beuzenberg, V., Holland, P.T. and Miles, C.O. Acute toxicity of gymnodimine to mice. *Toxicon*. 44: 173-178, 2004.

Newell, R.I.E. and Jordan, S.J. Preferential ingestion of organic material by the American oyster *Crassostrea virginica*. *Mar. Ecol. Prog. Ser.* 47-53, 1983.

Nzoughet, J.K., Hamilton, J.T.G., Botting, C.H., Douglas, A., Devine, L., Nelson, J., *et al.* Proteomics identification of azaspiracid toxin biomarkers in blue mussels, *Mytilus edulis*. *Molecular and Cellular Proteomics*. 8: 1811-1822, 2009.

Nzoughet, K.J., Hamilton, J.T.G., Floyd, S.D., Douglas, A., Nelson, J., Devine, L., *et al.* Azaspiracid: First evidence of protein binding in shellfish. *Toxicon*. 51: 1255-1263, 2008.

- O'Driscoll, D., Škrabáková, Z., O'Halloran, J., van Pelt, F.N.A.M. and James, K.J. Mussels Increase Xenobiotic (Azaspiracid) Toxicity Using a Unique Bioconversion Mechanism. *Environmental Science & Technology*. 45: 3102-3108, 2011.
- Perez, S., Vale, C., Botana, A., Alonso, E., Vieytes, M. and Botana, L. Determination of toxicity equivalent factors for paralytic shellfish toxins by electrophysiological measurements in cultured neurons. *Chemical research in toxicology*. 24: 1153-1157, 2011.
- Plakas, S.M., El Said, K.R., Jester, E.L.E., Ray Granade, H., Musser, S.M. and Dickey, R.W. Confirmation of brevetoxin metabolism in the Eastern oyster (*Crassostrea virginica*) by controlled exposures to pure toxins and to *Karenia brevis* cultures. *Toxicon*. 40: 721-729, 2002.
- Plakas, S.M., Wang, Z., El Said, K.R., Jester, E.L.E., Granade, H.R., Flewelling, L., *et al.* Brevetoxin metabolism and elimination in the Eastern oyster (*Crassostrea virginica*) after controlled exposures to *Karenia brevis*. *Toxicon*. 44: 677-685, 2004.
- Puerto, M., Campos, A., Prieto, A., Cameán, A., Almeida, A.M.D., Coelho, A.V., *et al.* Differential protein expression in two bivalve species; *Mytilus galloprovincialis* and *Corbicula fluminea*; exposed to *Cylindrospermopsis raciborskii* cells. *Aquatic Toxicology*. 101: 109-116, 2011.
- Ramos, V. and Vasconcelos, V. Palytoxin and analogs: Biological and ecological effects. *Marine drugs*. 8: 2021-2037, 2010.
- Reigman, R. Species composition of harmful algal blooms in relation to macronutrient dynamics. In Anderson D. M., Cembella A. D. and M., H.G. (eds), *Physiological Ecology of HAB*, Springer-Verlag, New York, pp. 475-488, 1998
- Richardson, K. (1997) Harmful or exceptional phytoplankton blooms in the marine ecosystem. pp. 367-385.
- Roach, J.S., LeBlanc, P., Lewis, N.I., Munday, R., Quilliam, M.A. and MacKinnon, S.L. Characterization of a dispiroketal spirolide subclass from *Alexandrium ostenfeldii*. *Journal of Natural Products*. 72: 1237-1240, 2009.
- Rossignoli, A.E. and Blanco, J. Subcellular distribution of okadaic acid in the digestive gland of *Mytilus galloprovincialis*: First evidences of lipoprotein binding to okadaic acid. *Toxicon*. 55: 221-226, 2010.
- Rossignoli, A.E., Fernández, D., Regueiro, J., Mariño, C. and Blanco, J. Esterification of okadaic acid in the mussel *Mytilus galloprovincialis*. *Toxicon*. 57: 712-720, 2011.
- Ryan, J., Morey, J., Bottein, M., Ramsdell, J. and Van Dolah, F. Gene expression profiling in brain of mice exposed to the marine neurotoxin ciguatoxin reveals an acute anti-inflammatory, neuroprotective response. *BMC Neuroscience*. 107, 2010.

Sala, G.L., Bellocci, M. and Rossini, G.P. The cytotoxic pathway triggered by palytoxin involves a change in the cellular pool of stress response proteins. *Chemical Research in Toxicology*. 22: 2009-2016, 2009.

Sala, G.L., Ronzitti, G., Sasaki, M., Fuwa, H., Yasumoto, T., Bigiani, A., *et al.* Proteomic analysis reveals multiple patterns of response in cells exposed to a toxin mixture. *Chemical Research in Toxicology*. 22: 1077-1085, 2009.

Sasso, S., Pohnert, G., Lohr, M., Mittag, M. and Hertweck, C. Microalgae in the postgenomic era: A blooming reservoir for new natural products. *FEMS Microbiology Reviews*. 36: 761-785, 2012.

Satake, M., Ofuji, K., Naoki, I., James, K.J., Furey, A., McMahon, T., *et al.* Azaspiracid, a new marine toxin having unique spiro ring assemblies, isolated from Irish mussels, *Mytilus edulis*. *Journal of the American Chemical Society*. 120: 9967-9968, 1998.

Shumway, S.E. and Cucci, T.L. The effect of the toxic dinoflagellate *Protogonyaulax tamarensis* on the feeding behavior of bivalve molluscs. *Aquat. Toxicol.* 9-27, 1987.

Shumway, S.E., Pierce, F.C. and 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*. 87: 1021-1023, 1987.

Silvestre, F. and Tosti, E. Impact of marine drugs on cytoskeleton-mediated reproductive events. *Marine drugs*. 8: 881-915, 2010.

Smith, E.A., Grant, F., Ferguson, C.M.J. and Gallacher, S. Biotransformations of Paralytic Shellfish Toxins by Bacteria Isolated from Bivalve Molluscs. *Applied and Environmental Microbiology*. 67: 2345-2353, 2001.

Smolowitz, R. and Doucette, G. Immunohistochemical localization of saxitoxin in the siphon epithelium of the butter clam, *Saxidomus giganteus*. *The Biological bulletin*. 189: 229-230, 1995.

Sobel, J. and Painter, J. Illnesses caused by marine biotoxins. *Clinical Infectious Diseases*. 41: 1290-1296, 2005.

Sullivan, J.J., Iwaoka, W.T. and Liston, J. Enzymatic transformation of PSP toxins in the littleneck clam (*Protothacastaminea*). *Biochemical and Biophysical Research Communications*. 114: 465-472, 1983.

Suzuki, T., Igarashi, T., Ichimi, K., Watai, M., Suzuki, M., Ogiso, E., *et al.* Kinetics of diarrhetic shellfish poisoning toxins, okadaic acid, dinophysistoxin-1, pectenotoxin-6 and yessotoxin in scallops &Patinopecten yessoensis&/i>. *Fisheries Science*. 71: 948-955, 2005.

Suzuki, T., Mitsuya, T., Matsubara, H. and Yamasaki, M. Determination of pectenotoxin-2 after solid-phase extraction from seawater and from the dinoflagellate *Dinophysis fortii* by liquid

chromatography with electrospray mass spectrometry and ultraviolet detection Evidence of oxidation of pectenotoxin-2 to pectenotoxin-6 in scallops. *Journal of Chromatography A*. 815: 155-160, 1998.

Suzuki, T., Ota, H. and Yamasaki, M. Direct evidence of transformation of dinophysistoxin-1 to 7-O-acyl-dinophysistoxin-1 (dinophysistoxin-3) in the scallop *Patinopecten yessoensis*. *Toxicon*. 37: 187-198, 1999.

Svensson, S. and Förlin, L. Analysis of the importance of lipid breakdown for elimination of okadaic acid (diarrhetic shellfish toxin) in mussels, *Mytilus edulis*: results from a field study and a laboratory experiment. *Aquatic Toxicology*. 66: 405-418, 2004.

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

Talarmin, H., Droguet, M., Penneç, J.P., Schröder, H.C., Muller, W.E.G., Gioux, M., *et al.* Effects of a phycotoxin, okadaic acid, on oyster heart cell survival. *Toxicological and Environmental Chemistry*. 90: 153-168, 2008.

Thessen, A.E., Soniat, T.M., Dortch, Q. and Doucette, G.J. *Crassostrea virginica* grazing on toxic and non-toxic diatoms. *Toxicon*. 55: 570-579, 2010.

Todd, E.C.D. Domoic acid and Amnesic Shellfish Poisoning- a review. *Journal of Food Protection*. 69-83, 1993.

Torgersen, T., Sandvik, M., Lundve, B. and Lindegarth, S. Profiles and levels of fatty acid esters of okadaic acid group toxins and pectenotoxins during toxin depuration. Part II: Blue mussels (*Mytilus edulis*) and flat oyster (*Ostrea edulis*). *Toxicon*. 52: 418-427, 2008.

Tubaro, A., Dell'Ovo, V., Sosa, S. and Florio, C. Yessotoxins: A toxicological overview. *Toxicon*. 56: 163-172, 2010.

Twarog, B.M., Hidaka, T. and Yamaguchi, H. Resistance to tetrodotoxin and saxitoxin in nerves of bivalve molluscs: A possible correlation with paralytic shellfish poisoning. *Toxicon*. 10: 273-278, 1972.

Twiner, M., Hanagriff, J., Butler, S., Madhkoor, A. and GJ, D. Induction of apoptosis pathways in several cell lines following exposure to the marine algal toxin azaspiracid. *Chemical research in toxicology*. 25: 1493-1501, 2012.

Valdiglesias, V., Méndez, J., Pàsaro, E., Cemeli, E., Anderson, D. and Laffon, B. Assessment of okadaic acid effects on cytotoxicity, DNA damage and DNA repair in human cells. *Mutation Research* 74-79, 2010.

Vale, P. and De M. Sampayo, M.A. Esterification of DSP toxins by Portuguese bivalves from the Northwest coast determined by LC-MS - A widespread phenomenon. *Toxicon*. 40: 33-42, 2002.

- Van Dolah, F.M. Marine algal toxins: Origins, health effects, and their increased occurrence. *Environmental Health Perspectives*. 108: 133-141, 2000.
- Venier, P., De Pittà, C., Bernante, F., Varotto, L., De Nardi, B., Bovo, G., *et al.* MytiBase: A knowledgebase of mussel (*M. galloprovincialis*) transcribed sequences. *BMC Genomics*. 10, 2009.
- Venier, P., Varotto, L., Rosani, U., Millino, C., Celegato, B., Bernante, F., *et al.* Insights into the innate immunity of the Mediterranean mussel *Mytilus galloprovincialis*. *BMC Genomics*. 12, 2011.
- Vilariño, N. Marine toxins and the cytoskeleton: azaspiracids. *FEBS Journal*. 275: 6075-6081, 2008.
- Wang, D.Z. Neurotoxins from marine dinoflagellates: A brief review. *Marine drugs*. 6: 349-371, 2008.
- Wang, J., Wang, Y.Y., Lin, L., Gao, Y., Hong, H.S. and Wang, D.Z. Quantitative proteomic analysis of okadaic acid treated mouse small intestines reveals differentially expressed proteins involved in diarrhetic shellfish poisoning. *Journal of Proteomics*. 75: 2038-2052, 2012.
- Yasumoto, T., Murata, M. and Oshima, Y. Diarrhetic shellfish toxins. *Tetrahedron*. 41: 1019-1025, 1985.
- Yasumoto, T., Oshima, Y. and Yamaguchi, M. Occurrence of a New Type of Shellfish Poisoning in the Tokohu District. *Bull. Jap. Soc. Sci. Fish.* 44: 1249-1255, 1978.
- Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*. advance online publication, 2012.

Latimeria menadoensis:

I dati trascrittomici in questo organismo, ottenuti da studi di RNA-seq eseguiti con tecnologia Illumina, provengono da campioni di fegato e testicoli e sono il frutto di una collaborazione tra l'Università di Trieste, Ancona e Viterbo.

Questo lavoro ci ha inoltre offerto l'opportunità di collaborare anche con il Broad Institute e di entrare a far parte del progetto di sequenziamento del genoma di *Latimeria chalumnae*.

Data l'eccezionalità dell'organismo, il nostro è stato il quinto esemplare mai pescato di questa specie, i dati trascrittomici sono derivati da un unico individuo.

Lo studio, inoltre, a differenza degli altri, è stato esplicitamente pensato per creare il trascrittoma di questo organismo i cui campioni disponibili sono così rari.

Il sequenziamento dei due tessuti di *L. menadoensis* ha permesso di ottenere 145.435.156 *reads paired-end*.

Oltre ad essere state rimosse le sequenze originate dall'RNA ribosomale, anche in questo caso le sequenze sono state filtrate per eliminare adattatori e basi a bassa qualità. Il set di sequenze è stato così ridotto a un totale di 88.872.414 *reads*.

Il Broad Institute, grazie alla collaborazione in corso, ci ha offerto la possibilità di assemblare le nostre *reads* con Trinity, software di assemblaggio da loro sviluppato, permettendoci così di elaborare dati provenienti da diversi metodi di assemblaggio.

La combinazione di differenti metodi e algoritmi di assemblaggio può essere considerata come la strategia migliore per l'ottenimento di trascritti di alta qualità.

Poiché i *contig* ottenibili con Trinity presentano spesso un'elevata ridondanza, a causa della tendenza del software di creare *contig* diversi per ogni *splicing* alternativo, abbiamo cercato di limitare questa ridondanza utilizzando l'assembler MIRA e il software CLC Genomic Workbench.

Innanzitutto le sequenze sono state assemblate con Trinity creando, così, 306.882 *contig* che sono stati successivamente utilizzati come sequenze di input per il programma di assemblaggio MIRA.

Questo programma si basa sull'uso interattivo di strategie *multipass*, utilizzando regioni ad alta similarità e strategie *fallback* per usare regioni a bassa similarità.

Questo metodo di assemblaggio nonostante richieda molta più RAM e tempi più lunghi rispetto a programmi di assemblaggio che utilizzano l'algoritmo *de Bruijn*, permette di assemblare tra loro sequenze che altri *assembler* non riescono ad allineare.

Le sequenze ottenute sono state filtrate per lunghezza, mantenendo un *cut-off* di 250 pb, riducendo l'assemblaggio a 105.653 trascritti, riducendo così del 19.21% la ridondanza nei *contig* creati da Trinity.

Contemporaneamente le *reads* sono state assemblate utilizzando il CLC Genomic Workbench, generando 149.339 *contig*.

All'interno delle sequenze contigue create sono state ricercate le *open reading frame* (ORF) e solamente quelle che presentavano almeno una ORF di almeno 70 codoni sono stati mantenute.

I trascritti derivanti dai diversi assemblaggi sono stati quindi allineati tramite BLASTn utilizzando parametri molto restrittivi.

Mediamente Trinity utilizza per l'assemblaggio un maggior numero di *reads* rispetto al CLC Genomic Workbench. Per questo motivo, nel caso in cui fossero presenti sequenze molto simili tra quelle create dalla coppia Trinity/MIRA e CLC Genomic Workbench, si è preferito scartare i *contig* creati da quest'ultimo. Sono stati, infine, mantenuti i *contig* generati da CLC solo nel caso in cui non ci fosse similarità di sequenza oppure, nel caso ci fosse una similarità significativa, questi fossero di almeno 200pb più lunghi.

Anche in questo caso, l'opera di filtraggio e selezione è stata fatta utilizzando uno *script* in Python creato appositamente. I *contig* risultanti sono stati filtrati per lunghezza e i trascritti con bassa copertura sono stati scartati ottenendo un set di 66.308 sequenze di alta qualità.

Anche in questo caso è stata effettuata l'analisi per il calcolo dell' Ortholog Hit Ratio utilizzando il BLASTx contro NR. Le successive fasi di annotazione sono state effettuate utilizzando il software di annotazione BLAST2GO (Conesa et al., 2005) che ha permesso di annotare le sequenze con Gene Ontology, BLASTx ed Interpro.

Un'ulteriore analisi è stata fatta per ricercare eventuali elementi trasponibili utilizzando il software Repeatmasker. (<http://www.repeatmasker.org>)

Avendo a disposizione sequenze provenienti da due tessuti diversi è stata effettuata un'analisi di RNA-seq utilizzando le metodiche descritte precedentemente per *M. galloprovincialis* per identificare i trascritti maggiormente espressi nei due tessuti.

I *contig* creati sono stati inoltre confrontati con i trascritti ottenuti dal Broad Institute dall'assemblaggio di sequenze trascrittomiche di muscolo di *L. chalumnae*.

Il muscolo è un tessuto molto specializzato che esprime un numero di geni molto minore rispetto ai nostri tessuti ed infatti il 50% dell'espressione genica totale in muscolo è data da solo 12 geni.

Dal confronto è emerso come i tessuti da noi analizzati esprimano un range di trascritti molto maggiore, e ha permesso di valutare quanto divergente sia il contributo nell'espressione genica dei tre tessuti.

Facendo parte del progetto di sequenziamento del genoma abbiamo potuto anche mappare, a livello nucleotidico, le *reads* di *L. menadoensis* a nostra disposizione all'interno delle regioni codificanti del genoma annotato da Ensembl di *L. chalumnae*.

Questa analisi è stata svolta utilizzando il programma CLC Genomic Workbench.

Questo ha permesso a noi di avere delle statistiche sulla profondità di sequenziamento e a loro di testare la qualità delle annotazioni.

I risultati hanno dimostrato che la profondità dei dati di RNA-seq di *liver* e *testis* a nostra disposizione può essere considerata un fondamentale strumento per l'identificazione di nuovi geni e in particolare dei trascritti non codificati non annotati.

Inoltre, grazie a questo confronto, è stato possibile stimare a 99,73% la similarità tra *L. menadoensis* e *L. chalumnae*.

Per avere una stima della divergenza evolutiva tra le due specie, inoltre, abbiamo selezionato un set di 25 geni ortologi altamente conservati, con identità di sequenza superiore all'80%, le cui sequenze fossero disponibili per *L. menadoensis*, *L. chalumnae*, *Takifugu rubripes* e *Tetraodon nigroviridis*.

Il tasso di sostituzione nelle due specie di Latimeria è risultato essere pari a 0.49/100pb mentre è risultato essere circa 16 volte più alta nella coppia Takifugu/Tetraodon (8,25/100).

Poiché il tempo stimato di divergenza tra Tetraodon e Takifugu, basato su evidenze paleolitiche, è tra i 32,25 e i 56 milioni di anni (Benton and Donoghue, 2007), abbiamo ipotizzato che la datazione della divergenza tra il celacanto africano e indonesiano potrebbe essere stimata tra 1,9 e 3,3 milioni di anni.

Questo lavoro ha portato alla scrittura di diverse pubblicazioni derivanti direttamente dal lavoro di analisi e assemblaggio dei dati trascrittomici e ha posto le basi per eventuali lavori futuri.

Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*.

Alberto Pallavicini, Adriana Canapa, Marco Barucca, Jessica Alfoldi, Maria Assunta Biscotti, Francesco Buonocore, Gianluca De Moro, Federica Di Palma⁵, Anna Maria Fausto, Mariko Forconi, Marco Gerdol, Daisy Monica Makapedua, Jason Turner-Meier, Ettore Olmo, Giuseppe Scapigliati.

Keywords: coelacanth, *Latimeria menadoensis*, transcriptome, *de novo* assembly, RNA-seq

Abstract

Latimeria menadoensis is a coelacanth species first identified in 1997 in Indonesia, at 10,000 Km of distance from its African congener. In the present work we describe the *de novo* transcriptome assembly obtained from liver and testis samples collected from the fifth specimen ever caught of this species.

The deep RNA sequencing performed with Illumina technologies generated 145,435,156 paired-end reads, accounting for ~14 GB of sequence data, which were *de novo* assembled using a Trinity/CLC combined strategy. The assembly output was processed and filtered producing a set of 66,308 contigs, whose quality was thoroughly assessed. The comparison with the recently sequenced genome of the African congener *Latimeria chalumnae* and with the available genomic resources of other vertebrates revealed a good reconstruction of full length transcripts and a high coverage of the predicted full coelacanth transcriptome.

The RNA-seq analysis revealed remarkable differences in the expression profiles between the two tissues, allowing the identification of liver- and testis-specific transcripts which may play a fundamental role in important biological processes carried out by these two organs.

Given the high genomic affinity between the two coelacanth species, the here described *de novo* transcriptome assembly can be considered an unmatched support tool for the improvement of gene prediction within the genome of *L. chalumnae* and a valuable resource for investigation of many aspects of tetrapod evolution.

Introduction

One of the most important transitions in vertebrate evolution was the arising of terrestrial vertebrates, which entailed considerable morphological changes related to the acquisition of novel functions by pre-existing and, in several cases, pre-adapted structures, like the evolution of lobe fins into tetrapod limbs. The terrestrial vertebrates would have derived from fossil forms of lobe-finned fishes, a highly successful group in the Devonian (400 Mya), with hundred species populating the Gondwana supercontinent's oceans and river systems (Maisey 1996; Cloutier and Ahlberg 1997).

Until 1938 only two sarcopterygian taxa were considered to have survived post-Devonian extinction: the dipnoi (lungfish), with three extant genera, and the tetrapods, with ~23,500 species. Hence the clamour when the first living coelacanth (*Latimeria chalumnae*), a fish considered extinct, was found off the estuary of river Chalumna, in South Africa (Smith 1939; Smith 1989). In 1997 a specimen of another *Latimeria* population was identified by Mark V. Erdmann in a fish market in Manado Tua (Sulawesi, Indonesia) (Erdmann et al. 1999). The distance between the two sites (more than 10,000 Km) and the early molecular findings (Holder et al. 1999; Pouyaud et al. 1999) led to the identification of the latter specimen as belonging to a distinct species, called *L. menadoensis*. Several individuals of *L. chalumnae* have been fished to date, as opposed to only six individuals of *L. menadoensis*. Specimens of Indonesian coelacanth are therefore very rare and constitute a valuable scientific resource and a mine of precious genetic information.

The main molecular and morphological studies of the genus *Latimeria* have addressed the evolutionary relationships linking lungfishes, coelacanths and tetrapods. Their results have however been discordant, since different datasets have sustained different hypotheses (Gorr et al. 1991a; Gorr et al. 1991b; Meyer and Dolven 1992; Yokobori et al. 1994; Meyer 1995; Zardoya and Meyer 1996; Zardoya et al. 1998; Holder et al. 1999; Pouyaud et al. 1999; Tohyama et al. 2000; Venkatesh et al. 2001; Brinkmann et al. 2004; Takezaki et al. 2004; Inoue et al. 2005; Shan and Gras 2011).

L. menadoensis has also been the subject of other molecular investigations aimed at characterizing some genes of evolutionary interest: Hox genes (Koh et al. 2003; Shashikant et al. 2004; Amemiya et al. 2010), ParaHox genes (Mulley and Holland 2010), the Protocadherine cluster (Noonan et al. 2004), the RAG1 and RAG2 genes (Brinkmann et al. 2004), Sonic hedgehog gene and its enhancers (Hadzhiev et al. 2007), visual pigments (Yokoyama and Tada 2000), a Heat Shock Protein 70 (Modisakeng et al. 2009), neurohypophyseal hormones (Gwee et al. 2008), and vitellogenins (Canapa et al. 2012). Furthermore some transposable elements were characterized in this species (Bejerano et al. 2006; Nishihara et al. 2006; Xie et al. 2006; Smith et al. 2012). The importance of *L. menadoensis*

has led to the study of its whole mitochondrial genome (Inoue et al. 2005; Sudarto et al. 2010) as well as to partial sequencing of a BAC library, which has made ~5 Mb of genomic sequences available to databases (Danke et al. 2004). Furthermore, correlations between quantitative and compositional characteristics of the genome of *L. menadoensis* were considered in Makapedua et al. (2011).

Over the last few years next generation sequencing technologies (NGS) have revolutionized the fields of genomics and transcriptomics, providing the opportunity to analyze genomes and transcriptomes with high sequencing depth in a relatively short time in comparison with Sanger sequencing. The molecular data obtained with such technologies, applied to a rising number of organisms, are proving steadily important to study their relationships at the macro- and micro-evolutionary levels. In this respect, having in mind that genes are targets of evolution-driven changes that lead to the different morphology of animals, in the framework of studies of genomic features of coelacanth (Amemiya et al. 2012), we examined the *L. menadoensis* transcriptomes of liver and testis using deep-sequencing techniques.

The liver is one of the most suitable tissues in that it participates more than other organs in a range of physiological processes and contains cell types endowed with distinct roles and functions. Considering the different sex determination mechanisms occurring across vertebrates, the expression in testis is interesting to better understand the genes involved in these processes and their evolution since the systematic position of *Latimeria*. Furthermore, this tissue proved to be particularly suitable and useful for deep RNA-seq, as it expresses a broad range of different transcripts, permitting the assembly of a high number of full length sequences. In fact, about a half of the sequences included in the high quality set of 66,308 contigs was estimated to have been correctly assembled to the full length.

The *de novo* transcriptome assembly was able to significantly enhance the global view of the sequences expressed in coelacanth, overcoming the limitations linked to the automated and conservative, protein coding gene-focused, prediction by Ensembl, and providing a remarkable amount of information concerning expressed sequences produced by non-annotated genes. This knowledge provided significant information not only to investigate important biological processes and metabolic pathways in *Latimeria*, but also to acquire information on the origin of tetrapods and on the possible evolutionary dynamics relative to the genes involved in the transition from aquatic to terrestrial vertebrates.

Results

Latimeria menadoensis transcriptome sequencing

The Illumina sequencing procedure generated 145,435,156 raw nucleotide paired-end reads (76,932,818 and 68,502,338 reads from liver and testis, respectively). The average read length for liver was 97.28 bp, corresponding to a complete dataset of 7.48 GB of sequence data. Deep RNA-seq of testis produced reads slightly shorter, with an average length of 96.22 bp, accounting to 6.59 GB of sequence data.

Following the processing steps involving the trimming of adapters and low quality bases, and the removal of short reads and of reads originated by ribosomal RNA, the two sequence sets were significantly reduced. 47,470,578 and 41,401,836 high quality sequencing reads were selected from liver and testis, respectively. Therefore a total of 88,872,414 sequencing reads were used for the *de novo* assembly. A summary of the trimming step statistics is reported in **Table 1**. A detailed report of quality and statistics for the reads used for the *de novo* transcriptome assembly is presented in **Supplementary file S1**.

De novo assembly

The *de novo* transcriptome assembly performed with Trinity (Grabherr et al. 2011) by using both liver and testis reads generated a total of 306,882 contigs. The filtering step used to select only the longest transcript per gene produced 223,365 contigs, and the additional step applied to remove redundant sequences by MIRA 3.4.0 (Chevreux et al. 2004) and to filter sequences shorter than 250 bp, further reduced the Trinity assembly to a set of 105,653 transcripts.

The *de novo* assembly produced with the CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark) generated 149,339 raw contigs. The high quality subset of protein-coding sequences selected to integrate the Trinity assembly, as described in the materials and methods section, comprised 48,846 sequences. A total of 8,496 CLC contigs were detected by BLASTn as matching existing Trinity contigs and significantly longer than them. The corresponding Trinity contigs were therefore replaced. The remaining 40,350 CLC contigs were discarded, as they could not significantly improve the Trinity assembly.

A total of 105,653 contigs was obtained following the combining of the data generated by the two *de novo* assemblers. Finally, the filtering step applied to remove poorly covered sequences, resulting from the fragmentation of transcripts expressed at particularly low levels, reduced the contig number

to a final high quality set of 66,308 sequences. A detailed graphical summary of the strategy used and of the results obtained by the *de novo* assembly of *L. menadoensis* transcriptome is shown in **Figure 1**.

Assembly quality assessment

The goal of these assembly processing steps was to reduce redundancy without losing any valuable sequence data (**Figure 2**). Despite making use of a large fraction of the original sequencing reads (65.41% of the intact sequence pairs -fragments- could be mapped to the contigs), the raw Trinity assembly was largely redundant. Moreover, the mapping of the reads on the assembled contigs revealed 75% of non-specific matches. On the contrary the raw CLC assembly showed virtually no redundancy (~0.01%) but only 33% of sequenced fragments were used to produce the assembly.

The sequence redundancy was drastically reduced to 19.21% after the removal of Trinity redundant contigs by MIRA. Furthermore, no sequence data were lost, as the total number of reads mapped on the updated assembly slightly increased (+1.19%): this was due to the elongation of 8,496 Trinity contigs by CLC.

Although a large portion of contigs with low expression was discarded (39,342 contigs, accounting for 37.24% of the 105,653 contigs), this did not significantly affect the total number of mapped reads (which only decreased by 0.34%) and contributed to a further reduction of sequence redundancy (which dropped to 17.39%).

The comparison between sequence length categories based on average coverage, before and after the contig filtering step (**Figure 3**), revealed that this procedure was able to sensibly reduce the amount of short sequences, especially those shorter than 500 bp, moving the distribution of contig length towards longer and more reliable sequences.

Transcript fragmentation was assessed with the Ortholog Hit Ratio method (O'Neil et al. 2010), which relies on the comparison between the observed length of contigs and the full length of known ortholog sequences of other species, detected by BLASTx. This method is strongly influenced by inter-species divergence and by the different substitution rates observed among genes and can often lead to an under-estimation of transcript integrity (Ewen-Campen et al. 2011). To overcome this imperfection of the method, we applied a correction, by only considering in the analysis highly conserved genes (characterized by a BLAST identity higher than 90%, independently from the hit length). By these

means, a sufficiently large set of sequences was analyzed (6,024 coelacanth contigs), permitting to obtain a reliable estimate of fragmentation within the high quality liver and testis transcripts.

About a half of the contigs resulted to be assembled to their full length, when compared with known ortholog sequences (**Figure 4**). The mean and median ratios resulted to be 0.72 and 0.86, respectively. Approximately a quarter of the high quality transcript set is expected to be composed by highly fragmented contigs (covering less than 50% of the expected length).

The average length of the contigs obtained, ranging from 250 (the minimum length allowed) to 20,815 bp, was 1,080 bp. The N50 statistic of the assembly was 1,761 and 1,081 contigs longer than 5 Kb were obtained (80 contigs were longer than 10 Kb). A summary of the final assembly statistics is shown in **Table 2**.

Transcript annotation

The annotation performed with BLASTx to the NCBI non-redundant (nr) protein database revealed that 23,564 of the assembled contigs (35.54%) had at least one positive hit. 42,744 contigs did not give any BLAST hit by the cutoff of $1e-6$. The BLAST top hit species distribution is shown in **Figure 5**.

The BLAST2GO annotation, directly performed on the high quality set of transcripts translated into the six possible reading frames, revealed 42,667 out of 66,308 total sequences bearing at least one Interpro domain, accounting for 64.35% of annotated transcripts. The list of the 25 most abundant Interpro domains is displayed in **Table 3**.

The assembled sequences were also annotated with Gene Ontology (GO) terms as described in the materials and methods section, according to the three major GO categories: Cell Component, Molecular Function, and Biological Process. A total of 28,502 transcripts (42.98%) were associated with at least one GO term; concerning the second level of ontology, 6,698 were assigned to a Cell Component category, 13,061 to a Molecular Function category, and 13,030 to a Biological Process category. The summary of Gene Ontology mappings is reported in **Figure 6**.

TE elements in the coelacanth transcriptome

The analysis carried out with RepeatMasker (Smit 1996-2012) to identify transcribed repetitive elements of *L. menadoensis* revealed that 11.17% of the assembled contigs harbors at least one repeat and that 1.87% of sequenced bases matches to a RepeatMasker entry. The major part of matching contigs harbors a transposable element (98.9%): SINEs (79.9%), LINEs (12.0%), LTR retrotransposon families (0.5%), and DNA transposons (6.5%). The types of repetitive elements less represented are small RNAs (0.9%) such as tRNAs, srpRNAs, snRNAs, and 7SK RNAs; Unknown and Satellite elements (0.2%) (**Figure 7**).

Furthermore the Interpro domain analysis on the 66,308 high quality contigs allowed to identify 119 transcripts containing the IPR000477 domain (Reverse transcriptase), 72 contigs with IPR004244 domain (Transposase, L1), and 17 sequences harboring IPR001584 domain (Integrase, catalytic core).

RNA-seq mapping on the African coelacanth genome

Globally, the 61.64% of the sequence data generated by the Illumina paired-end sequencing of liver and testis RNA could map to the genes annotated of *L. chalumnae*. The 93.03% of the counted fragments mapped within exons (51.63% were mapped on exon-exon junctions), whereas just 6.97% of the counted fragments mapped within introns (2.98% mapped on exon-intron junctions). The observed redundancy was very low, highlighted by a match specificity of 98.1%. 17,129 out of the 22,819 annotated gene models were found to have a positive mapping, meaning that the 75.06% of the coelacanth predicted genes were expressed in liver or in testis.

A larger proportion of reads could map to the full assembled genome (85,682,920), revealing that 34.77% of the reads generated by the RNA-seq of liver and testis account for the expression of genes which are still not annotated. Overall, 3,189,494 reads (3.59% out of the total) could not be mapped to the *L. chalumnae* genomic scaffolds. The summary of the RNA-seq data mapping on the African coelacanth genome is presented in **Table 4**.

RNA-seq mapping on *L. menadoensis* transcriptome

The RNA-seq mapping performed to calculate the expression levels of the assembled transcripts in both analyzed organs mapped the majority of paired-end reads to the assembled contigs (**Table 5**). In fact, the percentage of counted fragments was 67.20% in liver and slightly lower in testis, 64.57%.

The fraction of mapping reads was very similar in the two organs, being 78.12% in liver and 76.90% in testis, indicating that slightly more than 20% of the sequence data generated by the NGS sequencing could not be mapped to the final set of assembled contigs. Furthermore, it was possible to estimate the number of fragments which were not used at all by the assembly procedure, by comparing the number of paired-end reads mapping in broken pairs with the number of uncounted fragments. Only about 5.5% out of the total number of fragments produced by sequencing did not show any mapping on the assembled contigs, neither as intact nor as broken pairs (5.47% in liver and 5.38% in testis).

The RNA-seq mapping revealed that a higher number of transcripts were expressed in testis in respect to liver. In fact the expression of 55,975 contigs (84.42%) was found in liver, whereas the expression of 61,633 sequences (92.95%) was detected in testis. The comparison between the two organs highlighted that 51,302 contigs (77.37%) were expressed in both. Nevertheless, the two transcriptomes resulted to be remarkably divergent when comparing expression levels, which for most genes were largely divergent as shown by the expression scatter plot in **Figure 8**.

The list of the 20 most expressed transcripts in liver and testis is reported in **Table 6** and **7** respectively. The majority of the most expressed genes in both samples likely play important tissue-specific functions. With a few exceptions (most notably the elongation factor 1 α and the subunit 6 of ATP synthase F₀, whose expression is important for the correct maintenance of all cell types) the 20 genes characterizing the two tissues show great differences in expression.

The transcriptome richness was further graphically inspected in **Figure 9** comparing *L. menadoensis* liver and testis transcriptomes to the RNA-seq of *L. chalumnae* muscle. A steep curve, reaching quickly the asymptote (corresponding to the 100% of the transcription observed in each tissue), means that a low number of genes are expressed at high levels in that tissue. On the contrary, the later the curve approaches the asymptote, the more genes are expressed, indicating higher transcriptome richness. Among the 3 tissues, muscle is definitely the one characterized by the steepest curve. In fact the 50% of the total gene expression in muscle is given by 12 genes, whereas the 80% is given by 206 genes. These data are consistent with observations previously collected in other organisms (Lanfranchi et al. 1996).

The two tissues used for the deep RNA-seq of *L. menadoensis* were both richer than muscle, although testis resulted to be, by far, the one expressing a broader range of transcripts. In testis 321 genes contribute to the 50% of gene expression, while in liver the same number of genes accounts for almost the 75%. The 1,000 most expressed genes in liver contribute to the 83% of total transcription (88% in muscle), whereas the same number of genes in testis contribute to the 65%.

The overlap between liver, testis and muscle transcriptomes was further investigated by analyzing how many common genes were found among the 1,000 most expressed in each tissue. A schematic representation of transcriptomes overlap is given in the Venn diagram in **Figure 10**. 172 sequences, likely representing housekeeping genes, whose expression at rather elevated levels is important in all tissues, were found in all the 3 sets. In all the three organs analyzed, about 2/3 of the transcripts were identified as tissue specific, highlighting once again the strong link between the biological function of different tissues and gene expression.

Discussion

***De novo* transcriptome assembly**

The advent of NGS technologies has had an outstanding impact on many fields of biology, including genetics (Mardis 2008), functional and comparative genomics (Morozova and Marra 2008; Zhang et al. 2011) and molecular ecology (Ekblom and Galindo 2011). The remarkable potential range of application of these techniques will likely move the focus of high throughput sequencing in the near future from genome and transcriptome sequencing to the use in clinical medicine and diagnostics (Meyerson et al. 2010; Pallen et al. 2010; Majewski et al. 2011). Due to its potential application to deep RNA-seq, NGS has been praised as a cost-effective and revolutionary tool for transcriptomics since the very early stages of its development (Wang et al. 2009).

Although great technical advances have been made in a relatively short lapse of time in the improvement of both sequencing technologies and sequencing data management, significant challenges linked with RNA-seq still remain unsolved. The major computational issues in the management of NGS data is represented by the reliable *de novo* assembly of transcriptomes (Martin and Wang 2011; Cahais et al. 2012). This is a complex task, due to presence of alternatively spliced transcript variants, gene duplications, allelic polymorphisms and noise due to suboptimal sequence quality, which often leads to the generation of a high number of short and poorly assembled contigs (Feldmeyer et al. 2011).

The massive amount of sequencing reads obtained from *L. menadoensis* liver and testis allowed us to apply stringent filtering criteria, both in the processing of raw sequencing reads and in the filtering of assembled contigs, in order to achieve a final set of high quality transcripts and to overcome the most common pitfalls of NGS assemblies.

Many different algorithms for *de novo* assembly have been developed, but so far none of them has conclusively proved to be most effective than the others. The accuracy and speed of the assembly, as well as the ability to detect and efficiently reconstruct alternatively spliced transcripts and to avoid sequence redundancy, are all factors which have to be taken into account while considering the assembly algorithm to be used. As a rule of thumb, the performance of different assemblers may significantly vary depending on the size and the quality of the sequence set to be assembled (Feldmeyer et al. 2011). The combination of different assemblies should be considered as the best strategy to obtain a more credible final product (Kumar and Blaxter 2010).

We chose to use the Trinity assembler, able to efficiently recover full length transcripts across a broad range of expression levels but somewhat redundant because of the inclusion of alternatively spliced variants (Grabherr et al. 2011). The Trinity assembly was used as a reference sequence set to be appropriately refined and enriched, whenever possible, by a second *de novo* assembly performed with the assembler included in the CLC Genomic Workbench. The choice of integrating the Trinity output with the CLC assembly was made because of the empirical observation of a more effective reconstruction of full length transcripts and because of the operational speed of its assembly algorithm, based on de Bruijn graph instead of the Overlap-Layout-Consensus (OLC). As this method, although extremely fast, is known to produce assemblies which are quite fragmented in comparison with other assemblers (Kumar and Blaxter 2010), only a selected set of assembled contigs was used to improve the Trinity assembly, with a particular emphasis on protein-coding transcripts.

***De novo* assembly quality assessment**

One of the problems most commonly arising from the *de novo* assembly of RNA-seq data is represented by sequence fragmentation (Feldmeyer et al. 2011). This issue can be, in the first place, the direct consequence of regions poorly covered by sequencing because of a low expression level or because of an insufficient sequencing depth applied.

In order to minimize this problem, as described in the methods section, all the contigs, whose average coverage resulted to be lower than 5, were removed prior to further analysis, reducing the number of contigs from 105,653 to a final set of 66,308 high quality contigs. This processing step reduced the fraction of short sequences with a proportional enrichment in longer transcripts (**Figure 3**). Furthermore, the contig processing strategy we used, graphically summarized in **Figure 1**, contributed to significantly reduce the sequence redundancy of the assembly, in the final set of contigs (which was calculated to be 17.39%) in respect with the Trinity output (**Figure 2**). Nevertheless,

several factors can negatively influence the outcome of a *de novo* transcriptome assembly, affecting the reconstruction of full length sequences. The presence of highly similar paralog genes resulting from recent gene duplication and the existence of allelic variants, combined with the possible presence of repetitive nucleotide stretches and low quality reads, can result in local mis-assemblies and breaking points in contigs construction. Therefore, a certain amount of fragmented contigs was expected, despite the good quality of the reads generated by Illumina sequencing and the stringent parameters used both in the raw sequencing reads processing and in the assembled contigs filtering.

The ortholog hit ratio analysis highlighted good mean and median ratio values and a high proportion of transcripts assembled to their full length (**Figure 4**). Therefore, despite the inevitable presence of broken transcripts, the results of the *de novo* assembly were extremely satisfying, highlighting that about half of the sequences, contained in the final set of transcripts, was assembled to the full length or very close to it and that just about a quarter of the contigs were resulting from highly fragmented transcripts.

Transcript annotation

The analysis of the top hit species distribution resulting from BLAST (**Figure 5**) reveals *Gallus gallus* as the first species, followed by *Xenopus tropicalis*. The first teleost fish of the list, zebrafish, ranked at the sixth place of the list, after the mammal *Monodelphis domestica*. These results clearly show that organisms, whose genome has been largely and deeply studied and annotated, are ranked quite high in the list, mainly because of the higher quality of genome assemblies, of the more accurate prediction of genes and of the higher number of protein sequences deposited in public sequence databases.

Almost the double number of contigs were annotated by Interproscan (42,667 contigs). Since the presence of Interpro domains is a strong indication of coding sequences, these data point out that 64.35% of the coelacanth *de novo* assembled contigs are coding for proteins characterized by known Interpro domains.

The most abundant Interpro domains (**Table 3**) are all extremely common in metazoans, with IPR000719 (Protein kinase, catalytic domain) being the most abundant one, with 2,041 annotated transcripts, followed by IPR007087 (Zinc finger, C2H2) and IPR002290 (Serine/threonine- / dual-specificity protein kinase, catalytic domain).

Moreover, slightly less than a half of the high quality contigs was assigned to at least one Cell Component, Molecular Function or Biological Process by the Gene Ontology mapping (**Figure 6**). Concerning the cellular localization, the majority of annotated transcripts was assigned to cell (GO:0005623), followed by organelle (GO:0043226) and macromolecular complex (GO:0032991). The largely predominant molecular functions resulted to be binding (GO:0005488) and catalytic activity (GO:0003824). Finally, concerning biological processes, cellular process (GO:0009987) and metabolic process (GO:0008152) were the two GO terms most represented.

TE discussion

In metazoans repeat elements cover a considerable part of genomes. Moreover, the transcriptome analysis allowed the evaluation of the transcriptional activity of transposable elements (TE) which play a key role in gene evolution and genome plasticity. TE are divided in two classes: Class I is composed of Long Terminal Repeat retrotransposons (LTRs) and Non-LTRs (subdivided in LINEs and SINEs); Class II is composed of DNA transposons.

The RepeatMasker analysis revealed that 11.17% of contigs harbors a repeat and the most represented elements belong to SINE families. The latter result is in line with the studies performed in the Indonesian coelacanth genome (Bejerano et al. 2006; Nishihara et al. 2006; Xie et al. 2006), in which the activity of SINE elements in *Latimeria* was inferred. The identification of LF-SINE and DeuSINE in *L. menadoensis* transcriptome might confirm that these elements are actually active. Moreover, since their conservation in higher vertebrates, this movement might predate the common ancestor of crossopterygians, for more than 400 Myr.

On the other hand the occurrence of complete SINEs in contigs bearing protein-coding sequence might reveal the gain of new functional roles (exaptation) (Gould and Vrba 1982), as previously described in tetrapod genomes.

Concerning the activity of LINEs, the second most represented interspersed elements, the Interproscan analysis identified amino acidic domains linked to these autonomous retrotransposons.

Chicken Repeat 1 (CR1) are the most abundant elements among LINEs. In contrast to the *Gallus gallus* genome where these elements are predominant but, with very few exceptions, nonfunctional (Wicker et al. 2005), in *Latimeria* they seem to be active.

Fragmented LTRs and ERVs were identified in the transcriptome. This result is in agreement with the analyses on Foamy-like retroviral elements recently discovered in *L. chalumnae* genome by Han and Worobey (2012) showing many frame-shifts and stop codons.

The abundance of the *Harbinger* DNA transposons in *L. menadoensis* genome (Smith et al. 2012) suggests that Class II elements represent a remarkable fraction of the coelacanth TEs, however our analysis indicates that few DNA elements are expressed. This discordance may be related to the lack of coelacanth specific sequences belonging to this class in the RM database or to their propagation mode.

The identification of mobile elements in transcriptomes sheds light on an unexpected genome dynamicity in an organism considered to be a living fossil even from a molecular point of view.

RNA-seq mapping on the African coelacanth genome

More than half of the sequence data generated by the RNA-seq of *L. menadoensis* liver and testis mapped on the genes annotated by Ensembl on the *L. chalumnae* genome (**Table 4**). This data revealed an overall good annotation of the African coelacanth transcripts, even though in some cases the RNA-seq data produced in this study could provide some evidence of additional exons, given that the 6.97% of the reads corresponded to regions annotated as introns.

Nevertheless, a rather high proportion of reads, close to 40%, could not be mapped on the genes annotated by Ensembl. This sticks with the strategy adopted by Ensembl for the annotation pipeline, which is automated and is mainly focused on protein-coding gene models. In fact, almost the 35% of the sequencing reads could map on the assembled genomic scaffolds outside from the annotated gene boundaries, revealing that a relevant portion of the transcripts expressed in the Indonesian coelacanth liver and testis might correspond to genes which were not annotated by the Ensembl RNA-seq annotation pipeline (**Table 4**). Therefore, the deep RNA-seq of liver and testis can be considered as a fundamental tool for the discovery of novel genes, and in particular, of many not yet annotated non-coding transcripts.

Slightly more than 3 million reads did not map on the genomic scaffolds. These sequence data could either correspond to mitochondrial RNA (which was esteemed to account for 3.03% and 2.08% of the reads in liver and testis, respectively) or to coding genes harbored in *L. chalumnae* genomic regions which were not successfully assembled.

Liver and testis transcriptomes comparison

The expression profile of the two organs analyzed was expected to be quite different, considering the largely different tasks they perform and the highly specialized cellular types involved. This difference was immediately evident by the graphical representation of the expression scatter plot (**Figure 8**).

Among the 20 most expressed transcripts in liver, a large fraction is constituted by plasma proteins, whose synthesis is carried out by liver (such as the three chains constituting fibrinogen, α -2 macroglobulins, apolipoproteins, hemopexin, vitronectin, lipocalin, serum amyloid P and serum albumin), constitute the core of the highly expressed genes in this tissue (**Table 6**).

On the other hand testis invests a significant portion of transcription in genes involved in chromatin and cytoskeletal rearrangements. In particular, a testis-specific histone results to be expressed almost 25 times more than the second most expressed gene, prostaglandin H2D isomerase, and accounts for about 18% of the global testis transcription. A significant amount of the total gene expression is derived from the synthesis of messengers of protamines, used for the replacement of histones and the effective packaging of DNA in the sperm acrosome (Balhorn 2007).

The expression of genes involved in chromatin rearrangement is strictly regulated, as testis-specific histones are transiently and selectively expressed only during specific phases of spermiogenesis (Martianov et al. 2005). In fact, also sperm nuclear basic protein PL-I and histone H1x-like figure among the most representative testis genes. Furthermore a relevant number of other testis-specific genes can be linked to the meiotic process carried out in the testicular germinal cells and to the cytoskeletal rearrangements consequently required (tubulin α chain testis-specific, tubulin β 2-C and centrin-1). Moreover, specific types of microtubules are required for the correct assembly of mitotic and meiotic spindles and of the flagellar axoneme of spermatozoa (Kemphues et al. 1982; Villasante et al. 1986). The tubulin genes specifically expressed in testis are likely linked to these functions.

Although the expression of a large fraction of genes was clearly strictly tissue-specific, thanks to the sequencing depth applied a relevant overlap between the two transcriptomes (77.37%) was observed.

The issue of transcriptome richness was addressed by analyzing the relative contributions of the expression of each contig to the total observed transcription in the two tissues, and in RNA-seq data of *L. chalumnae* muscle (**Figure 9**). Highly specialized tissues are expected to invest the most gene expression in a selected set of genes, thus being transcriptionally poor, whereas tissues involved in many different biological functions, characterized by the coexistence of many different cell types are expected to be transcriptionally rich, as they express a broader range of transcripts. Within this picture, muscle is a classic example of a highly specialized tissue, expressing at particularly high

levels a limited set of genes involved in the structural organization of muscle fibers and responsible of contraction.

Testis expresses a broader range of transcripts, which is in agreement with the assumption that cells in this organ are characterized by drastic morphological and functional changes linked to gamete generation: in this scenario chromatin structure is constantly rearranged and gene expression may therefore be substantially variable during the different stages of spermatogenesis (Tanaka and Baba 2005).

Despite being transcriptionally poorer than testis, the RNA-seq of liver likely brought a remarkable amount of additional data as pointed out by the extent of the overlap between the two transcriptomes (**Figure 10**). Therefore, although the RNA-seq of two different organs like testis and liver was particularly effective to approach the coverage of a complete transcriptome, the scarce overlap observed between the two *L. menadoensis* transcriptomes and the *L. chalumnae* muscle suggests that the sequencing of RNAs obtained from additional samples would be useful in order to effectively describe the complete transcriptome of this organism.

Conclusion

The *de novo* assembly of the Indonesian coelacanth *L. menadoensis* liver and testis transcriptomes here described contains complete information concerning the expressed sequences involved in the important biological processes held by liver and testis tissues, such as metabolism and reproduction (Canapa et al. 2012). Furthermore, thanks to the high sequencing depth applied and to the broad range of transcripts expressed, the assembly also contains a great amount of sequence data originated from genes which are not directly linked to liver and testis, permitting to obtain a good overview of the overall coelacanth transcriptome.

In addition, the RNA-seq data generated in the present work provided a valuable resource for the Ensembl annotation of the recently sequenced genome of the African coelacanth *L. chalumnae*. In fact, the paired-end sequence data from liver and testis were processed through the Ensembl RNA-seq pipeline, generating 9,364 high confidence gene models, which permitted to improve the genome annotation by the addition of 547 new genes and 1,782 related transcripts (Amemiya et al. 2012). This was a considerable improvement with respect to the previous annotation, based both on sequence similarity and on the data provided by the RNA-seq of *L. chalumnae* muscle, a tissue transcriptionally poor if compared with liver and testis. Nevertheless, the mapping performed on the genomic scaffolds

revealed that a remarkable amount of sequence data remained not used for the gene predictions. These data likely include valuable information about non-coding transcripts and genes whose prediction by automated pipelines is particularly difficult.

The importance of the obtained results mainly regards the origin of terrestrial vertebrates since the key position of *Latimeria* as the unique extant representative of the lineage from which tetrapods should have arisen. The transcriptome data indicate a higher affinity of this species to several terrestrial vertebrates, even if only in few species the genome and transcriptome have been exhaustively analyzed.

Therefore, the *de novo* transcriptome assembly, for the quality of information it generated, may certainly be considered a step ahead in helping to understand the biology of this living fossil.

Methods

Samples collection

On 16th September 2009 a coelacanth was found in a shark net near Talise Island, Indonesia. This male, weighing 27 kg with a total length of 116 cm, was the fifth specimen of *L. menadoensis* ever caught since the discovery of this species in 1997. The animal was moved to the Faculty of Fisheries and Marine Science, University of Sam Ratulangi, Manado (Indonesia), where the liver and testis tissues used in this study were collected immediately after death and directly fixed in RNAlater (Applied Biosystems, Warrington, UK). Tissue samples were shipped to the Science Faculty, Università Politecnica delle Marche, Ancona, Italy, under the Convention on International Trade in Endangered Species (CITES; permit no. IT/IM/2009/MCE/01585-2009/19713).

RNA extraction

Total RNA was isolated from liver and testis using TRIzol reagent (Invitrogen, Carlsbad, CA). Following the treatment with DNase I Amplification Grade (Sigma, Steinheim, Germany), an aliquot of the extracts was used to assess the quality and quantity of RNA by spectrophotometric and capillary electrophoretic analysis. The liver RNA sample resulted to have a 260/280 nm absorbance ratio of 1.74, a 260/230 nm absorbance ratio of 0.94 and a RNA integrity number (RIN, estimated with an Agilent2100 Bioanalyzer) of 6.6. The testis RNA sample resulted to have a 260/280 and a 260/230 nm absorbance ratios of 1.89 and 1.23, respectively, with a RIN of 7.

Sequencing of the liver and testis transcriptomes

Messenger RNA selection and cDNA library preparation were performed by the Istituto di Genomica Applicata (IGA, Udine, Italy). The sequencing of the libraries was performed on an Illumina Genome Analyzer II platform (San Diego, California).

Briefly, the poly-A mRNAs were selected using magnetic beads-linked oligo (dT) probes. The fragmentation was obtained with divalent cations. cDNA was synthesized and Illumina sequencing adapters were then ligated to the fragments, according to the manufacturer's protocol. A smear of ligated fragments of 150 to 400 bp of length was selected by size and excised from an agarose gel. The sequencing of the cDNA libraries was performed on a flow cell using a 100-cycles paired-end strategy.

Data processing and *de novo* assembly of *Latimeria menadoensis* transcriptome

The raw sequencing reads were trimmed by removing Illumina adapter sequences and low quality bases (the quality limit was set to 0.05). The resulting trimmed sequences shorter than 75 bp were discarded. All the reads originated from ribosomal RNA were also removed prior to the assembly step.

The *de novo* assembly of the processed reads was performed with a combined approach, by integrating the outputs of two different methods, which have been specifically developed for *de novo* assembly of short reads: Trinity (Grabherr et al. 2011) and the commercially available CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark). At first, the two assemblies were performed separately using as input the same sequence set, comprising both liver and testis sequence data.

The *de novo* Trinity assembly was completed using the November 2011 version of Trinity. It was run using the strand-specific data option which was set to RF. All other options were set to their default values. Only the longest transcripts per each gene were selected for further analysis. Redundant contigs created by Trinity were collapsed by a MIRA 3.4.0 assembly (Chevreux et al. 2004).

The *de novo* CLC assembly was performed assuming a paired-end read distance comprised between 100 and 350 bp and the penalties for mismatches, insertions, and deletions were set at 2\3\3, whereas the parameters for the length fraction and similarity were set to 0.5 and 0.9, respectively. The paired-end read distance was empirically determined after several preliminary *de novo* assemblies followed by analysis of paired-end read mapping, which showed this range to be normally distributed with the highest frequency at 240 bp. The minimum allowed assembled contig length was set at 250 bp. Only contigs assembled with high confidence were kept and used for the implementation of the Trinity assembly whenever possible. A particular emphasis was put on protein-coding transcripts, as only contigs displaying an open reading frame (ORF) of a minimum of 70 codons were selected. The ORF prediction was carried out with the “Find Open Reading Frames” tool included in the CLC Genomic Workbench, considering AUG as a start codon and selecting the “open-ended sequence” option.

Identical or highly similar contigs generated by the two different *de novo* assemblers were detected by BLASTn, setting the cutoff to an e-value of 1e-100 and to an identity of 98%. Contigs generated by the CLC assembler identical to those created by Trinity were discarded, unless they were extending the Trinity contigs by at least 200 bp. In the latter case, Trinity contigs were replaced by their CLC counterparts. The schematic summary of the procedure used for integrating the outputs of the two assemblers is shown in **Figure 1**.

Finally, to ensure the creation of a highly reliable set of assembled transcripts, contigs covered by a low number of reads were discarded, following a global mapping of the complete set of both liver and testis filtered reads (CLC Genomic Workbench, mismatch/insertion/deletion costs set at 2/3/3, length fraction/similarity set to 0.75/0.95). All the transcripts showing an average coverage <5 were considered as possible fragments of longer transcripts, not reliable enough to be included in the high quality coelacanth transcript collection, and were therefore discarded. Only transcripts longer than 249 bp were kept.

Assembly quality assessment

In order to assess the quality of the contigs obtained with the filtering procedure in respect with the non-filtered set, the sequences were grouped into categories according to their sizes (intervals of 100 bp) and the relative abundance of each category was plotted in a histogram. The distributions of transcript lengths pre- and post-filtering were compared (**Figure 3**).

The sequence redundancy was estimated by the RNA-seq mapping of the reads from both tissues on the contigs created by the original Trinity assembly and to the filtered and non-filtered sets of contigs obtained with the Trinity and CLC combined approach. The RNA-seq analysis tool included in the CLC Genomic Workbench was used for this purpose (minimum length fraction and minimum similarity fraction were set at 0.75 and 0.95, respectively). The total number of reads mapped and the set of reads mapping non-specifically (matching on more than 1 contig) were compared, in order to evaluate the improvement of the assembly quality obtained with the processing steps. Sequence redundancy was calculated as the number of reads mapping not-specifically normalized on the total number of mapped reads (**Figure 2**).

The total number of reads originated from mitochondrial RNA was assessed by the mapping of the filtered reads set to the deposited mitochondrial DNA sequence of *L. menadoensis* (Genbank accession: NC_006921.2). The mapping was performed with the CLC Genomic Workbench, using the same settings described above to estimate sequence redundancy.

The approximate abundance of full length transcripts and the fragmentation in the collection were also estimated using the Ortholog Hit Ratio method (O'Neil et al. 2010), using the NCBI non-redundant (nr) protein database for the determination of the hit length regions through BLASTx. A correction was applied to the standard method in order to remove the bias given by inter-species divergence, as only contigs displaying BLASTx identity higher than 90%, independently from the alignment length, were considered as “true orthologs and selected for the analysis (**Figure 4**).

Transcript functional annotation

The filtered transcripts were annotated with Blast2GO (version 2.4.4, <http://www.blast2go.com/>), a tool specifically developed for the annotation of novel sequence sets (Conesa et al. 2005). Sequence similarity was evaluated with BLASTx (Altschul et al. 1990) against the NCBI non-redundant (nr) protein database using an e-value cutoff of 1e-6.

The presence of conserved domains was researched and annotated using Interproscan (Zdobnov and Apweiler 2001) on the six possible translation frames of each contig.

Contigs were functionally annotated according to the Gene Ontology (<http://www.geneontology.org/>) nomenclature. GO terms were assigned to each transcript and annotated according to the level 2 of the Cell Component, Molecular Function, and Biological Process categories.

Furthermore, in order to identify by homology transposable elements and repeated sequences from a database of vertebrate repeats, the contigs were analysed with RepeatMasker (Smit 1996-2012).

Mapping on *L. chalumnae* genome

The liver and testis sets of filtered reads were mapped on the annotated *L. chalumnae* genome Ensembl release e!67 using the Genomic Workbench 4.5.1 RNA-seq tool, assuming a minimum length fraction of 0.75 and a minimum similarity fraction allowed of 0.95. As the sequence similarity between *L. menadoensis* and *L. chalumnae* was estimated to be 99.73%, the mapping parameters used were supposed not to significantly influence the mapping outcome. The allowed paired-end read distance was set between 100 and 350 bp. Based on gene annotations, it was possible to categorize the fragments as mapping within exons, within introns and on exon-exon or exon-intron junctions. Furthermore, the number of reads mapping on non-annotated genomic regions was also calculated, to assess the amount of sequence data accounting for the expression of non-annotated genes.

RNA-seq analysis

The liver and testis filtered reads were separately mapped to the high quality set of the assembled contigs to assess the expression values in the two tissues. The mapping was carried out with the Genomic Workbench 4.5.1 RNA-seq tool, with a minimum length fraction allowed of 0.75 and a minimum similarity fraction allowed of 0.95. Paired-end read distance was considered to be comprised between 100 and 350 bp. Only intact sequence pairs (fragments) mapping were counted

and expression values were calculated as FPKM (Fragments Per Kilobase per Million fragments mapped).

Besides liver and testis, also RNA-seq data obtained from the African congener *L. chalumnae* muscle (Sequence Read Archive sample ID: SRS283232) were used for comparison purpose (the muscle transcriptome was *de novo* assembled with Trinity and processed to remove redundancy exactly as previously described for the liver and testis assembly).

The transcriptome richness was graphically inspected by plotting the cumulative number of reads mapped on each of the 1,000 most expressed transcripts in each tissue, normalized on the total number of reads mapped (**Figure 9**).

The overlap between liver and testis transcriptomes was estimated by the comparison of the sets comprising the 1,000 most expressed genes per tissue. The comparison was also extended to the *L. chalumnae* muscle transcript set generated in the frame of the African coelacanth genome sequencing project (Amemiya et al. 2012) (**Figure 10**).

Data Access

The raw sequence data generated by Illumina sequencing of *L. menadoensis* liver and testis samples were deposited at the NCBI Sequence Read Archive and are accessible at the study ID SRS362269-70.

Acknowledgments

The authors of the Dipartimento di Scienze della Vita e dell'Ambiente, Università Politecnica delle Marche (Ancona, Italy) are affiliated to Istituto Nazionale Biosistemi e Biostrutture (INBB).

Figures

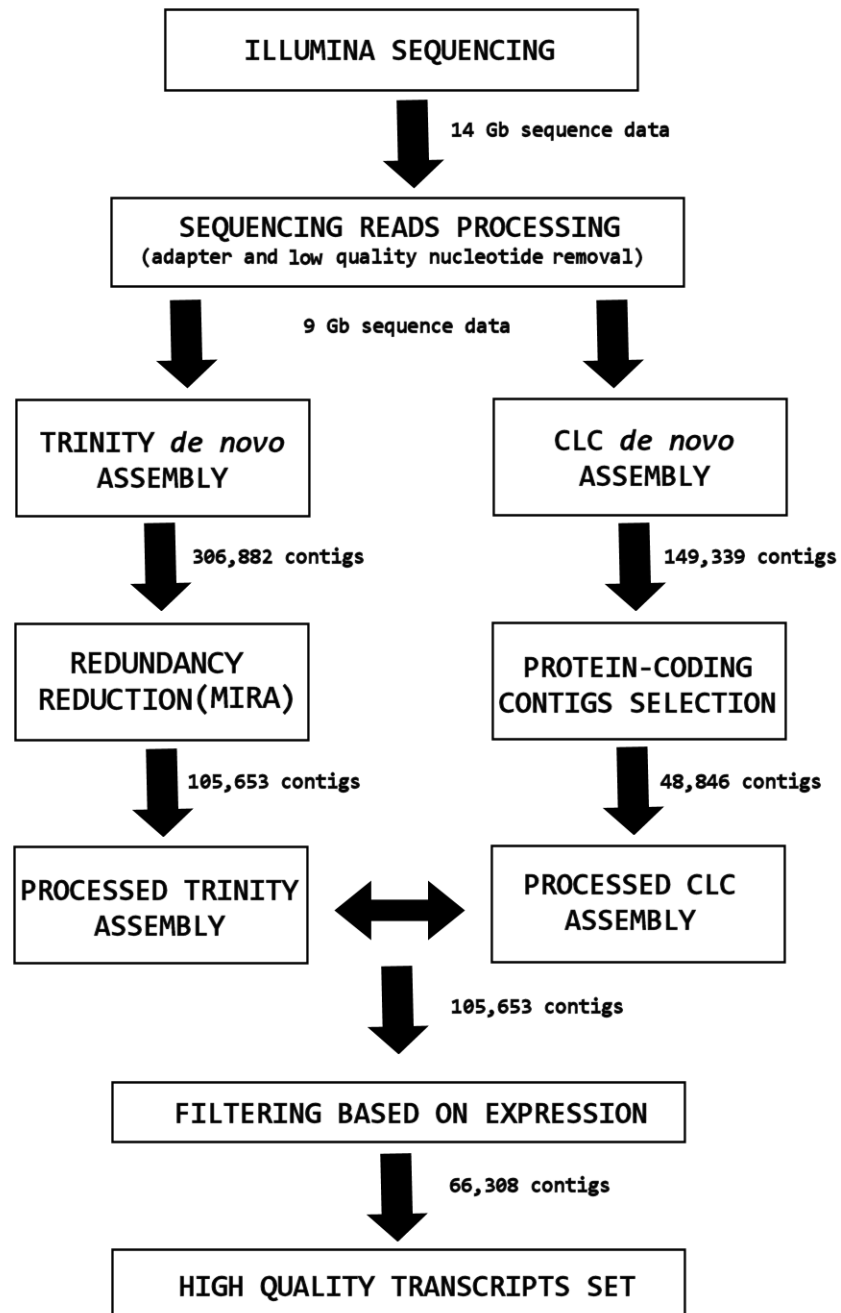


Figure 1: Graphic summary of the combined *de novo* assembly strategy and filtering steps applied to generate the final high quality transcripts set comprising 66,308 sequences.

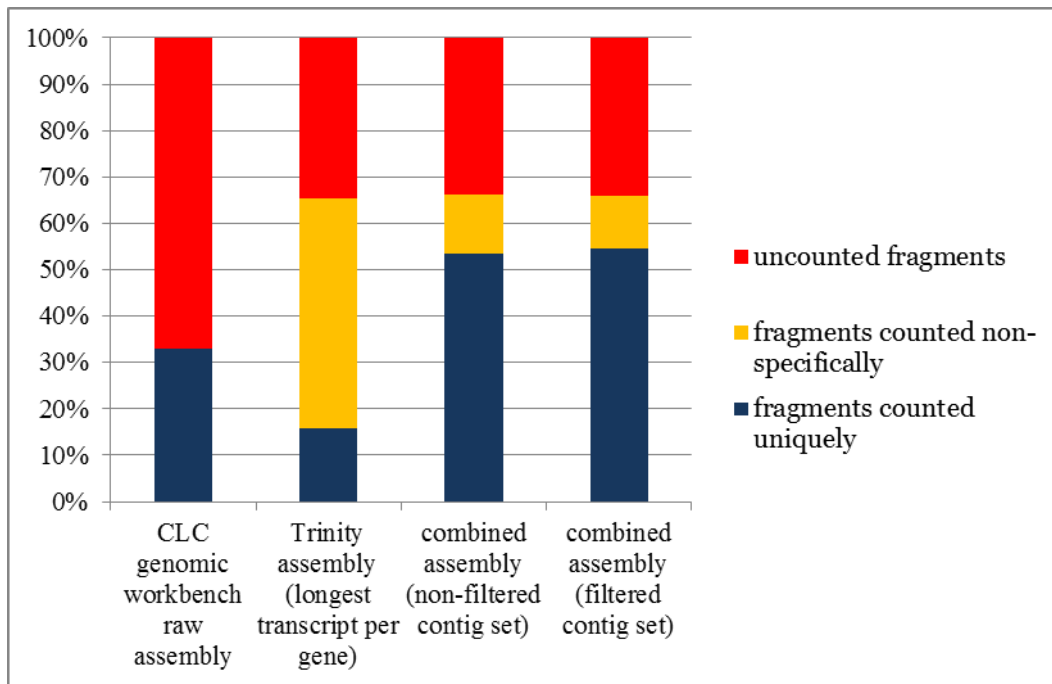


Figure 2: Sequencing read usage in the assemblies and assembly redundancy estimated by RNA-seq mapping. Redundancy is calculated as the number of fragments mapping non-specifically on multiple contigs. Fragments mapping on contigs as broken read pairs were not counted.

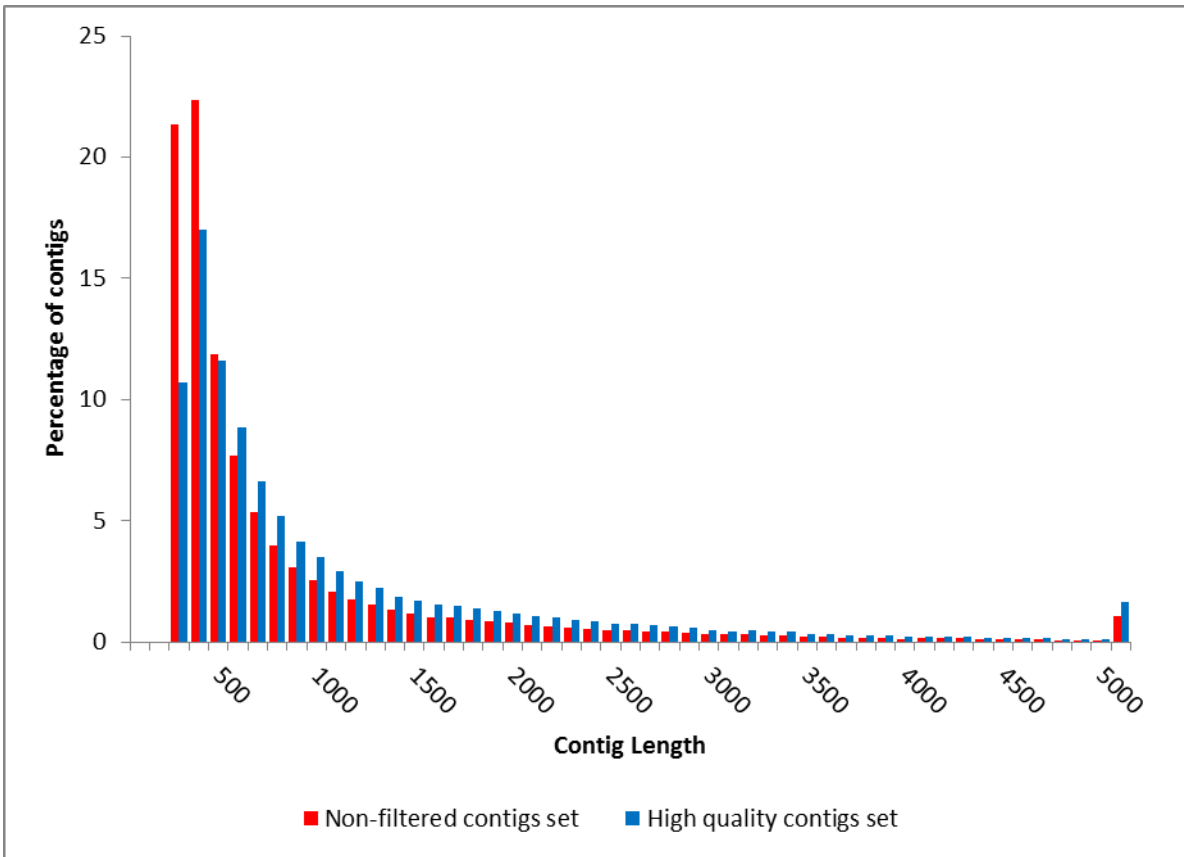


Figure 3: Comparison of contig length distribution before (red) and after (blue) the filtering step based on average sequence coverage. The reduction of the fraction of short contigs is represented by the shift of distribution towards the right side of the graph. X: Length categories, organized in 100 bp intervals. Y: number of contigs observed per category, normalized on the total number of contigs assembled.

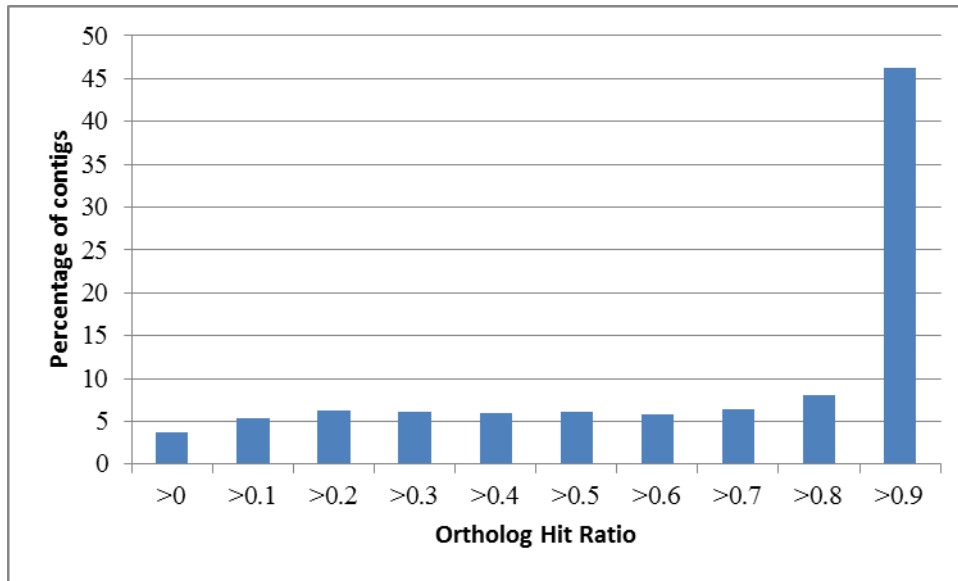


Figure 4: Ortholog Hit Ratio, calculated on the high quality set of liver and testis transcripts. The ratio of length between assembled contigs and the full length orthologs is reported on the X axis, the number of contigs observed in each ratio category, normalized on the total number of contigs used in the analysis is shown on the Y axis.

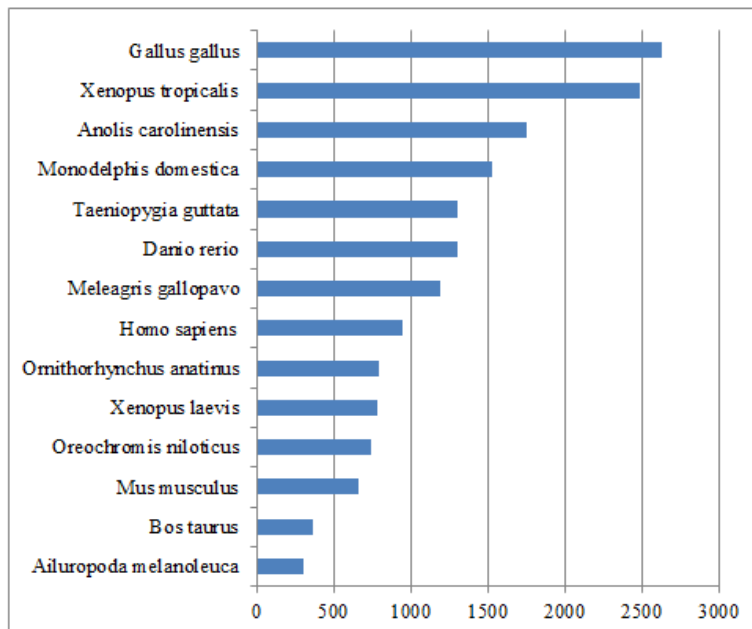


Figure 5: Top BLAST hit species distribution, obtained by BLASTx against the NCBI non-redundant (nr) protein database. Only the 15 most represented species are shown. The complete number of top hits of other organisms is 7,572.

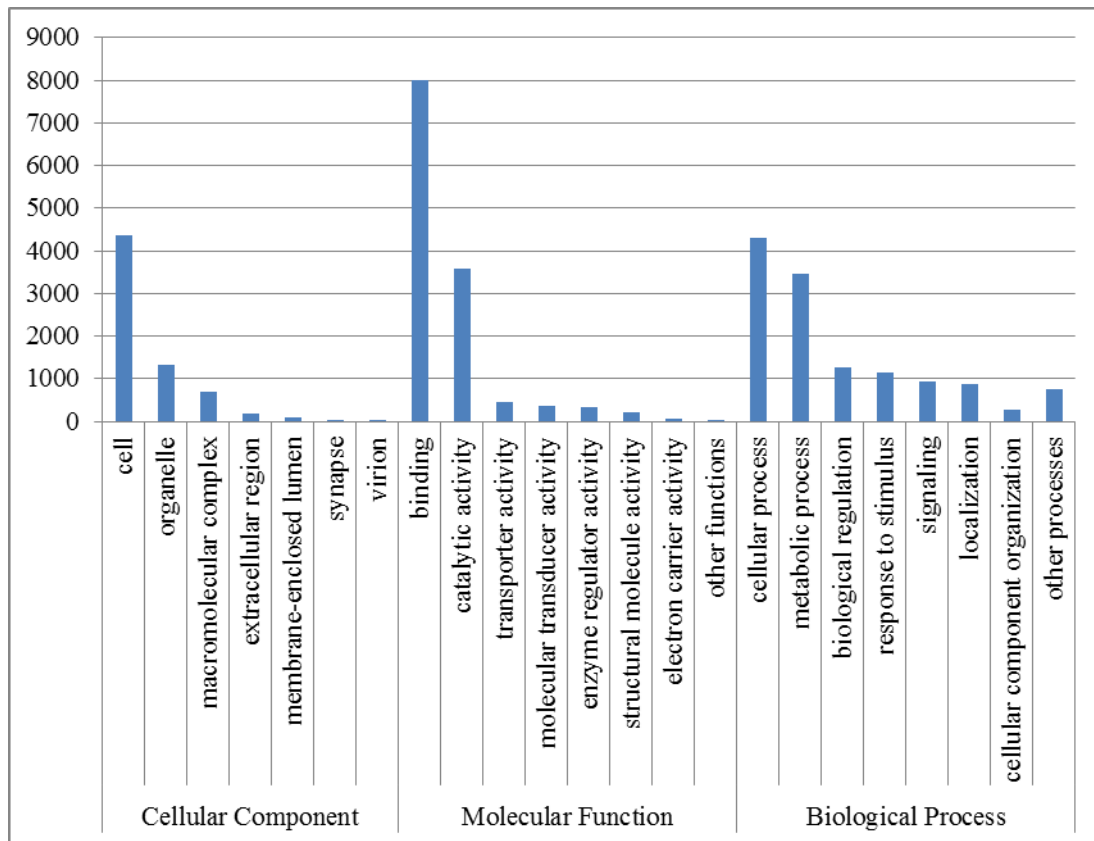


Figure 6: Gene Ontology mapping performed on the high quality transcript set. The mapping summary takes into account annotations at Level 2 of Cell Component, Molecular Function and Biological Process.

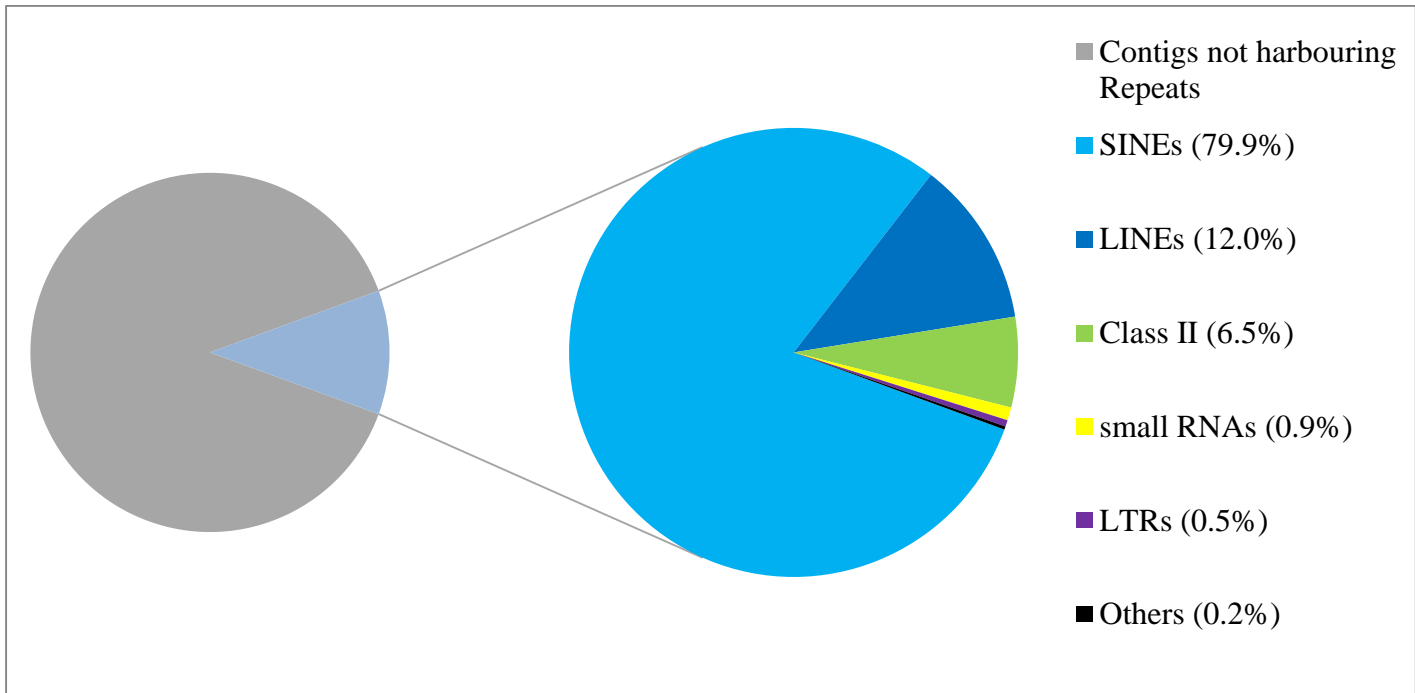


Figure 7: Contigs harbouring a repeat element identified by RepeatMasker. SINEs (Short interspersed elements); LINEs (Long interspersed elements); Class II (DNA transposons); small RNAs (non-coding RNAs: tRNAs, srpRNAs, snRNAs, 7SK RNAs); LTRs (Long terminal repeats); Others (Unknown and Satellite elements).

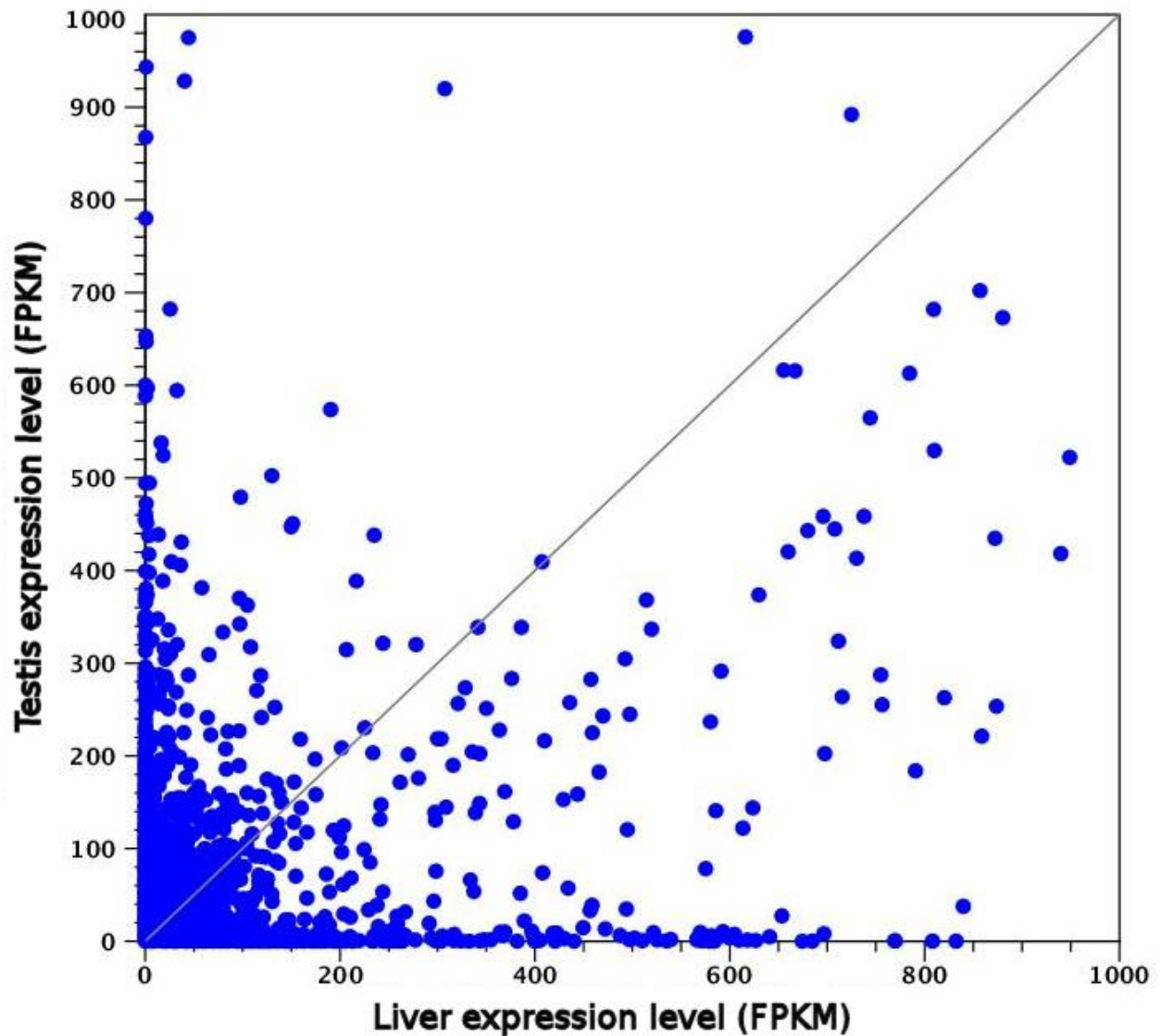


Figure 8: Scatter plot depicting the expression levels (calculated as FPKM, Fragments Per Kilobase per Million fragments mapped) in liver and testis. Genes whose expression levels are identical in the two organs are located on the bisector. For graphical representation convenience, only genes whose expression was lower than 1,000 FPKM in both tissues are shown (therefore only 79 genes are not shown in the graph).

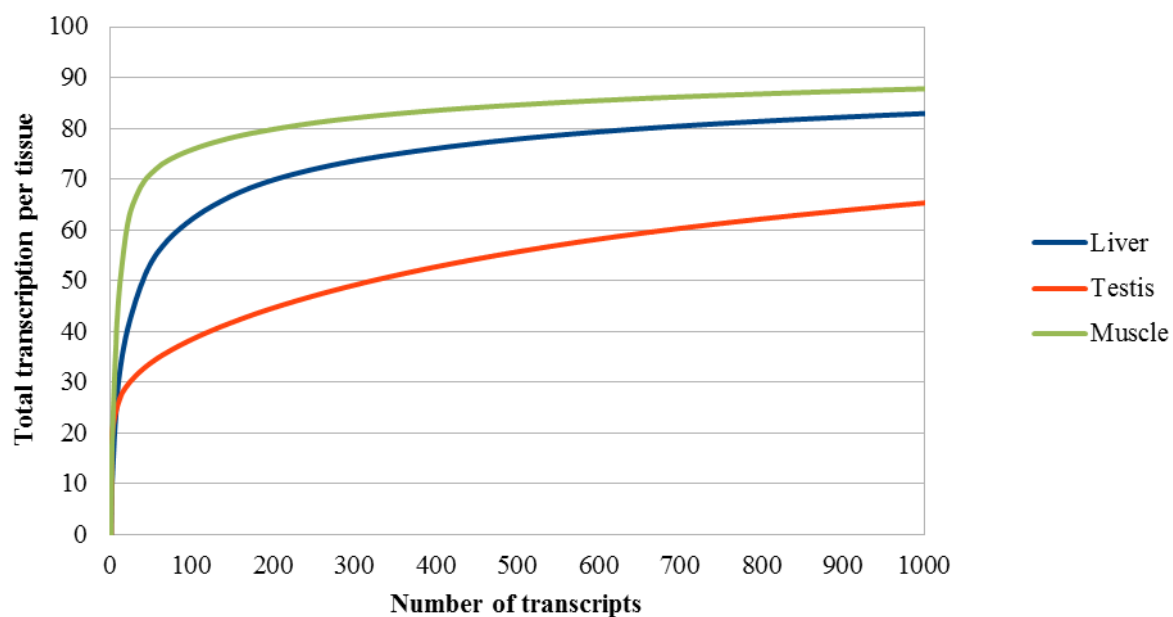


Figure 9: Transcriptomic richness of *L. menadoensis* liver and testis and of *L. chalumnae* muscle, shown as the cumulative number of reads mapping on the 1,000 most expressed transcripts per each tissue, normalized on the total number of reads mapped on all transcripts (Y axis).

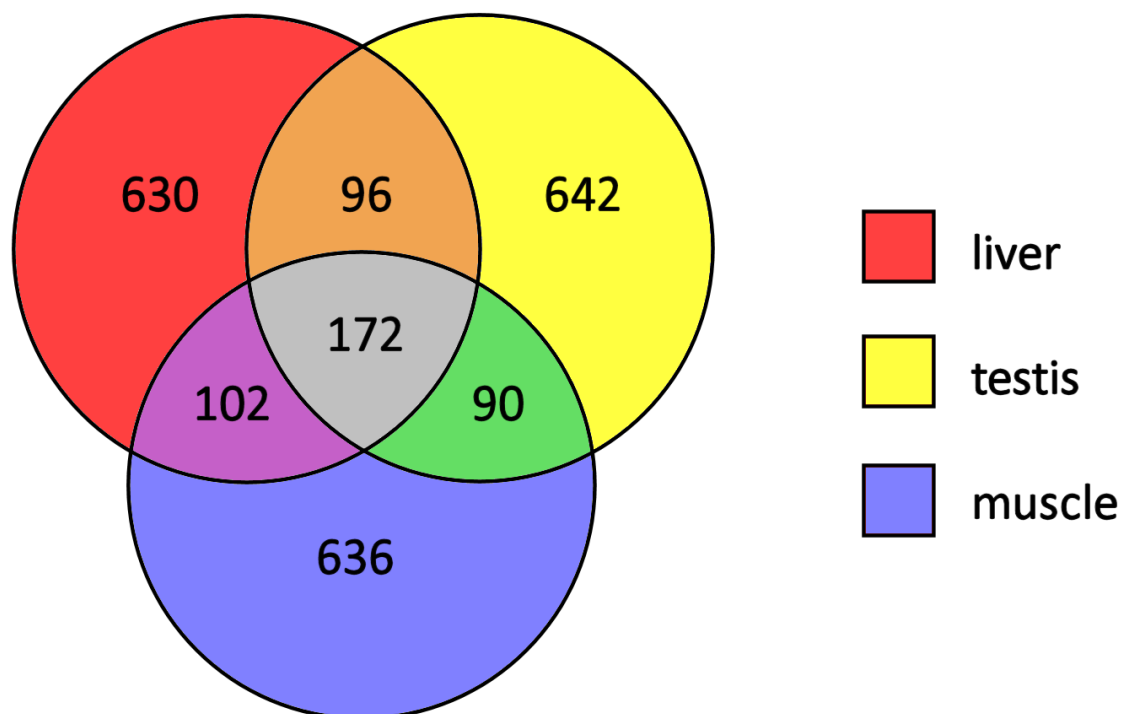


Figure 10: Venn diagram depicting the overlap between liver, testis and muscle transcriptomes evaluated on the 1,000 most expressed transcripts in each tissue.

Tables

Table 1: Trimming report

	Liver	Testis
Number of reads before trimming	76,932,818	68,502,338
Reads kept after trimming	64,099,318	55,326,118
Percentage of discarded reads	10.96%	13.69%
Reads average length before trimming	97,28	96,22
Reads average length after trimming	103,4	102,9
Ribosomal RNA reads	16,628,740	13,924,282
Percentage of ribosomal RNA reads	21.61%	20,33%
Number of high quality reads	47,470,578	41,401,836

Table 2: Assembly statistics

Total number of high quality assembled reads	88,872,414
Number of created contigs	66,308
Number of bases in contigs	71,621,287
Average length (bp)	1,080
Median length (bp)	626
N50	1,761
N80	662
N90	438
Longest contig (bp)	20,815
Number of contigs > 5 Kb	1,081
Number of contigs > 10 Kb	80

Table 3: the 25 most abundant Interpro domains revealed by the Interproscan annotation of the high quality coelacanth transcript set.

Interpro domain	Description	Number of detected contigs
IPR000719	Protein kinase, catalytic domain	2041
IPR007087	Zinc finger, C2H2	1778
IPR002290	Serine/threonine- / dual-specificity protein kinase, catalytic domain	1472
IPR013783	Immunoglobulin-like fold	1130
IPR015880	Zinc finger, C2H2-like	1056
IPR020635	Tyrosine-protein kinase, catalytic domain	981
IPR011009	Protein kinase-like domain	946
IPR020683	Ankyrin repeat-containing domain	927
IPR001680	WD40 repeat	845
IPR001849	Pleckstrin homology domain	834
IPR003961	Fibronectin, type III	822
IPR001452	Src homology-3 domain	792
IPR008271	Serine/threonine-protein kinase, active site	696
IPR001841	Zinc finger, RING-type	687
IPR007110	Immunoglobulin-like	684
IPR017986	WD40-repeat-containing domain	675
IPR020849	Small GTPase superfamily, Ras type	674
IPR013087	Zinc finger C2H2-type/integrase DNA-binding domain	662
IPR000504	RNA recognition motif domain	661
IPR002110	Ankyrin repeat	659

IPR013083	Zinc finger, RING/FYVE/PHD-type	614
IPR001478	PDZ domain	600
IPR015943	WD40/YVTN repeat-like-containing domain	600
IPR001650	Helicase, C-terminal	575
IPR016024	Armadillo-type fold	564

Table 4: RNA-seq mapping on *L. chalumnae* genome statistics

Mapping on annotated genes	
Counted fragments	21,589,809 (48.59%)
Uncounted fragments	22,846,398 (51.41%)
Match specificity	98.1%
Reads mapped in pairs	43,179,618 (48.59%)
Reads mapped in broken pairs	11,603,481 (13.06%)
Unmapped reads	34,089,315 (38.36%)
Fragments mapped on exon-exon junctions	11,147,648 (51.63%)
Fragments mapped on exon-intron junctions	642,299 (2.98%)
Total fragments mapped on exons	20,084,744 (93.03%)
Total fragments mapped on introns	1,505,065 (6.97%)

Mapping on genomic scaffolds

Total number of reads mapping on genomic scaffolds	85,682,920 (96.41%)
Reads mapped on non-annotated genes	30,899,821 (34.77%)
Unmapped reads	3,189,494 (3.59%)

Table 5: RNA-seq statistics

Liver	Liver	Testis
Counted fragments	15,949,179	13,363,810
Percentage of counted fragments	67.20%	64.57%
-uniquely	13,479,204	10,734,166
-non-specifically	2,469,975	2,629,644
Uncounted fragments	7,786,110	7,337,108
Reads mapped in pairs	31,898,358	26,727,620
Reads mapped in broken pairs	5,187,130	5,111,602
Percentage of mapped reads	78.12%	76.9%
Reads not mapped	10,385,090	9,562,614

Table 6: List of the 20 most expressed genes in liver and testis

Gene name	Liver expression level (FPKM*)	Testis expression level (FPKM)
α -2-macroglobulin like 1	20,842.949	4.422
Apolipoprotein AIV-like	13,160.855	0.678
Inner centromere protein A	12,979.576	29.031
Vitellin layer outer membrane 1	11,014.570	1.324
Fibrinogen α chain	8,185.427	0.045
Fibrinogen β chain	7,773.433	1.786
Hemopexin	7,131.137	0.832
Elongation factor 1- α	6,729.351	2,516.552
Serine proteinase inhibitor Kazal type 2	5,697.525	0.439
Ferritin H	5,000.387	613.380
ATP synthase F0 subunit 6	4,647.673	1,680.129
Lipocalin	4,278.524	314.124
Apolipoprotein E	4,210.166	46.594
Ferritin heavy polypeptide 1	3,826.063	339.725
Riboflavin-binding protein	3,675.538	12.429
Serum albumin	3,601.101	0.400
α -2 macroglobulin	3,547.877	0.136
Fibrinogen gamma polypeptide	3,482.420	13.593
Vitronectin	3,481.011	0.111
Serum amyloid P	3,344.509	15.287

*Fragments Per Kilobase per Million fragments mapped.

Table 7: List of the 20 most expressed genes in testis and liver

Gene name	Testis expression level (FPKM*)	Liver expression level (FPKM)
Testis-specific histone	156,927.735	38.907
Prostaglandin H2D isomerase	6,494.277	4.650
Y-box transcription factor	4,264.142	358.797
Sjogren syndrome nuclear autoantigen 1	3,898.979	0.000
Tubulin α chain, testis-specific	3,317.532	25.942
Elongation factor 1- α	2,516.552	6729.351
Histone H1x-like	2,033.078	7.469
H\ACA ribonucleoprotein complex, subunit 2	1,992.342	3.647
Unknown	1,952.952	1.078
Tubulin β 2-C	1,757.980	8.536
ATP synthase F0 subunit 6	1,680.129	4647.673
Sperm nuclear basic protein PL-I	1,438.904	0.573
Centrin-1	1,346.417	4.526
Ferritin heavy chain	1,223.551	649.598
HSP90- β	1,213.755	507.986
Ubiquitin	1,211.773	424.873
Cra-B	1,151.852	0.838
TP-53 target gene protein-like	1,139.390	0.400
Ribosomal protein S6	1,012.800	1,690.661
High mobility group protein B2	975.940	615.996

*Fragments Per Kilobase per Million fragments mapped.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3): 403-410.
- Amemiya CT et al., Comparative analysis of the genome of the African coelacanth, *Latimeria chalumnae*, sheds light on tetrapod evolution. *Nature* 2012 (submitted).
- Amemiya CT, Powers TP, Prohaska SJ, Grimwood J, Schmutz J, Dickson M, Miyake T, Schoenborn MA, Myers RM, Ruddle FH et al. 2010. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proceedings of the National Academy of Sciences of the United States of America* 107(8): 3622-3627.
- Balhorn R. 2007. The protamine family of sperm nuclear proteins. *Genome Biology* 8(9): 227.
- Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441(1): 87-90.
- Brinkmann H, Venkatesh B, Brenner S, Meyer A. 2004. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* 101(14): 4900-4905.
- Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources*.
- Canapa A, Olmo E, Forconi M, Pallavicini A, Makapedua MD, Biscotti MA, Barucca M. 2012. Composition and Phylogenetic Analysis of Vitellogenin Coding Sequences in the Indonesian Coelacanth *Latimeria menadoensis*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 318(5): 404-416.
- Chevreur B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S. 2004. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research* 14(6): 1147-1159.
- Cloutier R, Ahlberg PE. 1997. Morphology, characters, and the interrelationships of basal sarcopterygians. In *Interrelationship of fishes*, (ed. LJ Stiassny, LR Parenti, G Johnson), pp. 445-479. Academic Press, San Diego, CA.

- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18): 3674-3676.
- Danke J, Miyake T, Powers T, Schein J, Shin H, Bosdet I, Erdmann M, Caldwell R, Amemiya CT. 2004. Genome resource for the Indonesian coelacanth, *Latimeria menadoensis*. *Journal of Experimental Zoology Part A: Comparative Experimental Biology* 301(3): 228-234.
- Eklom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107(1): 1-15.
- Erdmann MV, Caldwell RL, Jewett SL, Tjakrawidjaja A. 1999. The second recorded living coelacanth from north Sulawesi. *Environmental Biology of Fishes* 54(4): 445-451.
- Ewen-Campen B, Shaner N, Panfilio K, Suzuki Y, Roth S, Extavour C. 2011. The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. *BMC Genomics* 12(1): 61.
- Feldmeyer B, Wheat C, Krezdorn N, Rotter B, Pfenninger M. 2011. 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* 12(1): 317.
- Gorr T, Kleinschmidt T, Fricke H. 1991a. Close tetrapod relationships of the coelacanth *Latimeria* indicated by haemoglobin sequences. *Nature* 351(6325): 394-397.
- Gorr T, Kleinschmidt T, Sgouros JG, Kasang L. 1991b. A "living fossil" sequence: primary structure of the coelacanth (*Latimeria chalumnae*) hemoglobin--evolutionary and functional aspects. *Biological Chemistry Hoppe-Seyler* 372(8): 599-612.
- Gould SJ, Vrba ES. 1982. Exaptation - a missing term in the science of form. *Paleobiology* 8(1): 4-15.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7): 644-652.
- Gwee PC, Amemiya CT, Brenner S, Venkatesh B. 2008. Sequence and organization of coelacanth neurohypophysial hormone genes: Evolutionary history of the vertebrate neurohypophysial hormone gene locus. *BMC Evolutionary Biology* 8(1).

- Hadzhiev Y, Lang M, Ertzer R, Meyer A, Strähle U, Müller F. 2007. Functional diversification of sonic hedgehog paralog enhancers identified by phylogenomic reconstruction. *Genome Biology* 8(6).
- Han G-Z, Worobey M. 2012. An Endogenous Foamy-like Viral Element in the Coelacanth Genome. *PLoS Pathog* 8(6): e1002790.
- Holder MT, Erdmann MV, Wilcox TP, Caldwell RL, Hillis DM. 1999. Two living species of coelacanths? *Proceedings of the National Academy of Sciences of the United States of America* 96(22): 12616-12620.
- Inoue JG, Miya M, Venkatesh B, Nishida M. 2005. The mitochondrial genome of Indonesian coelacanth *Latimeria menadoensis* (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. *Gene* 349: 227-235.
- Kemphues KJ, Kaufman TC, Raff RA, Raff EC. 1982. The testis-specific α -tubulin subunit in *Drosophila melanogaster* has multiple functions in spermatogenesis. *Cell* 31(3): 655-670.
- Koh EGL, Lam K, Christoffels A, Erdmann MV, Brenner S, Venkatesh B. 2003. Hox gene clusters in the Indonesian coelacanth, *Latimeria menadoensis*. *Proceedings of the National Academy of Sciences of the United States of America* 100(3): 1084-1088.
- Kumar S, Blaxter M. 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* 11(1): 571.
- Lanfranchi G, Muraro T, Caldara F, Pacchioni B, Pallavicini A, Pandolfo D, Toppo S, Trevisan S, Scarso S, Valle G. 1996. Identification of 4370 expressed sequence tags from a 3'-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. *Genome Research* 6(1): 35-42.
- Maisey JG. 1996. *Discovering fossil fishes*. Holt, New York.
- Majewski J, Schwartzenuber J, Lalonde E, Montpetit A, Jabado N. 2011. What can exome sequencing do for you? *Journal of Medical Genetics* 48(9): 580-589.
- Makapedua DM, Barucca M, Forconi M, Antonucci N, Bizzaro D, Amici A, Carradori MR, Olmo E, Canapa A. 2011. Genome size, GC percentage and 5mC level in the Indonesian coelacanth *Latimeria menadoensis*. *Marine Genomics* 4(3): 167-172.
- Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24(3): 133-141.

- Martianov I, Brancorsini S, Catena R, Gansmuller A, Kotaja N, Parvinen M, Sassone-Corsi P, Davidson I. 2005. Polar nuclear localization of H1T2, a histone H1 variant, required for spermatid elongation and DNA condensation during spermiogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 102(8): 2808-2813.
- Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics* 12(10): 671-682.
- Meyer A. 1995. Molecular evidence on the origin of tetrapods and the relationships of the coelacanth. *Trends in Ecology and Evolution* 10(3): 111-116.
- Meyer A, Dolven SI. 1992. Molecules, fossils, and the origin of tetrapods. *Journal of Molecular Evolution* 35(2): 102-113.
- Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* 11(10): 685-696.
- Modisakeng K, Jiwaji M, Pesce E-R, Robert J, Amemiya C, Dorrington R, Blatch G. 2009. Isolation of a <i>Latimeria menadoensis</i> heat shock protein 70 (<i>Lmhs70</i>) that has all the features of an inducible gene and encodes a functional molecular chaperone. *Molecular Genetics and Genomics* 282(2): 185-196.
- Morozova O, Marra MA. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92(5): 255-264.
- Mulley JF, Holland PWH. 2010. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. *Molecular Biology and Evolution* 27(10): 2386-2391.
- Nishihara H, Smit AFA, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Research* 16(7): 864-874.
- Noonan JP, Grimwood J, Danke J, Schmutz J, Dickson M, Amemiya CT, Myers RM. 2004. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Research* 14(12): 2397-2405.
- O'Neil S, Dzurisin J, Carmichael R, Lobo N, Emrich S, Hellmann J. 2010. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics* 11(1): 310.
- Pallen MJ, Loman NJ, Penn CW. 2010. High-throughput sequencing and clinical microbiology: Progress, opportunities and challenges. *Current Opinion in Microbiology* 13(5): 625-631.

- Pouyaud L, Wirjoatmodjo S, Rachmatika I, Tjakrawidjaja A, Hadiaty R, Hadie W. 1999. A new species of coelacanth. *Une nouvelle espece de coelacanth Preuves genetiques et morphologiques* 322(4): 261-267.
- Shan Y, Gras R. 2011. 43 genes support the lungfish-coelacanth grouping related to the closest living relative of tetrapods with the Bayesian method under the coalescence model. *BMC Research Notes* 4.
- Shashikant C, Bolanowski SA, Danke J, Amemiya CT. 2004. Hoxc8 early enhancer of the Indonesian coelacanth, *Latimeria menadoensis*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 302(6): 557-563.
- Smit AFA. 1996-2012. RepeatMasker.
- Smith JJ, Sumiyama K, Amemiya CT. 2012. A living fossil in the genome of a living fossil: Harbinger transposons in the coelacanth genome. *Molecular Biology and Evolution* 29(3): 985-993.
- Smith JLB. 1939. A living fish of Mesozoic type. *Nature* 143(3620): 455-456.
1989. A surviving fish of the order Actinistia. *Transactions of the Royal Society of South Africa* 47(1): 9-12.
- Sudarto, Lalu XC, Kosen JD, Tjakrawidjaja AH, Kusumah RV, Sadhotomo B, Kadarusman, Pouyaud L, Slembrouck J, Paradis E. 2010. Mitochondrial genomic divergence in coelacanths (*Latimeria*): Slow rate of evolution or recent speciation? *Marine Biology* 157(10): 2253-2262.
- Takezaki N, Figueroa F, Zaleska-Rutczynska Z, Takahata N, Klein J. 2004. The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of forty-four nuclear genes. *Molecular Biology and Evolution* 21(8): 1512-1524.
- Tanaka H, Baba T. 2005. Gene expression in spermiogenesis. *Cellular and Molecular Life Sciences* 62(3): 344-354.
- Tohyama Y, Ichimiya T, Kasama-Yoshida H, Cao Y, Hasegawa M, Kojima H, Tamai Y, Kurihara T. 2000. Phylogenetic relation of lungfish indicated by the amino acid sequence of myelin DM20. *Molecular Brain Research* 80(2): 256-259.
- Venkatesh B, Erdmann MV, Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* 98(20): 11382-11387.

- Villasante A, Wang D, Dobner P, Dolph P, Lewis SA, Cowan NJ. 1986. Six mouse alpha-tubulin mRNAs encode five distinct isotypes: testis-specific expression of two sister genes. *Molecular and Cellular Biology* 6(7): 2409-2419.
- Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1): 57-63.
- Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH et al. 2005. The repetitive landscape of the chicken genome. *Genome Research* 15(1): 126-136.
- Xie X, Kamal M, Lander ES. 2006. A family of conserved noncoding elements derived from an ancient transposable element. *Proceedings of the National Academy of Sciences of the United States of America* 103(31): 11659-11664.
- Yokobori SI, Hasegawa M, Ueda T, Okada N, Nishikawa K, Watanabe K. 1994. Relationship among coelacanths, lungfishes, and tetrapods: A phylogenetic analysis based on mitochondrial cytochrome oxidase I gene sequences. *Journal of Molecular Evolution* 38(6): 602-609.
- Yokoyama S, Tada T. 2000. Adaptive evolution of the African and Indonesian coelacanths to deep-sea environments. *Gene* 261(1): 35-42.
- Zardoya R, Cao Y, Hasegawa M, Meyer A. 1998. Searching for the closest living relative(s) of tetrapods through evolutionary analyses of mitochondrial and nuclear data. *Molecular Biology and Evolution* 15(5): 506-517.
- Zardoya R, Meyer A. 1996. Evolutionary relationships of the coelacanth, lungfishes, and tetrapods based on the 28S ribosomal RNA gene. *Proceedings of the National Academy of Sciences of the United States of America* 93(11): 5449-5454.
- Zdobnov EM, Apweiler R. 2001. InterProScan - An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9): 847-848.
- Zhang J, Chiodini R, Badr A, Zhang G. 2011. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* 38(3): 95-109.

Comparative analysis of the genome of the African coelacanth, *Latimeria chalumnae*, sheds light on tetrapod evolution

Chris T. Amemiya^{*1,2}, Jessica Alföldi^{*3}, Alison P. Lee⁴, Shaohua Fan⁵, Hervé Philippe⁶, Iain MacCallum³, Ingo Braasch⁷, Tereza Manousaki^{5,8}, Igor Schneider⁹, Nicolas Rohner¹⁰, Chris Organ¹¹, Domitille Chalopin¹², Jeramiah J. Smith¹³, Mark Robinson¹, Rosemary A. Dorrington¹⁴, Marco Gerdol¹⁵, Bronwen Aken¹⁶, Maria Assunta Biscotti¹⁷, Marco Barucca¹⁷, Denis Baurain¹⁸, Aaron M. Berlin³, Gregory L. Blatch¹⁹, Francesco Buonocore²⁰, Thorsten Burmester²¹, Michael S. Campbell²², Adriana Canapa¹⁷, John P. Cannon²³, Alan Christoffels²⁴, Gianluca De Moro¹⁵, Adrienne L. Edkins¹⁴, Lin Fan³, Anna Maria Fausto²⁰, Nathalie Feiner^{5,25}, Mariko Forconi¹⁷, Junaid Gamielien²⁴, Sante Gnerre³, Andy Gnirke³, Jared V. Goldstone²⁶, Wilfried Haerty²⁷, Mark E. Hahn²⁶, Uljana Hesse²⁴, Steve Hoffmann²⁸, Jeremy Johnson³, Sibel I. Karchner²⁶, Shigehiro Kuraku^{5,**}, Marcia Lara³, Joshua Z. Levin³, Gary W. Litman²³, Evan Mauceli^{3,***}, Tsutomu Miyake²⁹, M. Gail Mueller³⁰, David R. Nelson³¹, Anne Nitsche³², Ettore Olmo¹⁷, Tatsuya Ota³³, Alberto Pallavicini¹⁵, Sumir Panji²⁴, Barbara Picone²⁴, Chris P. Ponting²⁷, Sonja J. Prohaska³⁴, Dariusz Przybylski³, Nil Ratan Saha¹, Vydianathan Ravi⁴, Filipe J. Ribeiro^{3,****}, Tatjana Sauka-Spengler³⁵, Giuseppe Scapigliati²⁰, Stephen M. J. Searle¹⁶, Ted Sharpe³, Oleg Simakov^{5,36}, Peter F. Stadler³², John J. Stegeman²⁶, Kenta Sumiyama³⁷, Hakim Tafer³², Jason Turner-Maier³, Peter van Heusden²⁴, Simon White¹⁶, Louise Williams³, Mark Yandell²², Henner Brinkmann⁶, Jean-Nicolas Volff¹², Clifford J. Tabin¹⁰, Neil Shubin³⁸, Manfred Schartl³⁹, David Jaffe³, John H. Postlethwait⁷, Byrappa Venkatesh⁴, Federica Di Palma³, Eric S. Lander³, Axel Meyer^{5,8,25}, Kerstin Lindblad-Toh^{3,40}

Abstract

It was a zoological sensation when a living specimen of the coelacanth was first discovered in 1938, as this lineage of lobe finned fish was thought to have gone extinct 70 million years ago. The modern coelacanth looks remarkably similar to many of its ancient relatives from the Mesozoic era and its evolutionary proximity to our own fish ancestors provides a glimpse of the fish that first walked on land. Here we report the 2.9 Gb genome sequence of the African coelacanth, *Latimeria chalumnae*. Through a phylogenomic analysis, including RNA-Seq data from the lungfish, we conclude that the lungfish, and not the coelacanth, is the closest living relative of tetrapods. Coelacanth protein-coding genes are significantly more slowly evolving than those of tetrapods, whereas other genomic features such as abundance of transposable elements and the rate of genomic rearrangements do not indicate fewer overall genomic changes. Analyses of changes in genes, and their associated regulatory elements, during the vertebrate adaptation to land, highlight important gene families and functions. These include genes involved in immunity, nitrogen excretion and the development of fins, tail, ear, eye, brain, and smell. Functional assays of enhancers suspected to be involved in the fin-to-limb transition and in the emergence of extra-embryonic placental tissues demonstrate the importance of the coelacanth genome as a blueprint for understanding tetrapod evolution.

Introduction

It was just before Christmas 1938 when Ms. Marjorie Courtenay-Latimer, the curator of a small natural history museum in East London, South Africa, discovered a large, peculiar looking fish among the myriad specimens delivered to her by a local fish trawler. Unbeknownst to her, she was to become part of the biggest fish story in 70 million years. *Latimeria chalumnae*, named after its discoverer¹, was over one meter long, bluish in coloration, and had conspicuously fleshy fins that resembled the limbs of terrestrial vertebrates. This discovery turned out to be a biological sensation and is considered one of the greatest zoological finds of the 20th century. *Latimeria* is the only living member of an ancient group of lobe-finned fishes previously known only from fossils and believed to have been extinct since the Late Cretaceous period, about 70 million years ago (MYA)¹. It took almost 15 years before a second specimen of this elusive species was discovered in the Comoros Islands in the Indian Ocean, and only a total of 309 individuals, that are known to science, have been found in the past 75 years (Rik Nulens, personal communication)². The recent discovery in 1997 of a second coelacanth species in Indonesia, *L. menadoensis*, was equally surprising, as it had been assumed that living coelacanths were confined to small populations off the East African coast^{3,4,5}. Fascination with these fish is partly due to their prehistoric appearance – remarkably, their morphology is very similar to that of fossils that date back at least 300 million years, leading to the supposition that this lineage is especially slow-evolving among vertebrates^{1,6}. *Latimeria* has also been of particular interest to evolutionary biologists due to its hotly debated relationship to our last fish ancestor – the fish that first crawled up on land⁷. In the past 15 years, targeted sequencing efforts have yielded the sequences of the coelacanth mitochondrial genomes⁸, HOX clusters⁹, and certain gene families such as protocadherins¹⁰ and neuropeptides¹¹, but still, coelacanth research has felt the lack of large-scale sequencing data.

Here we describe the sequencing and analysis of the genome of *L. chalumnae*, the African coelacanth. Findings include a definitive placement of both the coelacanth and the lungfish in the vertebrate phylogeny, a conclusive demonstration that the gene content of the coelacanth is indeed slowly evolving, and a delineation of genes and regulatory elements associated with the vertebrate land transition.

Genome Assembly and Annotation

The African coelacanth genome was sequenced and assembled (LatCha1.0) using DNA from a *Latimeria chalumnae* specimen originating from the Comoros Islands (Supplementary Figure 1). It was sequenced by Illumina sequencing technology and assembled via ALLPATHS-LG¹². The *L. chalumnae* genome has previously been reported to have a karyotype of 48 chromosomes¹³. The draft assembly is 2.86 Gb in size and is composed of 2.18 Gb of sequence plus gaps between contigs. The coelacanth genome assembly has a contig N50 size of 12.7 kb and a scaffold N50 size of 924 kb and quality metrics comparable to other Illumina genomes (See Methods and Supplementary Note 1, Supplementary Tables 1,2).

The genome assembly was annotated separately by both the Ensembl gene annotation pipeline (Ensembl release 66, February 2012) and by MAKER¹⁴. The Ensembl gene annotation pipeline created gene models using Uniprot protein alignments, limited coelacanth cDNA data, RNA-seq data generated from *L. chalumnae* muscle (18 Gb of paired end reads were assembled by Trinity¹⁵, Supplementary Figure 2) as well as orthology with other vertebrates. This pipeline produced 19,033 protein coding genes containing 21,817 transcripts. The MAKER pipeline used the *L. chalumnae* Ensembl gene set, Uniprot protein alignments, and *L. chalumnae* (muscle) and *L. menadoensis* (liver and testis)¹⁶ RNA-seq to create gene models, yielding 29,237 protein coding gene annotations. As Ensembl's gene predictions are more conservative and MAKER's are more generous, we believe that the actual number of genes lies in-between these two estimates. In addition, 2,894 short non-coding RNAs, 1,214 lncRNAs and more than 24,000 conserved RNA secondary structures were identified (Supplementary Note 2, Supplementary Tables 3-4, Supp Data 1-3, Supplementary Figure 3). 336 genes were inferred to have undergone specific duplications in the coelacanth lineage (Supplementary Note 3, Supplementary Tables 5-6, Supp Data 4).

Determining the closest living fish relative of the tetrapod ancestor

The question of which living fish is the closest relative to 'the fish that first crawled up on land' has also long captured our imagination: between comparative morphologists and paleontologists the odds have been placed on either the lungfish or the coelacanth^{17,18}. Analyses of small amounts of sequence data for this important phylogenetic question (ranging from 1 to 43 genes) has tended to favor the lungfishes as the extant sister group to the land vertebrates^{19,20}, however, the alternative hypothesis that lungfish and coelacanth are equally closely related to the tetrapods could not be rejected with previous data sets²¹.

To seek a comprehensive answer we generated RNA-seq data from three samples (brain, gonad/kidney, gut/liver) from the West African lungfish, *Protopterus annectens*, and compared it to gene sets from 21 strategically chosen jawed vertebrate species. To perform a reliable analysis we selected 251 genes where 1-1 orthology was clear and used CAT-GTR, a complex site-heterogeneous model of sequence evolution known to reduce tree reconstruction artefacts^{22,23} (see Methods). The resulting phylogeny (**Figure 1**, PP=1.0 for the lungfish-tetrapod node) is fully resolved except for the relative positions of armadillo and elephant. It corroborates known vertebrate phylogenetic relationships and strongly supports the conclusion that tetrapods are more closely related to lungfish than to the coelacanth (Supplementary Note 4, Supplementary Figure 4).

How slowly evolving is the coelacanth?

The morphological resemblance of the modern coelacanth to its fossil ancestors has resulted in it being nicknamed ‘the living fossil’^{1,24}. This begs the interesting question: Is the genome of the coelacanth as slowly evolving as its outward appearance? Earlier work found that a few gene families, such as Hox and protocadherins, showed comparatively slower protein-coding evolution in coelacanth than in other vertebrate lineages^{9,10}. However, these genes may not be representative as the Hox genes are known to be highly conserved.

To address this question, we examined several types of genomic changes in the coelacanth compared to other vertebrates. Protein-coding gene evolution was examined using the 251 concatenated protein phylogenomics dataset (**Figure 1**). Pair-wise distances between taxa were calculated from the branch lengths of the tree using the Two-Cluster test proposed by Takezaki et al.²⁵ to test for equality of average substitution rates. Then, for each of the following species and species clusters (coelacanth, lungfish, chicken and mammals), we ascertained their respective mean distance to an outgroup consisting of three cartilaginous fishes (elephant shark, little skate and spotted catshark). Finally, we tested whether there was any significant difference in distance to the outgroup of cartilaginous fish for every pair of species and species clusters, using a Z-statistic. When these distances to the outgroup of cartilaginous fish were compared, we found that the coelacanth proteins tested were significantly more slowly evolving (0.890 substitutions/site) than the lungfish (1.05 substitutions/site), chicken (1.09 substitutions/site) and mammalian (1.21 substitutions/site) orthologues (Supp Data 5), in all cases with p-values $<10^{-6}$. And as can be seen in **Figure 1**, the substitution rate in coelacanth is approximately half that in tetrapods since the two lineages diverged. In addition, lungfish is also significantly more slowly evolving than the chicken and mammals, with p-values <0.0001 . A Tajima relative rate test²⁶ confirmed the coelacanth’s

significantly slower rate of protein evolution (Supp Data 6). As can be seen in **Figure 1**, the substitution rate observed on the coelacanth lineage is approximately half that of tetrapods. Because branch lengths may be underestimated in regions of a tree that have few species, here potentially confounding the analysis of the coelacanth branch, we examined the node-density effect^{27,28} in each tree of the Bayesian posterior distribution but found no evidence for this artifact.

Secondly, we examined the abundance of transposable elements (TEs) in the coelacanth genome. Theoretically, TEs might contribute most significantly to the evolution of a species by generating templates for exaptation to form novel regulatory elements and exons, and by acting as substrates for genomic rearrangement²⁹. We found that the coelacanth genome contains a wide variety of TE superfamilies and has a relatively high TE content (25%); this number is likely an underestimate due to the draft nature of the assembly (Supplementary Note 5, Supplementary Tables 7-10). Analysis of RNA-seq data and of the divergence of individual TE copies from consensus sequences show that 14 coelacanth TE super-families are currently active (Supplementary Note 6, Supplementary Table 10, Supplementary Figure 5). We conclude that the current coelacanth genome shows both an abundance and activity of TEs similar to many other genomes. This is in contrast to the slow protein evolution observed.

Analyses of chromosomal breakpoints in coelacanth genome and tetrapod genomes reveal extensive conservation of synteny and indicate that large-scale rearrangements have occurred at a generally low rate in the coelacanth lineage. (The coelacanth assembly provides the sensitivity necessary to detect fusions and other intrachromosomal rearrangements in the coelacanth lineage, and fissions in the other tetrapod lineages, but is less sensitive to other types of rearrangement (Supplementary Methods). Analyses of these rearrangement classes detected several previously published fission events that are known to have occurred in tetrapod lineages and at least 31 interchromosomal rearrangements that occurred in the coelacanth lineage or the early tetrapod lineage (0.063 fusions/million years), compared to 20 events (0.054 fusions/million years) in the salamander lineage and 21 events (0.057 fusions/million years) in the *Xenopus* lineage³⁰ (Supplementary Note 7, Supplementary Figure 6). Overall, these analyses indicate that karyotypic evolution in the coelacanth lineage has occurred at a relatively slow rate, similar to that of non-mammalian tetrapods³¹.

In a separate analysis we also examined the evolutionary divergence between the two species of coelacanth, *L. chalumnae* and *L. menadoensis*, found in African and Indonesian waters respectively. Previous analysis of mitochondrial DNA showed a sequence identity of 96%, but estimated divergence times range widely from 6 to 40 million years^{32,33}. When we compared the liver and

testis transcriptomes of *L. menadoensis*³⁴ to the *L. chalumnae* genome, we found an identity of 99.73% (Supplementary Note 8, Supplementary Figure 7), whereas alignments between 20 sequenced *L. menadoensis* BACs and the *L. chalumnae* genome showed an identity of 98.7% (Supplementary Table 11, Supplementary Figure 8). Both the genic and genomic divergence rates are similar to those seen between the human and chimpanzee genomes (99.5% and 98.8% respectively, divergence time 6-8 million years ago)³⁵, while the rates of molecular evolution in *Latimeria* are likely affected by multiple factors including the slower substitution rate seen in coelacanth, thereby suggesting a slightly larger divergence time for the two coelacanth species.

Vertebrate adaptation to land: clues from the coelacanth genome

As the sequenced genome closest to our most recent aquatic ancestor, the coelacanth provides a unique opportunity to identify genomic changes that were associated with the successful adaptation of vertebrates to an important new environment – land. However, given the draft status of the coelacanth genome assembly, it is most informative for detecting gene loss events associated with this transition, though both gene gain and loss in the tetrapod lineage are equally interesting. Therefore, here we examine which genes were lost and which conserved non-coding elements (CNEs) were gained in the tetrapod lineage, and discuss a specific adaptation: the emergence of the autopod (hand and wrist) from the fish fin.

Over the 400 MY interval that vertebrates have lived on land, genes that are unnecessary for existence in their new environment would have been eliminated. To understand this aspect of the water-to-land transition, we surveyed the *Latimeria* genome annotations to identify genes that were present in the last common ancestor of all bony fish (including coelacanth) but that are missing from tetrapod genomes. More than 50 such genes including components of the Fgf signaling, TGF-beta/Bmp signaling, and Wnt signaling pathways, as well as many transcription factor genes, were determined to be lost (Supp Data 7, Supplementary Figure 9). Previous studies of genes lost in this transition could only compare teleost fish to tetrapods, meaning that differences in gene content could have been due to loss in the tetrapod or in the lobe-finned fish lineages. We were able to confirm that four genes previously shown to be absent in tetrapods (*Actinodin* genes³⁶, *Fgf24*³⁷, *Asip2*³⁸), were indeed present and intact in *Latimeria*, supporting their loss in the tetrapod lineage. However, for 85% of the lost genes at least one other vertebrate paralog has been retained in tetrapods, potentially compensating for the gene losses.

We functionally annotated the >50 genes lost in tetrapods using zebrafish data (gene expression, knock-downs and knock-outs). Many genes were classified in important developmental categories: Fin development (13 genes), otolith and ear development (8 genes), kidney development (7 genes), trunk/somite/tail development (11 genes), eye (13 genes), and brain development (23 genes). This implies that critical characters in the morphological transition from water to land (fin-to-limb transition, remodelling of the ear, etc.) are reflected in the loss of specific genes along the phylogenetic branch leading to tetrapods. However, homeobox genes, which are responsible for the development of an organism's basic body plan, show only slight differences between *Latimeria*, ray-finned fish and tetrapods; it would appear that the protein-coding portion of this gene family, along with several others (Supplementary Note 9, Supplementary Tables 12-16, Supplementary Figure 10), have remained largely conserved during the vertebrate land transition. Although some of the lost genes are found in close proximity with each other in the genomes of coelacanth and zebrafish, most of the genes lost in tetrapods appear to have been lost individually rather than in large contiguous blocks (Supplementary Figure 11).

As vertebrates transitioned to a new land environment, changes occurred not only in gene content, but also in the regulation of existing genes. Regulatory changes have been shown to predominate in parallel adaptation in other vertebrates, such as sticklebacks³⁹ and are widely implicated as major facilitators of evolutionary change in a broader context. Conserved non-coding elements (CNEs) are strong candidates for gene regulatory elements and can act as promoters, enhancers, repressors and insulators^{40,41}. They can be computationally predicted by comparing related genome sequences. To identify CNEs that originated in the most recent common ancestor of tetrapods, we predicted CNEs that evolved in various bony vertebrate (i.e., ray-finned fish, coelacanth and tetrapod) lineages and assigned them to their likely branch points of origin. To detect CNEs, conserved sequences in the human genome were identified using MULTIZ alignments of bony vertebrate genomes, and then known protein-coding sequences, UTRs and known RNA genes were excluded. Our analysis identified 44,200 ancestral tetrapod CNEs that originated after the divergence of the coelacanth lineage. They represent 6% of CNEs that are under constraint in the bony vertebrate lineage. We compared the ancestral tetrapod CNEs to mouse embryo ChIP-seq data obtained using antibodies against p300, a transcriptional co-activator. This resulted in a 7-fold enrichment in the p300 binding sites for our candidate CNEs and confirmed that these CNEs are indeed enriched for gene regulatory elements.

Each tetrapod CNE was assigned to the gene it was physically closest to in the human genome and GO category enrichment was calculated for those genes. The most enriched categories were involved with smell perception (sensory perception of smell, detection of chemical stimulus,

olfactory receptor activity etc.). This is consistent with the notable expansion of olfactory receptor family genes in tetrapods compared with teleosts, and may reflect the necessity of a more tightly regulated, larger and more diverse repertoire of olfactory receptors for detecting airborne odorants as part of the terrestrial lifestyle. Other significant categories include morphogenesis (i. e., radial pattern formation, hind limb morphogenesis, kidney morphogenesis) and cell differentiation (including endothelial cell fate commitment and epithelial cell fate commitment), which is consistent with the body plan changes required for land transition, as well as immunoglobulin VDJ recombination, which reflects the presumed response differences required to address the novel pathogens that vertebrates would encounter on land (Supplementary Note 10, Supplementary Tables 17-24).

A major innovation of tetrapods is the evolution of limbs characterised by digits. The limb skeleton consists of a stylopod (humerus or femur), the zeugopod (radius/ulna and tibia/fibula), and an autopod (wrist/ankle and digits). There are two major hypotheses about the origins of the autopod – either it was a novel feature of tetrapods, or it has antecedents in the fins of fish^{42,43} (Supplementary Note 11, Supplementary Figure 12). We examine here the Hox regulation of limb development in ray-finned fish, coelacanth, and tetrapods to address these hypotheses.

In mouse, late phase digit enhancers are located in a gene desert located proximal to the HoxD cluster⁴⁴. Here we provide an alignment of the HoxD centromeric gene desert of coelacanth with tetrapods and ray-finned fishes (**Figure 2a**). Among the six cis-regulatory sequences previously identified in this gene desert⁴⁴, three sequences show sequence conservation restricted to tetrapods (Supplementary Figure 13). However, one regulatory sequence (Island 1) is shared between tetrapods and coelacanth, but not with ray-finned fish (**Figure 2bc**). When tested in a transient transgenic assay in mouse, the coelacanth sequence of Island 1 was able to drive reporter expression in a limb specific pattern (**Figure 2d**), making it likely that Island 1 was a lobe-fin developmental enhancer in the fish ancestor of tetrapods, that was then coopted into the autopod enhancer of modern tetrapods. In this case, the autopod developmental regulation was derived from an ancestral lobe-finned fish regulatory element. Further functional studies of Island 1 and other potential coelacanth cis-regulatory elements in this gene desert may provide insight into the evolution of HoxD regulation in appendages and the evolution of digits in tetrapods.

Evidence for selection in the urea cycle during the evolution of tetrapods

Changes in the urea cycle constitute an illuminating example of the adaptations associated with transition to land. Excretion of nitrogen is a major physiological challenge for terrestrial vertebrates. In aquatic environments, the primary nitrogenous waste product is ammonia, which is readily diluted by surrounding water before it reaches toxic levels, but on land, less toxic substances such as urea or uric acid must be produced instead (Supplementary Figure 14). The widespread and almost exclusive occurrence of urea excretion in amphibians, some turtles and mammals has led to the hypothesis that the use of urea as the main nitrogenous waste product was a key innovation in the vertebrate transition from water to land⁴⁵.

With the availability of gene sequences from coelacanth and lungfish, it becomes possible to test this hypothesis. We used a branch-site model in the HYPHY package⁴⁶, which estimates dN/dS (ω) values among different branches and among different sites (codons) across a multiple species sequence alignment. For the rate-limiting enzyme of the hepatic urea cycle, carbamoyl phosphate synthase I (CPS1), only one branch of the tree shows a strong signature of selection ($p = 0.02$), namely the branch leading to tetrapods and the branch leading to amniotes (**Figure 3**); no other enzymes in this cycle showed a signature of selection. Conversely, mitochondrial arginase (ARG2), which produces extrahepatic urea as a byproduct of arginine metabolism but which is not involved in the production of urea for nitrogenous waste disposal, did not show any evidence of selection in vertebrates (Supplementary Figure 15). This leads us to conclude that adaptive evolution occurred in the hepatic urea cycle during the vertebrate land transition. In addition, it is interesting to note that of the five amino acids of CPS1 that changed between coelacanth and tetrapods, three are in important domains (ATP-A site, ATP-B site, subunit interaction domain) and a fourth is known to cause a malfunctioning enzyme in human patients if mutated⁴⁷.

The coelacanth and placental evolution

The adaptation to a terrestrial lifestyle necessitated major changes in the physiological milieu of the developing embryo and fetus, resulting in the evolution and specialization of extraembryonic membranes of the amniote mammals⁴⁸. This acquisition is considered to be a major evolutionary innovation. The placenta, in particular, is a complex structure that develops from fusion of the allantois and chorion and is critical for providing gas and nutrient exchange between mother and fetus throughout the extended gestation of eutherian mammals, and is also a major site of hematopoiesis⁴⁹.

We have identified a region of the coelacanth HOX-A cluster that may have been involved in the evolution of extraembryonic structures in tetrapods, including the eutherian placenta. Global alignment of the coelacanth *Hoxa14-a13* region with the homologous regions of the horn shark, chicken, human and mouse yielded a potential conserved noncoding element (CNE) just upstream of the coelacanth *Hoxa14* gene (Supplementary Figure 16a, arrow). This conserved stretch is not found in teleost fishes but is highly conserved among horn shark, chicken, human and mouse despite the fact that the latter three have no *Hoxa14* orthologues, whereas the horn shark *Hoxa14* gene has become a pseudogene. This conserved region, HA14E1, corresponds to the proximal promoter-enhancer region of the *Hoxa14* gene in *Latimeria*. HA14E1 is >99% identical between mouse and human and all other sequenced mammals, and would thus be considered an ultraconserved element⁵⁰. The high level of conservation suggests that this element, which already possessed promoter activity, may have been coopted for other functions despite the loss of the *Hoxa14* gene in amniotes. The genomic landscape of the HA14E1 region confirms its conservation and inferred transcriptional activity (Supplementary Figure 16bc). Surprisingly, expression of human HA14E1 in a mouse transient transgenic assay did not give notable expression in the embryo *proper* at day 11.5⁵¹, which was unexpected since its location would predict that it would regulate axial structures caudally⁵². To validate this result, the chicken HA14E1 was used in similar enhancer expression assays in chick embryos. This experiment confirmed the lack of activity in the AP-axis; however, stunning expression was observed in the regions peripheral to the embryo *proper*, i.e., in the embryonically-derived *area vasculosa* of the chick embryo (**Figure 4a**). This extraembryonic region is where blood islands emerge that contribute to the developing vasculature of the chick embryo⁵³. Examination of a *Latimeria* BAC *Hoxa14*-reporter transgene in mouse embryos showed that the *Hoxa14* gene is specifically expressed in a subset of cells in an extraembryonic region at E8.5 (**Figure 4b**).

These findings suggest that the HA14E1 region may have been evolutionarily recruited to coordinate regulation of posterior HoxA genes (*Hoxa13*, *Hoxa11* and *Hoxa10*), which are known to be expressed in the mouse allantois and are critical for early formation of the mammalian placenta⁵⁴. Although *Latimeria* does not possess a placenta, it is a livebearer and has very large, vascularised eggs, but the relationship of *Hoxa14*, the HA14E1 enhancer, and blood island formation in the coelacanth remains unknown. Once again, the coelacanth genome has provided us a window into the formation of evolutionary innovations in the tetrapods.

Coelacanth lacks IgM

Immunoglobulin M (IgM), a class of antibodies, has been reported in all vertebrate species thus far characterised and is considered to be indispensable for adaptive immunity⁵⁵. Interestingly, IgM genes cannot be found in coelacanth despite an exhaustive search of the coelacanth sequence data, and even though all other major components of the immune system are present (Supplementary Note 12, Supplementary Figure 17). Instead, we found two IgW genes (Supplementary Figures 18-20), immunoglobulin genes only found in lungfish and cartilaginous fish and which are believed to have originated in the ancestor of jawed vertebrates⁵⁶ and to have been subsequently lost in teleosts and tetrapods. IgM was similarly absent from the *Latimeria* RNA-seq data, although both IgW genes were found as transcripts. To further characterise the apparent absence of IgM, we exhaustively screened large genomic *L. menadoensis* libraries using numerous strategies and probes and also performed PCR with degenerate primers that should universally amplify IgM sequences. The lack of IgM in *Latimeria* raises questions as to how coelacanth B cells respond to microbial pathogens and whether the IgW molecules can serve a compensatory function, even though there is no indication that the coelacanth IgW was derived from vertebrate IgM genes.

Discussion

Ever since its discovery, the coelacanth has been referred to as a ‘living fossil’ due to its morphological similarities to its fossil ancestors¹. However, questions have remained as to whether it truly is slowly evolving, as morphological stasis does not necessarily imply genomic stasis. In this study, we determined that *L. chalumnae*’s protein-coding genes are significantly more slowly evolving than those of other sequenced vertebrates. Nevertheless, its genome as a whole does not show evidence of unusually low rates of evolution – as evidenced by our analysis of TEs and of large-scale genomic rearrangement. The *L. chalumnae* genome has not shown a lower overall rate of evolution despite its decreased substitution rate in protein-coding genes. The reason for this lower substitution rate is still unknown, although a static habitat and a lack of predation over evolutionary timescales could be contributing factors to a lower need for adaptation.

A closer examination of gene families that show either unusually high or low levels of directional selection indicative of adaptation in the coelacanth, could tell us a great deal about which selective pressures, or lack thereof, shaped this evolutionary relict (Supplementary Note 13, Supplementary Figure 21).

The vertebrate land transition is one of the most important steps in our evolutionary history. The analysis presented here shows conclusively that the closest living fish to the tetrapod ancestor is the lungfish, not the coelacanth. However, the coelacanth is critical for our understanding of this transition, as the lungfish have intractable genome sizes (estimated at 50-100 Gb)⁵⁷. We have already learned a great deal about our adaptation to land through coelacanth whole genome analysis, and we have shown the promise of focused analysis of specific gene families involved in this process. Still, further study of the changes in limb morphology and locomotion, breathing, renal physiology, respiration and immunity between tetrapods and the coelacanth will undoubtedly yield important insights as to how a complex organism like a vertebrate can so drastically change its way of life.

Methods: Appear in the online supplement.

Acknowledgments

Acquisition and storage of *Latimeria chalumnae* samples was supported by grants from the African Coelacanth Ecosystem Programme of the South African National Department of Science and Technology. Generation of the *Latimeria chalumnae* and *Protopterus annectens* sequence by Broad Institute of MIT and Harvard was supported by grants from the National Human Genome Research Institute (NHGRI). KLT is the recipient of a EURYI award from the ESF. We would also like to thank the Genomics Sequencing Platform of the Broad Institute for sequencing the *L. chalumnae* genome and *L. chalumnae* and *P. annectens* transcriptomes, Said Ahamada, Robin Stobbs and the Association pour le Protection de Gombesa (APG) for their help in obtaining coelacanth samples, Yu Zhao for the use of data from *Rana chensinensis*, and Leslie Gaffney, Catherine Hamilton and John Westlund for assistance with figure preparation.

Author contributions

JA, CTA, AM and KLT planned and oversaw the project. RD and CTA provided blood and tissues for sequencing. ML prepared the DNA for sequencing. IM, SG, DP, FJR, TS and DJ assembled the genome. LF and JL made the *L. chalumnae* RNA-seq library. AC, MB, MAB, MF, FB, GS, AMF, AP, MG, GDM, JT-M and EO sequenced and analyzed the *L. menadoensis* RNA-seq library. BA, SMJS, SW, MC and MY annotated the genome. WH and CPP performed the lncRNAs annotation and analysis. PFS, SH, AN, HT, and SJP annotated ncRNAs. MG, GDM, AP, MR and CTA compared *L. chalumnae* and *L. menadoensis* sequence. HB, DB and HP performed the phylogenomic analysis. TMa and AM performed the gene relative rate analysis. DC, SF, OS, J-NV, MS and AM analysed transposable elements. JJS analysed large scale rearrangements in vertebrate genomes. IB and JP analysed genes lost in tetrapods. TMi analyzed actinodin and pectoral fin musculature. CO and MS analysed selection in urea cycle genes. AL and BV performed the conserved noncoding element analysis. IS, NR, VR, NS and CT performed the analysis of autopodial CNEs. KS, TS-S and CTA examined the evolution of a placenta-related CNE. NRS, GWL, MGM, TO and CTA performed the IgM analysis. JA, CTA, AM and KLT wrote the paper with input from other authors.

Figure

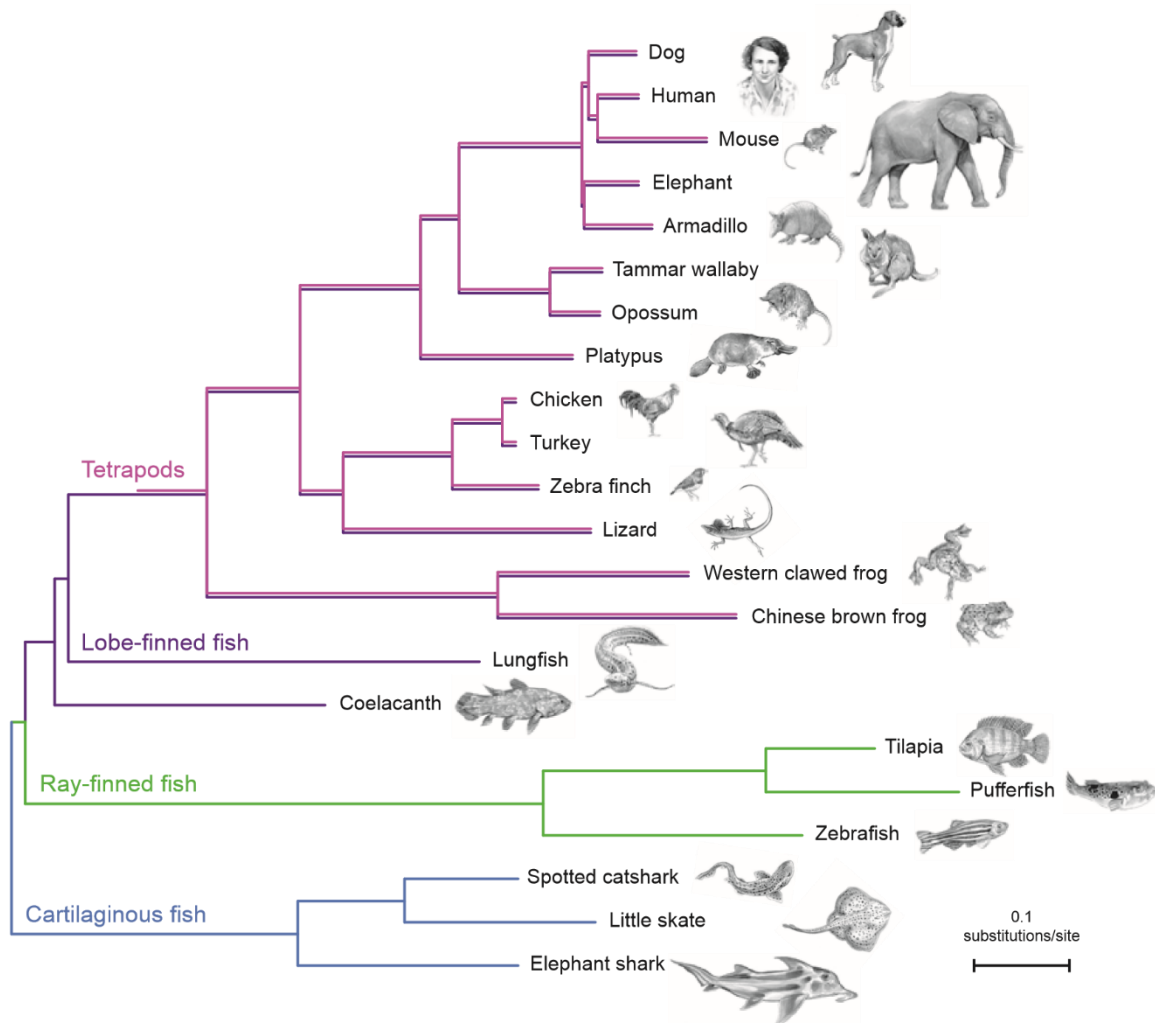


Figure 1. A phylogenetic tree of a broad selection of jawed vertebrates shows that lungfish, not coelacanth, is the closest relative of tetrapods. Multiple sequence alignments of 251 genes present as 1-to-1 orthologs in 22 vertebrates and with a full sequence coverage for both lungfish and coelacanth were used to generate a concatenated matrix of 100,583 unambiguously aligned amino acid positions. The Bayesian tree was inferred using PhyloBayes under the CAT+GTR+ Γ_4 model with confidence estimates derived from 100 jackknife tests (1.0 posterior probability)⁵⁸. The tree was rooted on cartilaginous fish, which are considered to be the outgroup to bony fish and tetrapods. It shows both that lungfish is more closely related to tetrapods than coelacanth and that the protein sequence of coelacanth is slowly evolving.

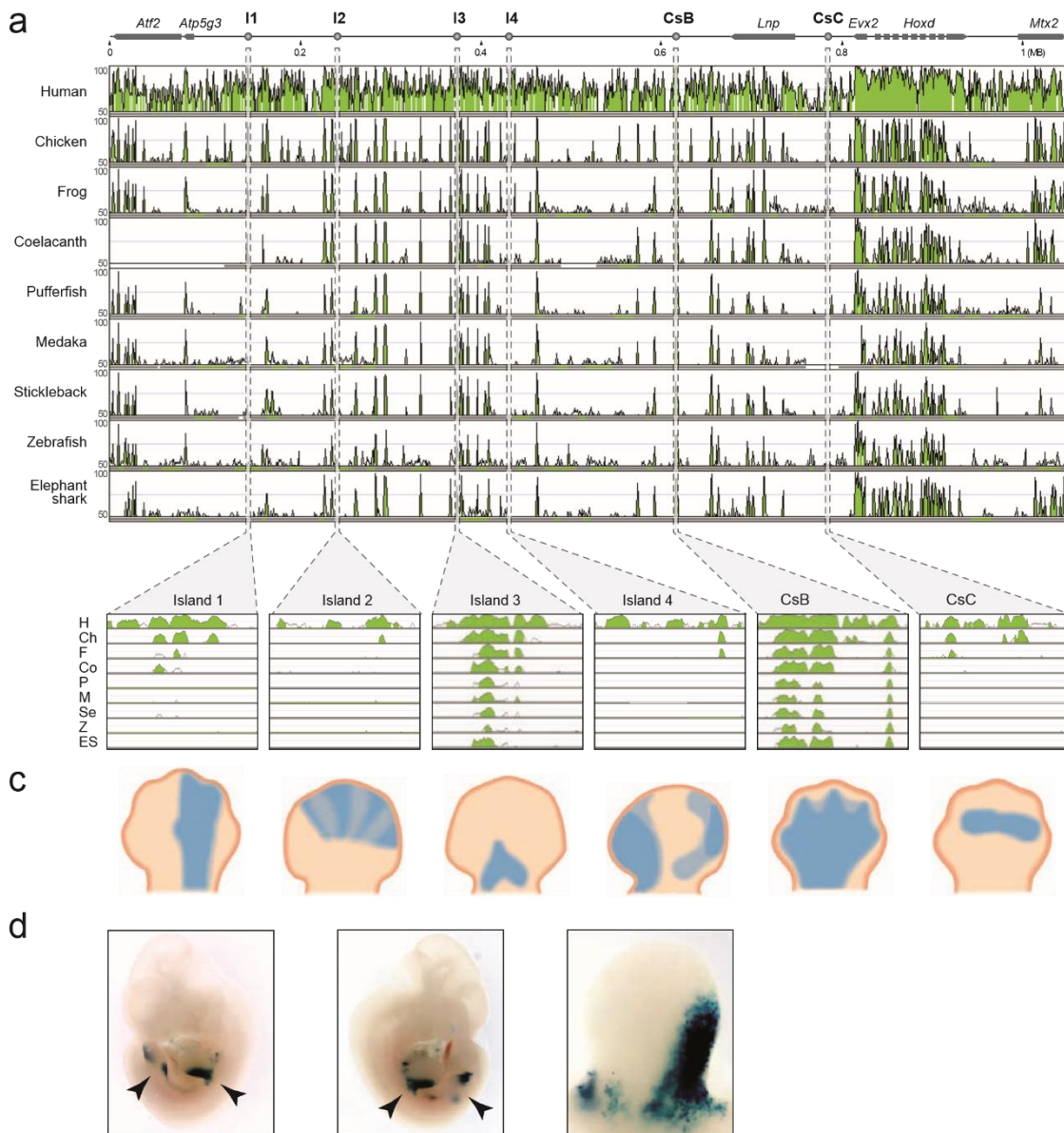


Figure 2. Alignment of the HoxD locus and upstream gene desert identifies conserved limb enhancers. (a) Organization of the mouse HoxD locus and centromeric gene desert, flanked by the ATF2 and MTX2 genes. Limb regulatory sequences (I1, I2, I3, I4, CsB and CsC) are noted. Using the mouse locus as a reference, corresponding sequences from human, chicken, frog, coelacanth, pufferfish, medaka, stickleback, zebrafish and elephant shark were aligned. Alignment (mVISTA program, homology threshold 70%) shows regions of homology between tetrapod, coelacanth and ray-finned fishes. (b) Alignment of vertebrate cis-regulatory elements I1, I2, I3, I4, CsB and CsC. (c) Expression patterns driven by each regulatory element assayed via mouse transgenesis. (d) Expression patterns of coelacanth Island I in a transgenic mouse. Limb buds indicated by arrowheads in the first two panels. The third panel shows a close-up of a limb bud.

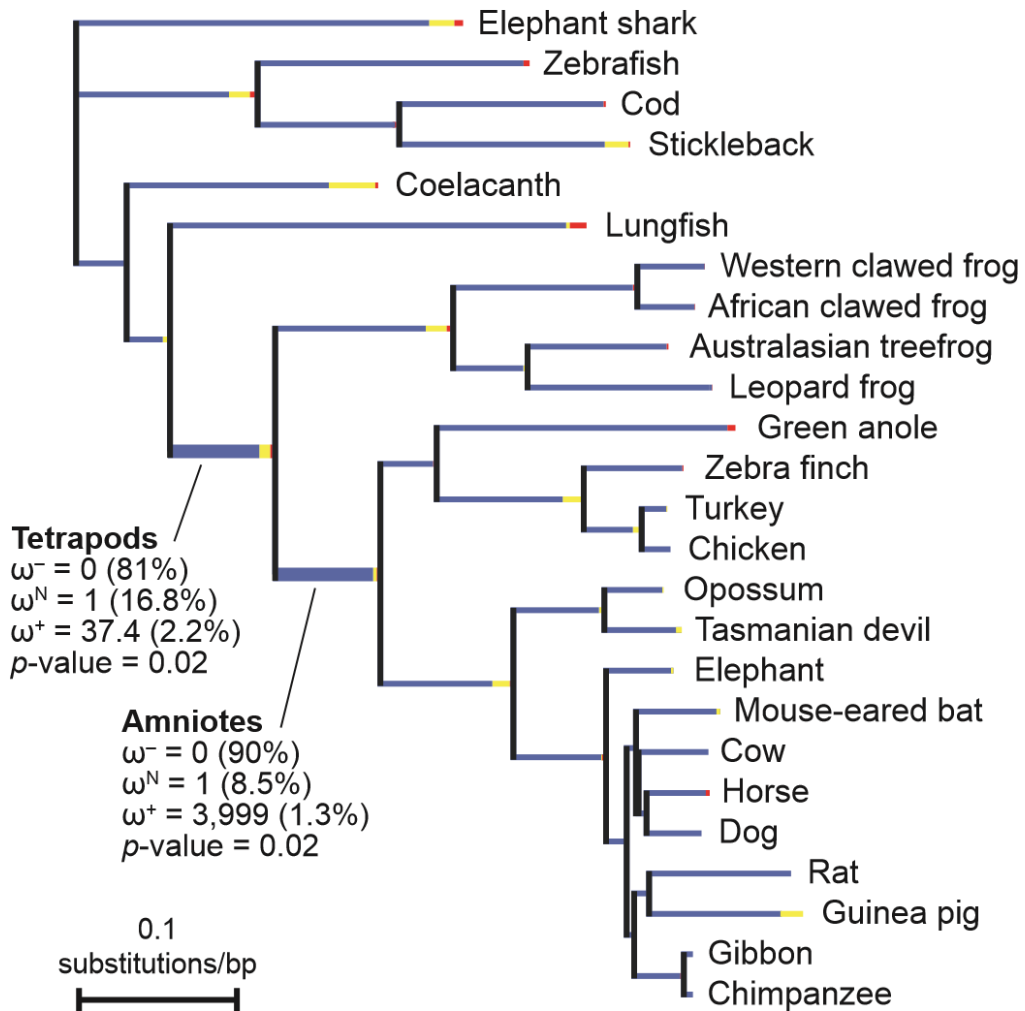


Figure 3. Phylogeny of *CPS1* coding sequences used to determine positive selection within the urea cycle. Branch lengths are scaled to the expected number of substitutions/nucleotide and branch color indicates the strength of selection (dN/dS or ω) with red corresponding to positive or diversifying selection ($\omega > 1$), blue to purifying selection ($\omega = 0$), and yellow to neutral evolution ($\omega = 1$). Thick branches indicate statistical support for evolution under episodic diversifying selection. The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection (shown as percentages for tetrapods and amniotes).

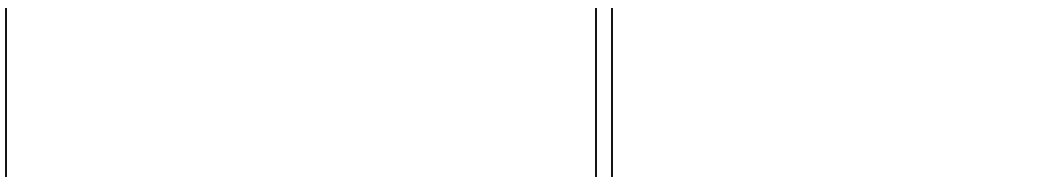
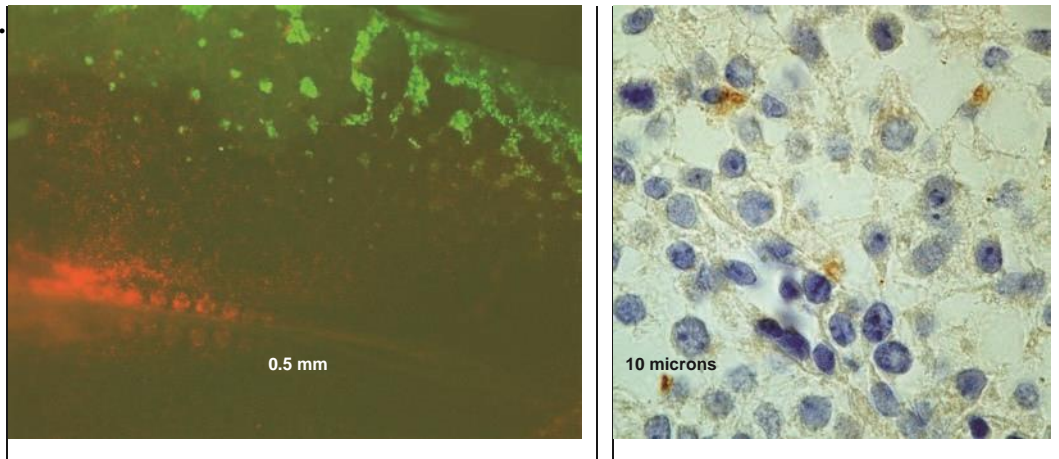


Figure 4.



Transgenic analysis implicates involvement of *Hox* CNE HA14E1 in extraembryonic activities in the chick and mouse. (A) Chicken HA14E1 drives reporter expression in blood islands in chick embryos. A construct containing chicken HA14E1 upstream of a minimal (TK) promoter driving eGFP was linearized and injected and electroporated in HH4 stage chick embryos together with a nuclear mCherry construct. GFP expression was analyzed at stage ~ HH11. The green aggregations and punctate staining are observed in the blood islands and developing vasculature. (B) Expression of *Latimeria Hoxa14* reporter transgene in the developing placental labyrinth of a mouse embryo. A field of cells from the labyrinth region of an E8.5 embryo from a BAC transgenic line containing coelacanth *Hoxa14-Hoxa9*⁵⁹ in which the *Hoxa14* gene had been supplanted with the gene for red fluorescence protein (RFP). Immunohistochemistry was used to detect RFP (brown staining in a small number of cells).

References

- Smith, J. L. B. A Living Fish of Mesozoic Type. *Nature* 143, 455-456, doi:10.1038/143455a0 (1939).
- Nulens, R., Scott, L. & Herbin, M. An Updated Inventory of All Known Specimens of the Coelacanth, *Latimeria* Spp: By Rik Nulens, Lucy Scott and Marc Herbin. (2010).
- Erdmann, M., Caldwell, R. & Kasim Moosa, M. Indonesian 'king of the sea' discovered. *Nature* 395, 335 (1998).
- Smith, J. L. Old Fourlegs: The story of the coelacanth. (Longmans, Green, 1956).
- McCabe, H. & Wright, J. Tangled tale of a lost, stolen and disputed coelacanth. *Nature* 406, 114, doi:10.1038/35018247 (2000).
- Zhu, M. et al. Earliest known coelacanth skull extends the range of anatomically modern coelacanths to the Early Devonian. *Nat Commun* 3, 772, doi:ncomms1764 [pii] 10.1038/ncomms1764 (2012).
- Zimmer, C. *At the Water's Edge: Fish with Fingers, Whales with Legs, and How Life Came Ashore but Then Went Back to Sea.* (Free Press, 1999).
- Zardoya, R. & Meyer, A. The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (*Latimeria chalumnae*). *Genetics* 146, 995-1010 (1997).
- Amemiya, C. T. et al. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc Natl Acad Sci U S A* 107, 3622-3627, doi:0914312107 [pii] 10.1073/pnas.0914312107 (2010).
- Noonan, J. P. et al. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res* 14, 2397-2405, doi:gr.2972804 [pii] 10.1101/gr.2972804 (2004).
- Larsson, T. A., Larson, E. T. & Larhammar, D. Cloning and sequence analysis of the neuropeptide Y receptors Y5 and Y6 in the coelacanth *Latimeria chalumnae*. *Gen Comp Endocrinol* 150, 337-342, doi:S0016-6480(06)00296-6 [pii] 10.1016/j.ygcen.2006.09.002 (2007).
- Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108, 1513-1518, doi:1017351108 [pii] 10.1073/pnas.1017351108 (2011).
- Bogart, J. P., Balon, E. K. & Bruton, M. N. The chromosomes of the living coelacanth and their remarkable similarity to those of one of the most ancient frogs. *J Hered* 85, 322-325 (1994).

- Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18, 188-196, doi:gr.6743907 [pii]10.1101/gr.6743907 (2008).
- Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29, 644-652, doi:nbt.1883 [pii] 10.1038/nbt.1883 (2011).
- Pallavicini, A. et al. Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. . submitted (2012).
- Schultze, H. P. & Trueb, L. *Origins of the Higher Groups of Tetrapods: Controversy and Consensus*. (Comstock Pub. Associates, 1991).
- Bemis, W. E., Burggren, W. W. & Kemp, N. E. The biology and evolution of lungfishes. *Suppl. J. of Morphology* 1 (1987).
- Meyer, A. & Wilson, A. C. Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. *J Mol Evol* 31, 359-364 (1990).
- Meyer, A. & Dolven, S. I. Molecules, fossils, and the origin of tetrapods. *J Mol Evol* 35, 102-113 (1992).
- Brinkmann, H., Venkatesh, B., Brenner, S. & Meyer, A. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proc Natl Acad Sci U S A* 101, 4900-4905, doi:10.1073/pnas.04006091010400609101 [pii] (2004).
- Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21, 1095-1109, doi:10.1093/molbev/msh112 msh112 [pii] (2004).
- Philippe, H. et al. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* 470, 255-258, doi:nature09676 [pii] 10.1038/nature09676 (2011).
- Thomson, K. S. *Living Fossil: The Story of the Coelacanth*. (W.W. Norton Press, 1991).
- Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol* 12, 823-833 (1995).
- Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135, 599-607 (1993).
- Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst Biol* 55, 637-643 (2006).

- Webster, A. J., Payne, R. J. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* 301, 478, doi:10.1126/science.1083202 301/5632/478 [pii] (2003).
- Bejerano, G. et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441, 87-90, doi:nature04696 [pii] 10.1038/nature04696 (2006).
- Voss, S. R. et al. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Res* 21, 1306-1312, doi:gr.116491.110 [pii] 10.1101/gr.116491.110 (2011).
- Smith, J. J. & Voss, S. R. Gene order data from a model amphibian (*Ambystoma*): new perspectives on vertebrate genome structure and evolution. *BMC Genomics* 7, 219, doi:1471-2164-7-219 [pii] 10.1186/1471-2164-7-219 (2006).
- Inoue, J. G., Miya, M., Venkatesh, B. & Nishida, M. The mitochondrial genome of Indonesian coelacanth *Latimeria menadoensis* (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. *Gene* 349, 227-235, doi:S0378-1119(05)00017-X [pii] 10.1016/j.gene.2005.01.008 (2005).
- Holder, M. T., Erdmann, M. V., Wilcox, T. P., Caldwell, R. L. & Hillis, D. M. Two living species of coelacanths? *Proc Natl Acad Sci U S A* 96, 12616-12620 (1999).
- Canapa, A. et al. Composition and Phylogenetic Analysis of Vitellogenin Coding Sequences in the Indonesian Coelacanth *Latimeria menadoensis*. *J Exp Zool B Mol Dev Evol* 318, 404-416, doi:10.1002/jez.b.22455 (2012).
- Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69-87, doi:nature04072 [pii] 10.1038/nature04072 (2005).
- Zhang, J. et al. Loss of fish actinotrichia proteins and the fin-to-limb transition. *Nature* 466, 234-237, doi:10.1038/nature09137 (2010).
- Jovelin, R. et al. Evolution of developmental regulation in the vertebrate FgfD subfamily. *Journal of experimental zoology. Part B, Molecular and developmental evolution* 314, 33-56, doi:10.1002/jez.b.21307 (2010).
- Braasch, I. & Postlethwait, J. H. The teleost agouti-related protein 2 gene is an ohnolog gone missing from the tetrapod genome. *Proceedings of the National Academy of Sciences of the United States of America* 108, E47-48, doi:10.1073/pnas.1101594108 (2011).

Jones, F. C. et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55-61, doi:nature10944 [pii] 10.1038/nature10944 (2012).

Navratilova, P. et al. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev Biol* 327, 526-540, doi:S0012-1606(08)01320-1 [pii] 10.1016/j.ydbio.2008.10.044 (2009).

Xie, X. et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* 104, 7145-7150, doi:0701811104 [pii] 10.1073/pnas.0701811104 (2007).

Shubin, N., Tabin, C. & Carroll, S. Fossils, genes and the evolution of animal limbs. *Nature* 388, 639-648, doi:10.1038/41710 (1997).

Shubin, N., Tabin, C. & Carroll, S. Deep homology and the origins of evolutionary novelty. *Nature* 457, 818-823, doi:nature07891 [pii] 10.1038/nature07891 (2009).

Montavon, T. et al. A regulatory archipelago controls Hox genes transcription in digits. *Cell* 147, 1132-1145, doi:S0092-8674(11)01273-6 [pii] 10.1016/j.cell.2011.10.023 (2011).

Wright, P. A. Nitrogen excretion: three end products, many physiological roles. *J Exp Biol* 198, 273-281 (1995).

Kosakovsky Pong, S. L. et al. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol* 28, 3033-3043, doi:msr125 [pii] 10.1093/molbev/msr125 (2011).

Haberle, J. et al. Molecular defects in human carbamoyl phosphate synthetase I: mutational spectrum, diagnostic and protein structure considerations. *Hum Mutat* 32, 579-589, doi:10.1002/humu.21406 (2011).

Carroll, R. L. *Vertebrate Paleontology and Evolution*. (W.H. Freeman and Company., 1988).

Gekas, C. et al. Hematopoietic stem cell development in the placenta. *Int J Dev Biol* 54, 1089-1098, doi:103070cg [pii] 10.1387/ijdb.103070cg (2010).

Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* 304, 1321-1325, doi:10.1126/science.1098119 1098119 [pii] (2004).

Vista Enhancer Browser, <http://enhancer.lbl.gov/cgi-bin/imagenodb3.pl?form=presentation&show=1&experiment_id=501&organism_id=1> (

Wellik, D. M. Hox patterning of the vertebrate axial skeleton. *Dev Dyn* 236, 2454-2463, doi:10.1002/dvdy.21286 (2007).

Sheng, G. Primitive and definitive erythropoiesis in the yolk sac: a bird's eye view. *Int J Dev Biol* 54, 1033-1043, doi:103105gs [pii] 10.1387/ijdb.103105gs (2010).

Scotti, M. & Kmita, M. Recruitment of 5' Hoxa genes in the allantois is essential for proper extra-embryonic function in placental mammals. *Development* 139, 731-739, doi:dev.075408 [pii] 10.1242/dev.075408 (2012).

Bengtén, E. et al. Immunoglobulin isotypes: structure, function, and genetics. *Curr Top Microbiol Immunol* 248, 189-219 (2000).

Ota, T., Rast, J. P., Litman, G. W. & Amemiya, C. T. Lineage-restricted retention of a primitive immunoglobulin heavy chain isotype within the Dipnoi reveals an evolutionary paradox. *Proc Natl Acad Sci U S A* 100, 2501-2506, doi:10.1073/pnas.0538029100 0538029100 [pii] (2003).

Gregory, T. R. *The Evolution of the Genome*. (Elsevier Academic Press, Inc., 2004).

Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21, 456-463, doi:bti191 [pii] 10.1093/bioinformatics/bti191 (2005).

Smith, J. J., Sumiyama, K. & Amemiya, C. T. A living fossil in the genome of a living fossil: Harbinger transposons in the coelacanth genome. *Mol Biol Evol* 29, 985-993, doi:msr267 [pii] 10.1093/molbev/msr267 (2012).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Characterization of sex determination and sex differentiation genes in *Latimeria*

Mariko Forconi¹, Adriana Canapa^{1§}, Marco Barucca¹, Maria A Biscotti¹, Francesco Buonocore², Anna M Fausto², Daisy M Makapedua³, Alberto Pallavicini⁴, Marco Gerdol⁴, Gianluca De Moro⁴, Giuseppe Scapigliati², Ettore Olmo^{1*}, Manfred Schartl^{5*}.

Abstract

Genes involved in sex determination and differentiation have been found in mouse, human, chicken, reptiles, amphibians and teleost fish. However, the conservation of their functions is yet largely unexplored and is not clear if there is a common set of genes to all vertebrates. The coelacanths as representatives of basal Sarcopterygians could help to delineate an ancestral inventory of genes involved in this important developmental process and should also give some first insights about components of the sex determination cascade and testes differentiation genes in these unique organisms, representing “living fossils”.

In this study we have identified and characterized 33 genes implicated in sex determination and differentiation from the *L. chalumnae* genome and from liver and testis transcriptomes of *Latimeria menadoensis*, assessing their expression levels in these tissues.

Among the analyzed genes higher interest is covered by *GSDF*, a gene so far known only from teleosts and here for the first time characterized in the sarcopterygian lineage; *FGF9*, a missing gene in teleosts; and *DMRT1* a gene whose expression has recently been linked to sexual identity maintenance in adult gonads.

The gene repertoire and testis specific expression signature of the coelacanth indicate more conserved features with modern fishes and point to unexpected changes in the gene regulatory network governing sexual development.

Introduction

In sexual development two processes can be distinguished: sex determination and sex differentiation. The first is defined as the process that determines whether the primordium of the bipotential gonad will develop into testis or ovary and thus decides over the sexual fate, while the latter takes place once the sex determination decision has been made and comprises the actual development of testes or ovaries from the undifferentiated gonad [1]. Sex determination is considered to be either a default pathway or a suppression of this default development and initiation of the opposite sexual development, while sex differentiation apparently results from the antagonistic relationship among genes influencing testis or ovary development [2,3]. Recently, it emerged that sex specific mechanisms, instrumental to maintain the male or female identity of the testis and ovary, operate even in the adult gonads of mammals [4-6].

In addition to a male- and female-specific development of the gonads, other organs can assume sometimes very elaborate differences as well. In vertebrates these secondary sex characters are generally believed to be exclusively instructed by the developing testes or ovary through sex steroids (with possible exceptions in birds, [7]), whereas in invertebrates it appears as a common rule that each somatic cell has its inherent sexual identity [8].

While in mammals sex steroids play a later role in development, in fish, amphibians, reptiles, birds, and marsupials early sex differentiation is influenced by sex steroids and the proteins involved in their metabolism and binding [9-20].

In vertebrates sex can be determined mainly by two different mechanisms: either sexual development is determined by the genetic constitution of the individual or by the environment e.g. due to the influence of temperature during development, nutrients or pH [21-23].

Many studies on sexual development of mammals have revealed that the consecutive processes of sex determination, gonad differentiation and identity maintenance are brought about by an elaborate network of transcription factor interactions and signalling molecules; a master regulator at the top then triggers the network towards male or female [24]. In most mammals, the Y-chromosomal SRY gene is the male determining gene, but this gene has not been detected outside the placental mammals [25]. In chicken (and possibly all birds), *Dmrt1*, and its homologs, *dmrt1bY* (or *DMY*) in the Japanese ricefish (medaka, *Oryzias latipes*) [26, 27] and *DM-W* in the frog *Xenopus laevis* [28] are the master regulators of sexual development, while Gonadal soma derived factor (*GSDF*) [29], Anti-Müllerian hormone (*AMH*) [30] the Anti-Müllerian hormone receptor (*AMHR2*) [31] or other genes have this function in several other fish species.

In contrast to this diversity of genetic sex determinants at the top, genome-wide searches and homology cloning in teleost fishes, amphibians, reptiles and birds revealed that the “downstream” components of the network are present and -based on spot-check expression studies- appear to have a conserved function. This led to the paradigm that for sex determination during evolution “masters change, slaves remain” [32-34]. It is, however, totally unclear how far back in evolutionary history this holds true; in particular when and how the vertebrate sex regulatory network evolved and whether the genes were repeatedly and independently recruited to this process or represent a conserved ancient mechanism. The unique opportunity to obtain high quality RNA for transcriptome analysis of testis and liver tissues from the Indonesian coelacanth *Latimeria menadoensis*, in combination with the availability of the whole genome sequence of the African coelacanth *L. chalumnae*, allows to get insights from an organism that is considered a nearest living relative of tetrapods.

The genes that emerged from many studies as the key components of the regulatory network of sexual development can be functionally grouped into (1) genes required for the development of the bipotential gonad (Wilm’s tumor suppressor-1 (*WT1*), Steroidogenic factor-1 (*SF-1*), and GATA-binding protein 4 (*GATA-4*)); (2) genes involved in male sex determination (Double sex and mab-3 related transcription factor 1 (*DMRT1*), SRY-related box 9 (*SOX9*), Dosage sensitive sex-reversal-adrenal hypoplasia congenital-critical region of X chromosome, gene 1 (*DAX1*), Fibroblast growth factor 9 (*FGF9*), Desert hedgehog (*DHH*)); (3) genes involved in male sex differentiation (Anti-Müllerian hormone (*AMH*), AMH-receptor2 (*AMHR2*), Androgen receptor (*AR*)); (4) genes involved in female sex determination (Wingless-type MMTV integration site family member 4 (*WNT4*), R-spondin-1 (*RSPO-1*), Catenin β -1 (*CTNNB1*), Forkhead box transcription factor L2 (*FOXL2*), Follistatin (*FST*)); (5) genes involved in female sex differentiation (Aromatase (also named *Cyp19A1* or *P450arom*), Estrogen receptor α (*ER α*), Estrogen receptor β (*ER β*)) (Figure 1).

The above mentioned genes and eleven others (*DMRT3*, *DMRT6*, *GSDF*, Platelet-derived growth factors (*PDGF*) α and β and their receptors (*PDGFR α* , *PDGFR β*), 11 β -hydroxylase (*CYP11B*), and 5 α -reductase 1, 2, and 3 (*SRD5A1*, *SRD5A2*, *SRD5A3*), where evidence of involvement in sex development in a specific group of organisms has been obtained [35-47], were searched in the *L. chalumnae* genome and in the transcriptome of *L. menadoensis*. Further on their expression levels were evaluated in the liver and testis of the adult specimen of the Indonesian coelacanth.

We find that the repertoire and expression profiles of the sex determination and sex differentiation genes in coelacanths are much more similar to that of the modern fish rather than to tetrapods, or even representing an intermediate situation, suggesting that unexpectedly also for gonad development major evolutionary novelties that accompanied the transition to terrestrial life were required.

Methods

The genome of the African coelacanth *L. chalumnae* has recently been sequenced (project accession PRJNA56111) [48] and is available as WGS scaffolds at <http://www.ncbi.nlm.nih.gov> and <http://www.ensembl.org>. The transcriptome of its Indonesian congener, *L. menadoensis*, is described in the work of Pallavicini and colleagues [49] and in Canapa and colleagues [50].

Briefly, a good quality RNA sample was used to generate a cDNA library for transcriptome sequencing on an Illumina Genome Analyzer II platform. After the filtering of high-quality reads, removal of reads containing primer/adaptor sequences, and trimming of read length, the assembly of the Illumina 100 bp paired-end reads was performed on a 4 cores server (72GB RAM). The commercially available CLC Genomics workbench (version 3.7.1, CLC bio, Aarhus, Denmark) and Trinity [51] were used for the *de novo* assembly of short reads. Contigs confirmed and improved by both methods were pooled in a high quality set.

Sampling location, CITES and other information on the Indonesian coelacanth specimen analysed are reported in Makapedua et al. [52].

To identify the coelacanth homologs of genes involved in sexual development, the corresponding *Xenopus tropicalis*, *Gallus gallus*, *Danio rerio* and *Homo sapiens* sequences were BLASTed on the *L. menadoensis* transcript dataset. The identity of each retrieved putative transcript was confirmed through NCBI BLAST by homology. BLASTx analyses allowed to define the completeness of the transcripts (Coding sequences, CDS).

The Indonesian coelacanth sequences were then BLASTed against the WGS dataset of *L. chalumnae*, in order to define the genomic scaffolds of the African coelacanth containing those genes. Divergence across the two species was calculated with PAUP on the matching sequences as p-distance percentage and Ka/Ks ratio was calculated with KaKs_calculator [53] using the Nei and Gojobori method [54]. The synonymous distance was calculated by MEGA5 [55] applying the uncorrected modified Nei and Gojobori method [56] on the concatenated coding sequences aligned with ClustalW2 (<http://www.ebi.ac.uk/Tools/msa/clustalw2/>; [57]).

Predicted transcripts of *L. chalumnae* were collected from ENSEMBL (http://www.ensembl.org/Latimeria_chalumnae/Info/Index). GSDF CDS was manually obtained from the alignment of *L. menadoensis* transcripts to the African coelacanth genome; *FGF9*, not present in the transcriptome and not annotated in ENSEMBL, was manually obtained by BLASTing

annotated amino acidic sequences of other species to the African coelacanth WGS. The two putative *L. chalumnae* transcripts were confirmed by homology through NCBI BLAST.

L. chalumnae and *L. menadoensis* transcripts were compared by ClustalW2 alignment and a graphical representation of each sequence pair is supplied in Figure S1A and Figure S1B.

Gene Ontology (GO) terms concerning sex determination and sex differentiation (GO0007530 and GO0007548, respectively) were selected and *L. menadoensis* orthologs to *D. rerio*, *X. tropicalis*, *G. gallus*, *Canis familiaris*, *Bos taurus*, *Sus scrofa*, *Mus musculus*, *Rattus norvegicus* and *H. sapiens* counterparts were counted.

Gene expression levels in *L. menadoensis* liver and testis were calculated using the CLC Genomic Workbench 4.5.1 (CLC Bio, Katrinebjerg, Denmark) on the basis of mapping paired reads from the transcriptome and are given in Fragments Per Kilobase of exon per Million sequenced fragments (FPKM). The lack of some transcripts in the assembled transcriptome might depend on a scarce expression of genes and thus on the limited number of reads not allowing the assembly of a contig. To determine the absence or the low expression levels of some transcripts in question in the *L. menadoensis* transcriptome, assessment of the FPKM of ENSEMBL transcript predictions was performed on *DMRT3*, *FOXL2*, *Aromatase*, *WNT4*, and *CYP11B*. The FPKM of the inferred sequence of *L. chalumnae FGF9* was also calculated.

Besides genes expected to be involved in sexual development, expression levels of some house-keeping (HK) genes, chosen according to Eisenberg and Levanon [58], were also evaluated. These include *phosphoglycerate kinase*, the *ribosomal proteins S27*, *RPL19*, *RPL11*, *RPL32*, and *HSPCB*.

Phylogenetic analyses were performed in order to check the proper assignment to evolutionary related gene groups. Sequences of SOXE, FGF9/16/20, and TGF- β groups of other vertebrates were retrieved from NCBI protein database and ENSEMBL. Multiple alignments were performed with ClustalW2 using default parameters. Bayesian Inference and Maximum Parsimony were used to build phylogenetic trees. Bayesian Inferences were performed by MrBayes-3.1.2 [59], the amino acidic model applied were Dayhoff [60] for the SOXE and for TGF- β groups, and Jones [61] for FGF9/16/20 group. In the analyses parameters were fixed to 1,000,000 generations, sampling every 100, burn-in was set as 2,500 and the stationarity defined when the average split of standard deviation reached a value lower than 0.009.

Maximum Parsimony analyses were performed using PAUP [62], applying heuristic search with tree bisection-reconnection (TBR) branch swapping and random stepwise additions with 100 replications. 1,000 bootstrap replicates were calculated. Only minimal trees were retained.

In each tree capture the designated outgroup, accession numbers, constant, parsimony informative, and non-informative sites are specified.

Conserved syntenic blocks were inferred from ENSEMBL annotation of putative *CYP11B* (Figure S2), *DMRT1*, *FGF9*, *FGF16*, and *FGF20* flanking regions in some sequenced vertebrate genomes. Sizes and distances of genes were calculated on the basis of annotated coordinates of each element. The scaffolds containing *FGF9* and flanking genes (*EFHA1* and *ZDHHC20*) conserved in tetrapods were identified by homology through tBlastN on *L. chalumnae* WGS data.

Results

The GO analyses of ‘sex determination’ and ‘sex differentiation’ term annotations to the transcriptome of *L. menadoensis* were conducted and compared to selected other vertebrate genomes (Table S1 and Table S2). 25 contigs were identified to be orthologs of a GO0007530 (Sex determination) annotation and 297 contigs were orthologs of GO0007544 (Sex differentiation) annotation.

In this study, thirty-three genes for which substantial evidence of being involved in sex determination and differentiation is available (Supplementary notes) were further analysed in more depth. CDSs were retrieved from the genome of *L. chalumnae* and from both testis and liver transcriptomes of *L. menadoensis* (Table 1 and Table 2), and their expression levels were assessed in these tissues. To confirm the putative orthology status for closely related genes, phylogenetic analyses were performed, and evolutionary relationships were inferred from the topology of the trees. Furthermore the micro-syntenic conservations previously described in other vertebrates for *DMRT1* [36] and *FGF9/16/20* [63, 64] were analyzed in *Latimeria*.

To evaluate whether sequence information from *L. menadoensis* and *L. chalumnae* can be combined, the genetic distance between the two coelacanths was determined comparing the transcripts of *L. menadoensis* to the genomic sequences of *L. chalumnae*. The value, calculated over all matching sequences, ranged between 0% and 0.826%. The divergence is mainly due to mutations, insertions or deletions in untranslated regions (UTRs). Point mutations affecting the coding region of transcripts are in most cases synonymous (Table 1 and Table 2). The synonymous distance calculated over the whole gene set was 0.0019 (standard error 0.0005). These findings allow to combine and investigate data from the two species together.

Genes in male sexual development

Twenty-five genes involved in male sexual development were analysed in *Latimeria*: three containing a Double sex and mab-3 (DM) Domain (*DMRT1*, *DMRT3*, and *DMRT6*); three belonging to the *SOXE* subfamily (*SOX8*, *SOX9*, and *SOX10*) of SRY-related HMG box transcription factors; other transcription factors including *WT1*, *DAX1*, *GATA-4*, *DHH*, *SF-1*; the signalling molecules *PDGF α* and β , *GSDF*, *AMH*, *FGF9*, and *FGF20*; four receptors comprising AR, AMHR2, and PDGFR α and β ; and the steroidogenic enzymes SRD5A1, SRD5A2, SRD5A3 and CYP11B (Table 1).

ENSEMBL prediction recovered in the *L. chalumnae* genome annotation twenty-three of the twenty-five genes. The two missing sequences were inferred manually from the genome assembly: *FGF9* was identified through comparison with orthologous sequences of other species and *GSDF* was identified through the alignment of a *L. menadoensis* transcript to whole genome shotgun (WGS) contigs of *L. chalumnae*. Fifteen of twenty-three *L. chalumnae* predicted transcripts comprised the complete CDS, while eight were partial. The manually inferred *L. chalumnae* *FGF9* covers the complete CDS while the *L. chalumnae* *GSDF* homolog is incomplete (about 75% of the CDS).

The testis and liver transcriptomes of *L. menadoensis* contained twenty-two transcripts. Half of the contigs of the Indonesian coelacanth carried a complete CDS, while the other half was partial or fragmented. Transcripts of three genes, *FGF9*, *CYP11B*, and *DMRT3*, were absent from both analysed tissues (Table 1). The comparison of *L. menadoensis* and *L. chalumnae* male sex development analysed sequences is depicted in Figure S1A.

Thirteen male sex development transcripts showed an expression value lower than one FPKM unit in testis and were considered as not expressed above background (Figure 2A).

With the exception of AR, eleven genes (*DMRT6*, *DMRT1*, *SOX9*, *WT1*, *GSDF*, *AMH*, *SRD5A1*, *DHH*, *SF-1*, *SRD5A3*, and *SOX10*) are higher expressed in testis than liver, but only three of them (*DMRT1*, *DMRT6*, and *SOX9*) have a marked differential expression between the two tissues presenting a FPKM difference higher than 10. In liver seven genes were found to be expressed above background: *SOX9*, *SRD5A1*, AR, *DAX1*, *PDGF α* , *GATA-4*, and *SRD5A2*.

The most highly expressed transcript among the twenty-five analysed male sex development genes is *DMRT6*, reaching in testis a value of 37.79 FPKM and ranking among the 2,000 most abundant transcripts of the over 61,000 contigs recorded in testis. In liver *DMRT6* expression is absent.

DMRT1, one of the most important genes in male development, plays a key function in fish [65, 66], chicken [67, 68], and reptiles [69]. The alignment of the Indonesian coelacanth transcripts to the

African congener genome (Figure 3A) identified the presence of 5 exons, exceeding the ENSEMBL predicted transcript of 1,572 bp at the 3' end (Figure 3B). The DM domain is encoded in the first annotated exon. The long 3'UTR harbours a region of 320 bp containing a low-copy interspersed repeat.

The size of the *DMRT1* gene in the *L. chalumnae* genome covers over 152 Kb (Figure 3A), close to 127 Kb gene in *H. sapiens* (ENSEMBL annotation), but spans a really long range if compared to the 3 Kb gene in *Crocodylus palustris* [70], to the 45 Kb gene in *D. rerio* [71], and 53-58 Kb gene in *G. gallus* ([72], ENSEMBL). Moreover, because of the lack of a 5' UTR (Figure 3B), which is encoded in other fish in the so-called exon 0 [73], in both sequences obtained from the transcriptome and the ENSEMBL prediction, the existence of another exon (which would elongate even more this genomic locus) is likely.

Brunner and colleagues [36] first identified a strict conservation in gene order surrounding the *DMRT1* gene, involving two other DM domain genes, *DMRT2* and *DMRT3*, and the gene *KANK1* (*KIAA0172*). We found that this conserved micro-synteny is also present in the *L. chalumnae* genome by comparing the genomic scaffold JH127237 (1,057,921 bp), from position 608,000 to 941,000, to other vertebrate chromosomes (Figure 3C). Interestingly, this region in *G. gallus* (where *DMRT1* is pivotal in male development) and in *Ornithorhynchus anatinus* is linked to a gonosome (Z and X5, respectively), while in other species of actinopterygian and sarcopterygian lineages it is located on an autosome. To date it was not yet possible to identify the presence of sex chromosomes in the *Latimeria* karyotype [74], nor anchoring the scaffold containing *DMRT1* to a chromosome.

Among the genes analysed, *DMRT1* is the second most highest expressed in testis with 11.84 FPKM units and belongs to the top 10% most expressed transcripts (Figure 2A).

SOX9 is a transcription factor activating *AMH* and together with *DMRT1* inhibiting *WNT4* and *FOXL2*. In mammals it is activated by another SOX family protein, *SRY*, while in other vertebrates it is regulated mainly by SF-1 and *DMRT1*. *SOX9* belongs to subgroup E of SOX proteins together with *SOX8* and *SOX10*. A phylogenetic analysis (Figure 4) of SOX E group proteins, carried out on several vertebrates, resulted in a tree topology that displays three major clades corresponding to the three different genes. In the *SOX9* and *SOX10* clades *Latimeria* sequences comprise a sister group of tetrapods, while the relationships of the coelacanth *SOX8* were not clearly resolved according to its phylogenetic position. *SOX9* and *SOX10* are much higher expressed in testis than liver but *SOX8* shows only a low expression in liver of *L. menadoensis* (Figure 2A; FPKM: 11.60 and 1.38 for *SOX9*, FPKM: 2.25 and 0.04 for *SOX10*).

In mammals FGF9 has an important function in male development creating a positive feedback cycle with *SOX9* and inhibiting the WNT4 pathway in testis [75]. In teleosts *FGF9* was not detected so far but *FGF20b* was proposed to display the same function as the missing gene *FGF9* [63, 64] and therefore to substitute the FGF9 action in sexual development. Interestingly, we found a *FGF9*-like sequence in coelacanths. In order to confirm the orthology relationships of the putative *Latimeria* *FGF9*, *FGF16*, and *FGF20*, sequence comparisons and conserved synteny arrangements of the flanking regions were investigated (Figure 5). In tetrapods both blocks harbouring either *FGF9* or *FGF20* are characterized by an *EFHA* and a *ZDHHC* gene upstream of the *FGF* genes. Downstream of *FGF9*, *16* and *20* different extended gene-deserted regions are noticed. In teleosts, where *FGF9* is absent, the other genes forming the micro-syntenic cluster are distributed over different chromosomes. In *L. chalumnae* the *FGF9* cluster is split on two scaffolds whose co-localization on the same chromosome is not yet possible to define. However, the proximity of a putative coding fragment of *EFHA1* upstream the 5' end of *FGF9* suggests that the coelacanth *FGF9* follows the tetrapod pattern.

The phylogenetic analysis of the FGF9/16/20 group (Figure 6) uncovers three major clades corresponding to the three different genes. The exact position of *L. chalumnae* FGF20 is unsolved. It is, like the *X. laevis* ortholog, paraphyletic to teleosts and tetrapods. FGF16 of coelacanth is basal to the tetrapods as expected. However, the placement of *Latimeria* FGF9, even if firmly nested within the FGF9 tetrapod clade, does not reflect the phylogenetic position in the taxonomic group.

Unexpectedly, neither *FGF9* nor *FGF20* were expressed in testis of *L. menadoensis*.

GSDF is a recently described gene that appears to be critically involved in teleost male development [29, 39, 40, 45]. It has not been found in tetrapods and no sarcopterygian homolog has been described so far. However, our BLAST analyses of teleost *GSDFs* on the *L. menadoensis* transcript database suggested the presence of a putative *GSDF* gene. A confirmation of its identity was performed through BLASTx analysis. Despite low similarity values (29% identity, 49% positive matching with *Oncorhynchus mykiss* GSDF NP_001118051.1, and 28% identity and 50% positive matching with *O. latipes* GSDF NP_001171213.1), Bayesian Inference and Maximum Parsimony reliably assigned the sequence to the GSDF clade of teleosts (Figure 7). For the phylogenetic analysis we included, besides GSDF, two other proteins of the TGF- β family chosen on the basis of their close relationships to GSDF, namely AMH and Inhibin- α [39]. A multiple alignment of the conserved TGF- β domain of the three genes revealed that *L. menadoensis* GSDF is a sister group of teleost GSDFs, with a posterior probability of 100 in the Bayesian Inference analysis and a bootstrap value of 97 in the Maximum Parsimony tree (Figure 8). The lack of a glycine in a cysteine knot, a diagnostic amino acid missing

in the GSDF protein as noticed by Shibata and colleagues [45], further confirms the inclusion of the *L. menadoensis* sequence to the GSDF group, being the first homolog described in the sarcopterygian lineage.

The Indonesian coelacanth *GSDF* was BLASTed on the *L. chalumnae* genome and identified a genomic counterpart partially on contig AFYH01270444 and partially on scaffold JH127632 with a gap of 171 bp among them.

GSDF in *L. menadoensis* is characterized by a remarkably high testis expression but no expression in liver (Figure 2A).

Genes in female sexual development

Eight female determining/differentiation genes were scrutinized in *Latimeria* (Table 2): three belonging to the WNT signalling pathway (*WNT4*, *RSPO-1*, and *CTNNB1*), a transcription factor (*FOXL2*), two estrogen receptors (*ER α* and *ER β*), a steroidogenic enzyme (*Aromatase*), and an activin binding protein (*FST*).

ENSEMBL prediction recovered all eight gene sequences in the *L. chalumnae* genome. Four transcripts (*ER β* , *CTNNB1*, *WNT4*, and *FOXL2*) have a complete CDS; for *FST* only two codons are missing at the 5' end; *RSPO-1* and *Aromatase* are partially complete while *ER α* could be only partially identified, being broken in 4 different scaffolds in the WGS.

The transcriptome analysis allowed to obtain 3 complete CDS sequences of *L. menadoensis* (*CTNNB1*, *ER β* , and *FST*), 2 mRNAs revealed fragmented (*RSPO-1* and *ER α*) and 3 transcripts were missing (*FOXL2*, *WNT4*, and *Aromatase*).

The comparison of *L. menadoensis* and *L. chalumnae* female sex development analysed sequences is depicted in Figure S1B.

Expression values in testis and liver of *L. menadoensis* are shown in Figure 9. *WNT4*, *FOXL2* and *Aromatase* (considered to be responsible for female development and pathway maintenance), as expected, did not show any expression in testis. Moreover *CTNNB1*, *FST*, and *ER β* show a remarkably high liver expression (56.08, 27.33, 12.93 FPKM, respectively). While hepatic expression of *FST* and *CTNNB1* was expected because these genes are described as ubiquitously expressed [76], the expression of *ER β* in liver of the male specimen was unexpected.

Discussion

In this study for the first time the basic set of genes that are generally believed to be critically involved in sexual development was isolated and characterized from coelacanths. Comparison of the genes between *Latimeria* species confirmed the very slow rate of evolution noted recently for the example of HOX genes [77], which are, however, known to evolve slowly by themselves. The genes studied here belong to very different gene families and thus give a more representative view on the aspect.

We are aware that many further interpretations, which can be made on the basis of the data reported here, are limited by the fact that only a single individual was studied. However, given the importance of this living fossil for understanding the evolution of tetrapods and fish and considering the exceptional opportunity to obtain RNA of suitable quality from an organism that is rare and listed as one of the most endangered species, some conclusions with the appropriate note of caution are nevertheless made here.

Based on different methods to calculate Ka/Ks values it can be concluded that none of the set of genes studied here is under positive selection in coelacanths.

A totally unexpected finding was the very high expression of *DMRT6* in testis, which in fact was the most abundant of all male-specific transcripts analyzed. *DMRT6* thus far was only known from amniotes and is absent in *Xenopus* and all fish genomes. This phylogenetic pattern could be explained by *DMRT6* being a newly arisen paralog of the *DMRT* gene family at the base of the amniote vertebrates. The presence of a bona fide *DMRT6* homolog in *Latimeria* speaks for a much earlier origin of this gene and supports a possible origin from the 1R/2R whole genome duplication events in the ancestral vertebrates [78]. Consequently, it was lost repeatedly in the teleost fish and the amphibian lineages, and even in the basal chordates. We can only speculate about some explanations, as only scarce data are available on expression of *DMRT6*. In mouse embryo the gene is expressed in the developing brain, but not in gonad [43]. In the human microarray database (<https://www.genevestigator.com>) an exclusive and high expression is recorded for ovary and testis while studies using mouse organs revealed a different picture. In fact only erythroblast and oocytes showed some elevated expression. Whatever the ancestral function of *DMRT6* was, it is reasonable to assume that this was taken over by other members of the gene family or is not required anymore in those lineages where it was lost. The persistence of the gene in *Latimeria* could be explained by an important function in male (and eventually also in female?) development, which, according to the current state of knowledge, was then at least partially conserved in the amniotes. Our findings in the

coelacanth certainly motivate to put now emphasis on *DMRT6* as a putative novel gene in the gene regulatory network governing gonad development.

The high expression of *DMRT1* in *Latimeria* testis and no expression in liver is in line with the known expression pattern and its important function for testes development and male gonad identity maintenance in vertebrates, from fish to mammals [65, 66]. In teleost fishes *DMRT1* in the adult testes is found in germ cells or somatic cell types or in both (for review see ref 65). Unfortunately, the RNA-Seq transcriptome data give no information on the cell type in coelacanth testis that express *DMRT1*. In medaka, a duplicated version of *DMRT1* on the Y-chromosome, designated *dmrt1bY*, is the master male sex determining gene (ref 26, ref 27)). Its major function appears to be the suppression of germ cell proliferation at the critical sex determining stage in males (Herpin et al. , BMC Dev. Biol 7: 99, 2007). In adult testes the expression is strongly downregulated (Hornung et al. Sex Dev 1:197, 2007) and only the autosomal *DMRT1* (designated *dmrt1a* in medaka) seems to operate in the mature testis because of its high expression. In *L. menadoensis* like for all other teleosts studied so far, only one copy of *DMRT1* was found. Thus, most likely *DMRT1* in coelacanths does not have a major role in primary sex determination but more in testis differentiation and adult testis function.

However, contrary to all other vertebrates studied [38, 43, 79-81] there was no expression of *DMRT3* in the male gonad of *L. menadoensis*.

The *TGF- β* family member *GSDF* is an important factor in gonad development of teleost fish with much higher expression in testis than ovary [39, 40]. In one species a duplicate of *GSDF* most likely became even the master male sex determination gene [29]. In medaka there is strong evidence that the master male sex determining gene *dmrt1bY* upregulates *GSDF* and that this is correlated with early testicular differentiation (Shibata et al. Gene Expr. Patterns 10:283, 2010). No homologue of *GSDF* has so far been identified outside teleosts. Our identification of a bona-fide *GSDF* sequence in coelacanths and its high expression in testis pointing to functional conservation as well, leads to the conclusion that this gene already arose at the base of the fish lineage, but was later lost during evolution of the tetrapods. *GSDF* thus appears to be an ancient male sex determining gene. Whether another *TGF- β* family member has taken over the function from *GSDF* in teleosts and coelacanths in testis development of tetrapods remains unclear in the absence of functional data on *GSDF* function in fish.

The high expression (at least compared to liver) of *SOX9*, *SOX10*, *WT1*, *AMH*, *DHH*, *SF-1* and *SDR5A1* and 3 as well as the low but testis-specific expression of *AMHR2* and the absence of the female factors *FST*, *RSPO-1*, *WNT4*, *FOXL2*, *Aromatase*, and estrogen receptor transcripts in the

testis transcriptome of *L. menadoensis*, are in line with the expression pattern of these genes known from many vertebrate species and their proposed function in sexual development.

In particular, the AMH/AMH-receptor system is of interest for sexual development in *Latimeria*. In mammals and most likely in all tetrapods AMH induces the regression of the Müllerian duct. Teleosts do not have Müllerian ducts, but Lungfish and *Latimeria* possess oviducts that are homologous to those of the tetrapods (Kapoor, BG, Khanna, B: *Ichthyology Handbook*, p 487, Springer 2004). Despite the absence of Müllerian ducts an important function for AMH/AMH-receptor in the manifestation of gonadal sex in teleosts has been shown, because in medaka AMH signalling is crucial for regulation of germ cell proliferation during early gonad differentiation (Nakamura et al. *Development* 139:2283, 2012). In adult teleosts the AMH signalling system is present and probably active in testis and ovary (Klüver et al, *Dev Dyn* 236:271, 2007, Pala et al, *Gene* 410: 249, 2008, Halm et al. *Gene* 388:148, 2006), while in mouse testis this system is downregulated before sexual maturity (Beau et al. *Mol. Reprod. Dev.* 56: 124, 2000). Given the robust expression of AMH and AMH-receptor in adult testis of *L. menadoensis*, which is more alike the situation in teleosts it will be interesting to know the expression patterns of both genes during the period of sex determination and early gonad formation in coelacanth and to compare this to the situation found in tetrapods or in teleosts.

In the liver transcriptome, several of the tested genes with reported function in sex determination and differentiation were found to be quite abundantly expressed. The high levels of *CTNNB1* are expected due to the ubiquitous function of this signal transducer of the WNT pathway. The high *FST* expression is in line with a generally broad expression profile observed in all vertebrates and with a finding in mice where this gene is required for homeostasis of liver cell growth [82]. This non-gonadal function of *FST* may be conserved in coelacanth. Similarly, the transcription factor *GATA-4* is besides its importance for regulating genes in testis development [83] also involved in the control of a number of liver genes, explaining why transcripts of the coelacanth homolog were found in both tissues. Contrary to coelacanth, where *5 α -reductase 2* is highly expressed in liver, in rat the type 1 isoform is differentially regulated by androgens and glucocorticoids in the liver resulting in high expression in this tissue, while type 2 is preferentially expressed in gonads [84]. This may indicate lineage specific sub-functionalization of these isozymes during evolution.

The absence of *SOX8* expression in *Latimeria* testis was unexpected. In other vertebrates, including teleost fish, *SOX8* expression is readily detected in this organ and has been assigned in mammals an important function in the FGF9/SOX9 interaction loop to maintain Sertoli cell identity by acting redundantly to *SOX9* [6, 85]. It appears that such a back-up function is not required in *Latimeria*

testis maintenance or that the redundant function has been lost in the extant coelacanth lineage. In medaka it was shown that SOX9 is required for germ cell proliferation and survival, but not for testis determination (Nakamura et al.2012, PLoS ONE). Together with our findings in *L. menadoensis* this may indicate that the sex determining role is a function that was acquired later in vertebrate evolution in the tetrapod lineage after the split of the teleost and coelacanth lineages.

An intriguing situation was found for *FGF9* and *20*, which constitute together with *FGF16* a subfamily of the paracrine *FGFs*. The critical role for *FGF9* in testis development is firmly established in mammals and appears to be well conserved in all tetrapods. On the other hand, this gene is absent from all teleost genomes ([63, 64], ENSEMBL), while *FGF16* and *20*, the latter being duplicated due to the teleost genome duplication, are present. In amphioxus there is one *FGF* gene that is basal to all three *FGFs* in tetrapods [86]. Thus *FGF9* could be a later duplicate of either *FGF16* or *20* and its role in testis development could be interpreted as a tetrapod innovation. The here reported presence of *FGF9* in *Latimeria*, however, supports an origin during the 1R/2R whole genome duplications in the ancestral chordates and a specific loss in the lineage leading to the teleosts. In the teleost *Oreochromis niloticus* (Tilapia) *FGF20b* and *FGF16* are both expressed in ovary and only a lower expression of *FGF16* was recorded in testis [64]. Together with the total absence of expression of *FGF9*, *FGF20* and *FGF16* in *L. menadoensis*, this indicates that the testis function of FGF signaling, in particular the central role of *FGF9*, was acquired later during evolution of the tetrapods.

Surprisingly, the *ERβ* gene was found to be expressed in the liver of the male coelacanth. Previously, it was noted that in the same individual the vitellogenin genes *vtgABI*, *II* and *III* were expressed [50]. Vitellogenins are yolk proteins that are physiologically expressed in the liver of reproductive females upon induction by estrogens. Thus expression of vitellogenins and estrogen receptor indicates the presence of estrogens in this male individual. Such estrogens could be derived from the environment pollutants as reported from a number of instances for fish from polluted waters. The fish analyzed here lived in Bunaken Marine Park in submarine caves at depths of 100 to 200 m, which can be considered a relatively protected environment. Alternatively, the *ERβ* expression could be the consequence of a pathological condition in the male, hormonal imbalance due to aging or a specific physiological feature of coelacanths.

Conclusions

In summary, this first analysis of a coelacanth testis transcriptome already revealed some information with respect to the question, how sexual development and testis differentiation may be regulated in this living fossil, and also gave new information about the evolution of this process in vertebrates. Interestingly, some genes that are generally considered as indispensable for testis development in all vertebrates, like *SOX8* or a fibroblast growth factor gene from the *FGF9/16/20* subfamily, are obviously not playing such a role in *Latimeria*. Together with the high *GSDF* expression the transcript profile is more alike that of the modern fish. The coelacanth testis transcriptome will help to reconstruct the ancestral tetrapod situation and gives hints, which evolutionary innovations for sexual development had occurred during the process of transition from water to land.

Acknowledgments

FM, CA, BMA, BM, and OE are affiliated to Istituto Nazionale Biosistemi e Biostrutture (INBB).

Figures

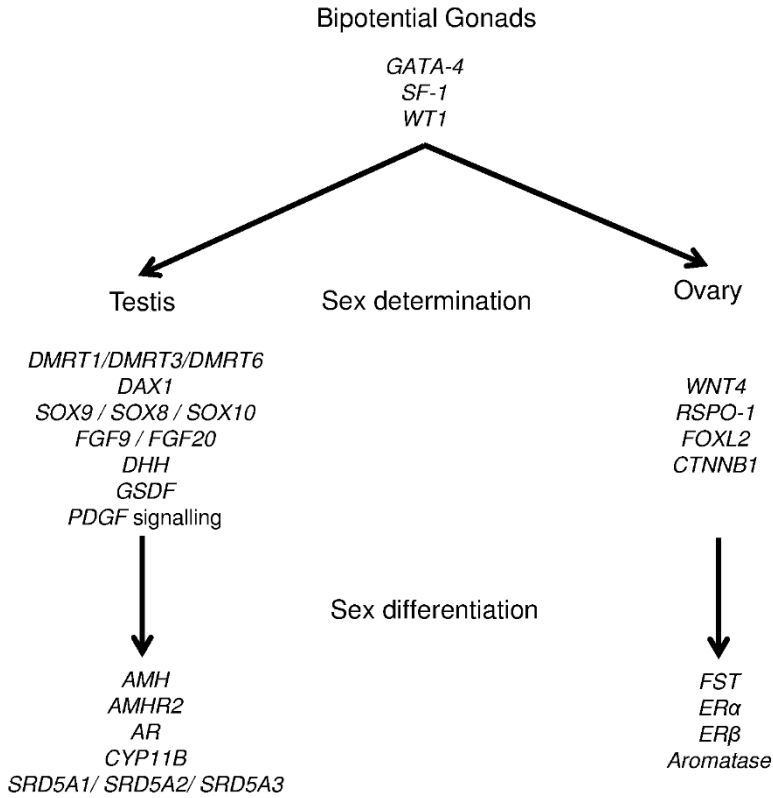


Figure 1. Genes involved in sexual development

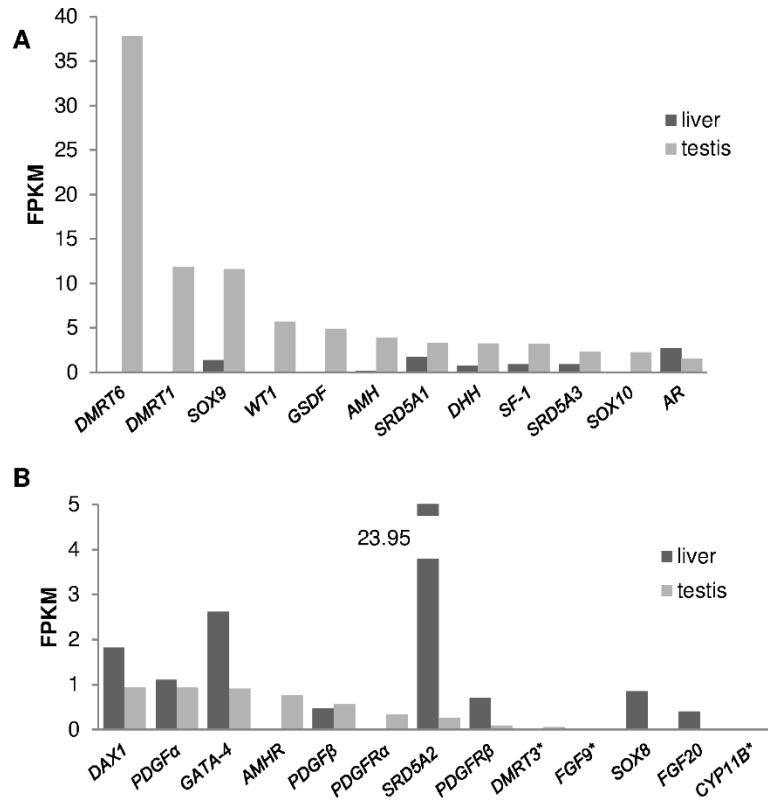


Figure 2. Male development gene expressions

Expression levels of male sex determining/differentiation genes in liver and testis transcriptomes of *L. menadoensis*. Values are expressed in FPKM (Fragments Per Kilobase of exon per Million sequenced fragments). A) genes highly expressed in testis B) genes poorly expressed in testis. The expression levels of some housekeeping genes (not represented) were also analysed: *phosphoglycerate kinase* 96.95 (liver), 342.41 (testis); *ribosomal protein S27a* 152.59 (liver), 128.43 (testis); *RPL19* 744.01 (liver) 64.89 (testis); *RPL11* 457.35 (liver), 282.59 (testis); *RPL32* 629.83 (liver), 373.75 (testis); *HSPCB* 507.99 (liver), 1213.75 (testis).

Threshold value= 1. * Expression level assessed on *L. chalumnae* ortholog.

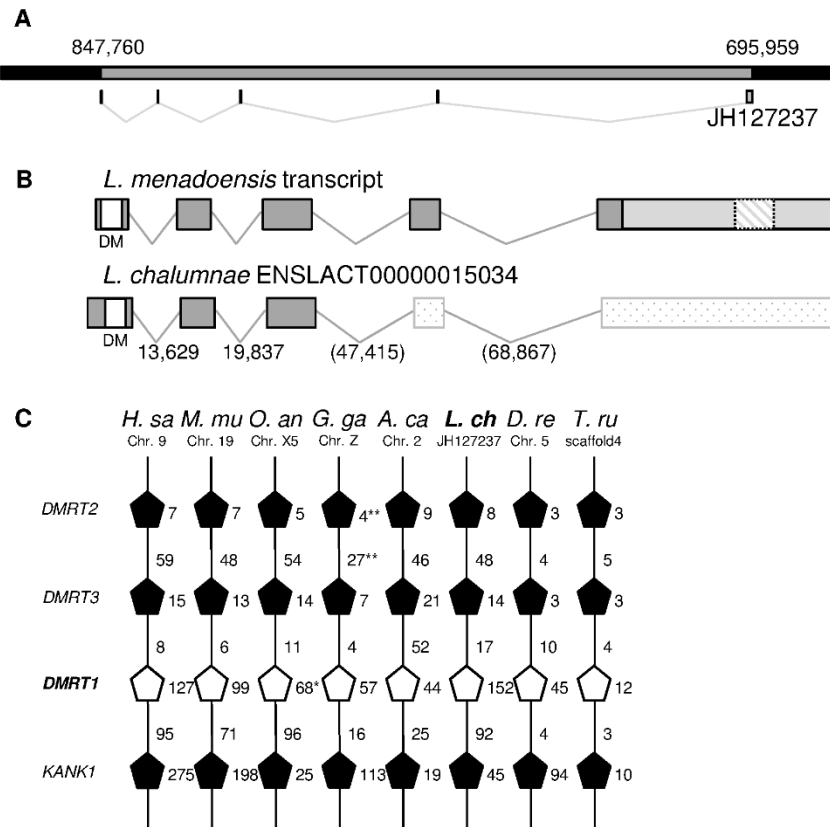


Figure 3. Conserved micro-synteny and structure of the *DMRT1* genomic locus and transcripts

A) Genomic representation of *DMRT1* on scaffold JH127237 of *L. chalumnae*. Grey box corresponds to gene. Small boxes and V signs represent the intron/exon map.

B) Transcript representation of *DMRT1* in *L. menadoensis* and in *L. chalumnae*. Boxes are the exons and V-signs are the introns. White box is the DM domain, light grey box is the 3'UTR. Dashed box is a putative transposable element contained in the 3'UTR, dotted boxes represent the missing exons in Ensembl transcript prediction.

C) Micro-syntenic conservation of genomic blocks containing the *DMRT1* gene. White pentagons represent *DMRT1* genes. The tip on the line indicates the relative orientation of the genes. Numbers near the pentagons are the gene size expressed in kb, numbers on lines represent inter-gene distances expressed in Kb.

Data from Ensembl: *H. sa* (*Homo sapiens*), *M. mu* (*Mus musculus*), *O. an* (*Ornithorhynchus anatinus*), *G. ga* (*Gallus gallus*), *A. ca* (*Anolis carolinensis*), *L. ch* (*Latimeria chalumnae*), *D. re* (*Danio rerio*), *T. ru* (*Takifugu rubripes*). *L. chalumnae* *DMRT1* position was defined through *L. menadoensis* transcript, elongating ENSLACT00000015034 coordinates.

*In *O. anatinus* *DMRT1* gene size was defined from comparison with other species.

**Values obtained in *G. gallus* from the annotation of NC_006127.3 accession.

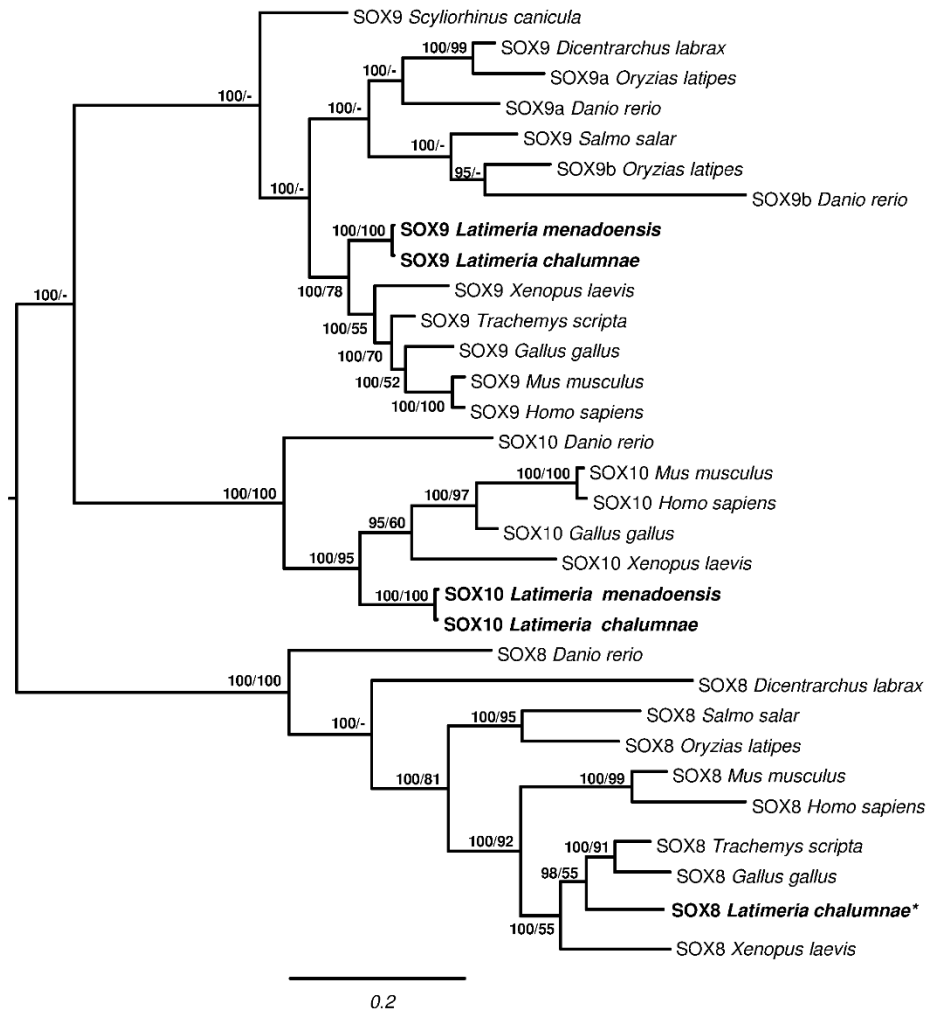


Figure 4. Phylogenetic tree of SOX8, SOX9, and SOX10

Phylogenetic analyses of vertebrate SoxE amino acidic sequences. Midpoint rooting. Total characters: 599, constant: 179, parsimony non-informative: 78, parsimony informative: 342. Numbers close to the nodes represents posterior probability in Bayesian Inference/bootstrap percentage in Maximum Parsimony.

Danio rerio (SOX8: AAX73357.1; SOX9a: NP_571718.1; SOX9b: NP_571719.1; SOX10: AAK84872.1); *Dicentrarchus labrax* (SOX8: CBN81184.1; SOX9: CBN81190.1); *Gallus gallus* (SOX8: AAF73917.1; SOX9: BAA25296.1; SOX10: AAD38050.2); *Homo sapiens* (SOX8: AAH31797.1; SOX9: CAA86598.1; SOX10: CAG30470.1); *Latimeria chalumnae* (SOX8: ENSLACP00000018883; SOX9: ENSLACP00000021343; SOX10: ENSLACP00000004990); *Latimeria menadoensis* (SOX9, SOX10: this study); *Mus musculus* (SOX8: AAF35837.1; SOX9: AAH23953.1; SOX10: NP_035567.1); *Oryzias latipes* (SOX8: NP_001158342.1; SOX9: BAC02947.1); *Salmo salar* (SOX8: ABC24688.1; SOX9: ACN10975.1); *Scyliorhinus canicula* (SOX8: ABA10785.1; SOX9: ABY71239.1); *Trachemys scripta* (SOX8: AAP59791.1; SOX9: ACG70782.1; SOX10: ENSLACP00000004990); *Xenopus laevis* (SOX8: AAI69525.1; SOX9: NP_001084276; SOX10: NP_001082358.1).

*Only a partial SOX8 sequence was retrieved in the transcriptome assembly of *L. menadoensis*, perfectly matching to the prediction of ENSEMBL for SOX8 gene in *L. chalumnae*.

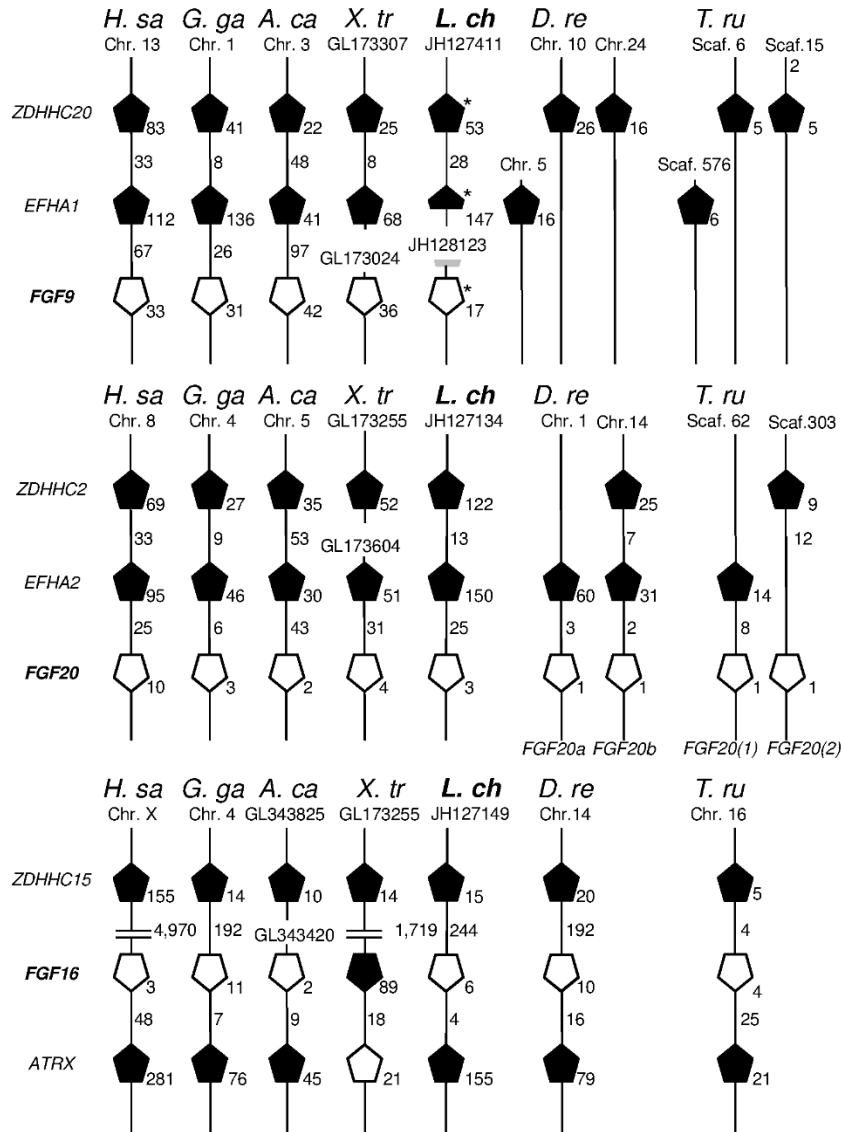


Figure 5. Analysis of micro-syntenic conservation in *FGF9*, *FGF16* and *FGF20* blocks

Micro-syntenic conservation of genomic regions containing *FGF9*, *FGF20* and *FGF16* genes. White pentagons represent *FGF* genes. The tip on the line indicates the relative orientation. Gray shape indicate putative sequences. Numbers near the pentagons are the gene size expressed in Kb, numbers on lines represent intergene distances expressed in Kb.

Data from Ensembl: *H. sa* (*Homo sapiens*), *G. ga* (*Gallus gallus*), *A. ca* (*Anolis carolinensis*), *X. tr* (*Xenopus tropicalis*), *L. ch* (*Latimeria chalumnae*), *D. re* (*Danio rerio*), *T. ru* (*Takifugu rubripes*).

Syntenic blocks for *FGF20* on *L. chalumnae* and *X. tropicalis*, and *FGF16* in *A. carolinensis* are broken on two different scaffolds. The *ZDHHC15* genes belonging to the syntenic block of *FGF16* in *H. sapiens* and *X. tropicalis* are maintained on the same chromosome or scaffold, but they are situated far from the genomic locus of *FGF16* and *ATRX*. *Genes missing in ENSEMBL prediction.

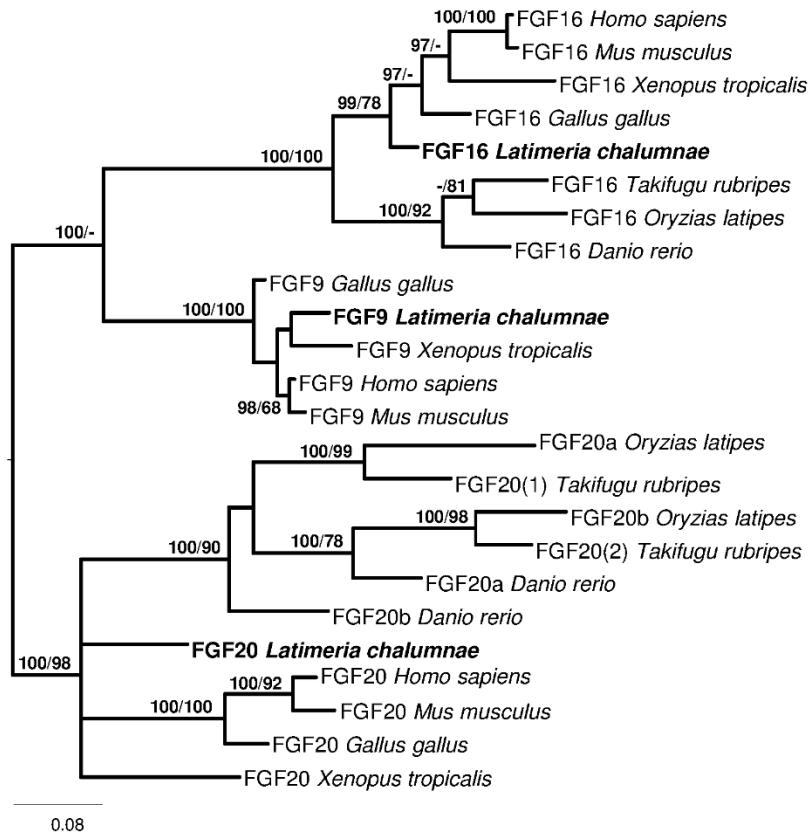


Figure 6. Phylogenetic tree OF FGF9, FGF16, and FGF20

Phylogenetic analysis of amino acid sequences of the vertebrate FGF9/16/20. Midpoint rooting. Total characters: 237, constant: 91, parsimony non-informative: 35, parsimony informative: 111. Numbers close to the nodes represents posterior probability in Bayesian Inference/bootstrapped percentage in Maximum Parsimony.

Danio rerio (FGF16: ENSDART00000061928; FGF20a: NP_001032180.1; FGF20b: NP_001034261.1); *Gallus gallus* (FGF9: NP_989730.1; FGF16: NP_001038115.1; FGF20: XP_426335.2); *Homo sapiens* (FGF9: NP_002001.1; FGF16: NP_003859.1; FGF20: NP_062825.1); *Latimeria chalumnae* (FGF9: manually inferred on JH128123; FGF16: ENSLACT00000011509; FGF20: ENSLACT00000014939); *Mus musculus* (FGF9: ADL60500.1; FGF16: BAB16405.1; FGF20: NP_085113.2); *Oryzias latipes* (FGF16: ENSORLT00000007651; FGF20a: ENSORLT00000012578; FGF20b: ENSORLT00000025767); *Takifugu rubripes* (FGF16: ENSTRUT00000021181; FGF20(1): ENSTRUT00000008788; FGF20(2): ENSTRUT00000039390); *Xenopus tropicalis* (FGF9: XP_002938621.1; FGF16: ENSXETT00000009790; FGF20: NP_001137399.1). *Latimeria menadoensis* is missing in this analysis because FGF9 and FGF20 are low or not expressed in the transcriptomes.

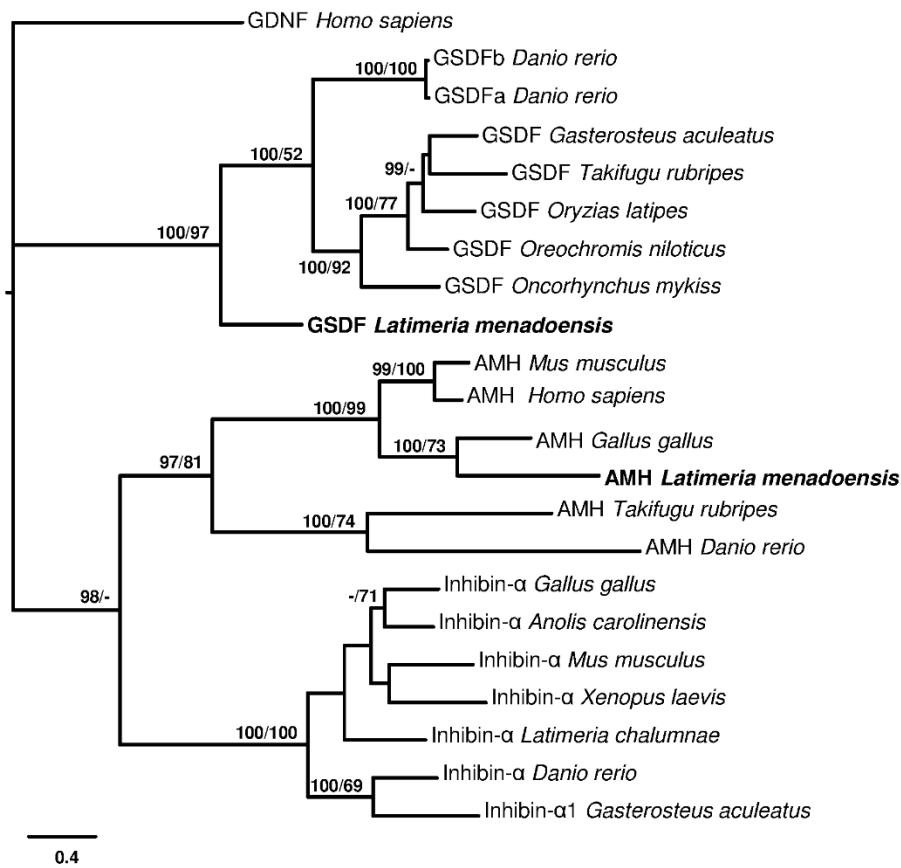


Figure 7. Phylogenetic tree of GSDF, AMH, and INHIBIN α

Phylogenetic analysis of amino acidic sequences of vertebrate GSDF, inhibin α and AMH. Total characters: 849, constant: 84, parsimony non-informative: 225, parsimony informative: 540.

Outgroup: human Glial-Derived Nerve growth Factor (GDNF). Numbers close to the nodes represent posterior probabilities in Bayesian Inference/bootstrap percentage in Maximum Parsimony. *Anolis carolinensis* (inhibin α : ENSACAT00000014331); *Danio rerio* (GSDFa: AEL99890.1; GSDFb: AEL99889.1; AMH: NP_001007780.1; inhibin α : ENSDART00000057348); *Gallus gallus* (AMH: NP_990361.1; inhibin α : NP_001026428.1); *Gasterosteus aculeatus* (GSDF: ENSGACT00000021595; inhibin α : ENSGACT00000018909); *Homo sapiens* (AMH AAC25614.1; GDNF: NP_000505.1); *Latimeria chalumnae* (inhibin α : ENSLACT00000017535); *Latimeria menadoensis* (GSDF, AMH this study); *Mus musculus* (AMH: AAI50478.1; inhibin α : AAH56627.1); *Oreochromis niloticus* (GSDF: BAJ78985.1); *Oryzias latipes* (GSDF: NP_001171213.1); *Oncorhynchus mykiss* (GSDF: ABF48201.1); *Takifugu rubripes* (GSDF: ENSTRUT00000036269; AMH: ENSTRUT00000045919); *Xenopus laevis* (inhibin α : NP_001106349.1). The reliability of the CDS in *L. menadoensis* is sustained by the two different assembly procedures applied resulting in the same sequence.

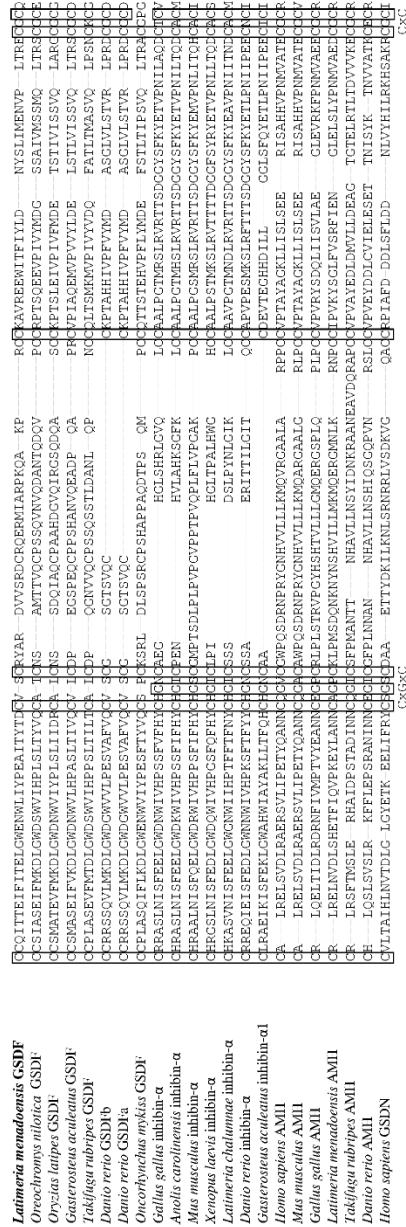


Figure 8. Multiple alignment of TGF-β domain in GSDf, AMH, and INHIBIN α

Conserved amino acids of the cysteine knot are boxed. *Anolis carolinensis* (inhibin α: ENSACAT00000014331); *Danio rerio* (GSDf: AEL99890.1, GSDf: AEL99889.1; AMH: NP_001007780.1; inhibin α: ENSDART00000057348); *Gallus gallus* (AMH: NP_990361.1; inhibin α: NP_001026428.1); *Gasterosteus aculeatus* (GSDf: ENSGACT00000021595; inhibin α: ENSGACT00000018909); *Homo sapiens* (AMH: AAC25614.1; GDNF: NP_000505.1); *Latimeria chalumnae* (inhibin α: ENSLACT00000017535); *Latimeria menadoensis* (this study); *Mus musculus* (AMH: AAI50478.1; inhibin α: AAH56627.1); *Oreochromis niloticus* (GSDf: BAJ78985.1); *Oryzias latipes* (GSDf: NP_001171213.1); *Oncorhynchus mykiss* (GSDf: ABF48201.1); *Takifugu rubripes* (GSDf: ENSTRUT00000036269; AMH: ENSTRUT00000045919); *Xenopus laevis* (inhibin α: NP_001106349.1).

The reliability of the CDS in *L. menadoensis* is supported by the two different assembly procedures applied in this study resulting in the same sequence.

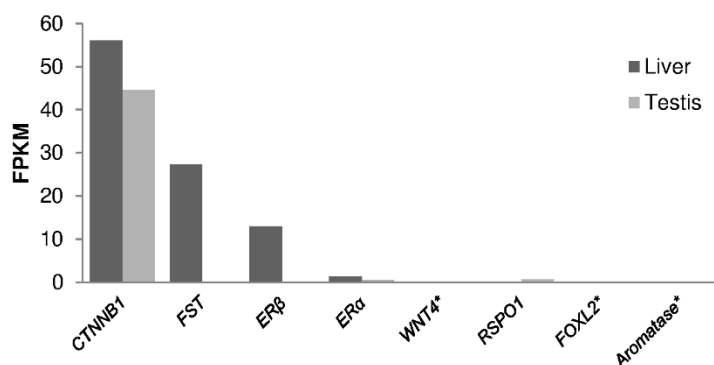


Figure 9. Female development gene expressions

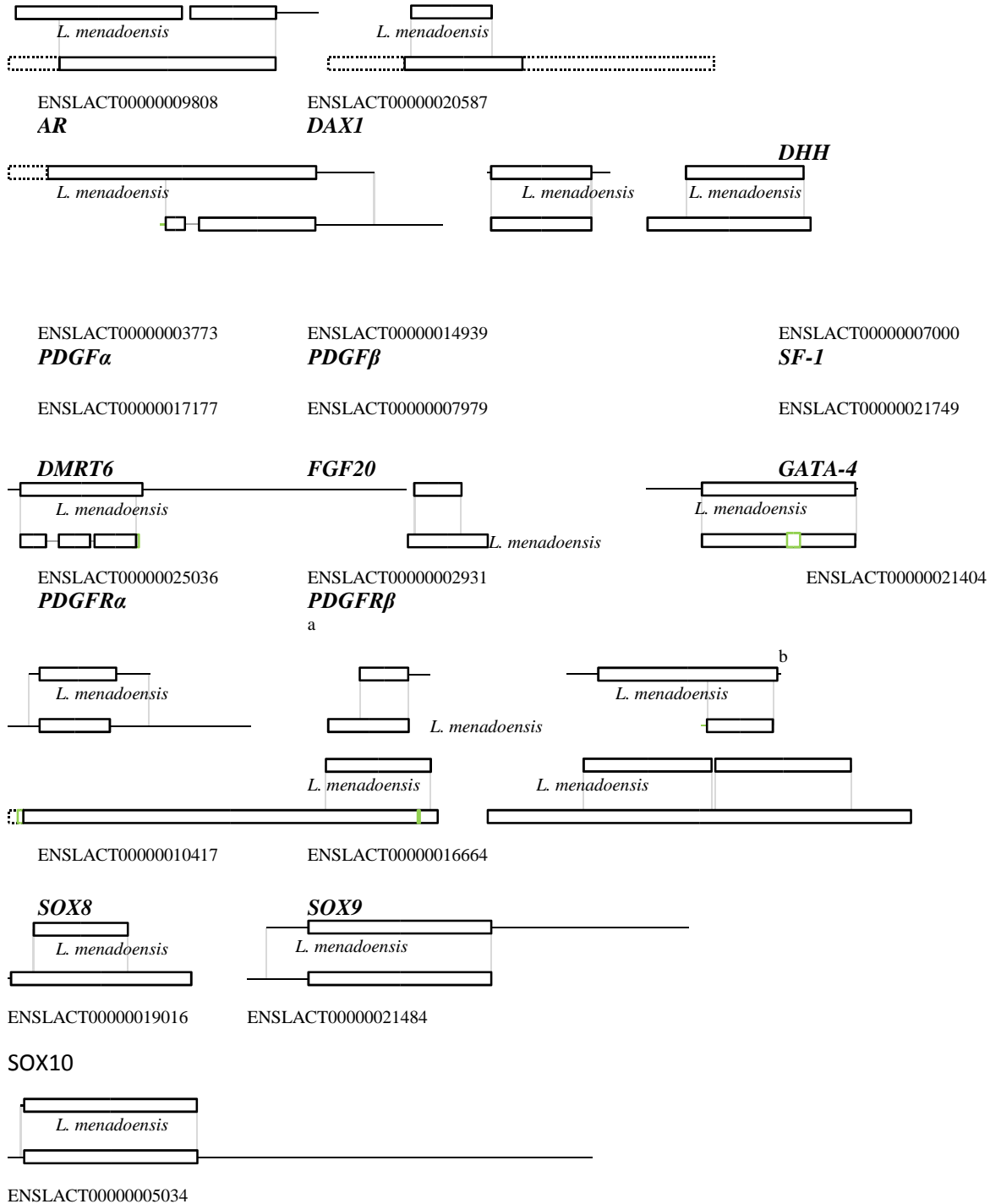
Expression of female determining/differentiation genes in liver and testis transcriptomes of *L. menadoensis*. Expression levels are reported in FPKM (Fragments Per Kilobase of exon per Million sequenced fragments). The expression levels of some housekeeping genes were also analysed: *phosphoglycerate kinase* 96.95 (liver), 342.41 (testis); *ribosomal protein S27a* 152.59 (liver), 128.43 (testis); *RPL19* 744.01 (liver) 64.89 (testis); *RPL11* 457.35 (liver), 282.59 (testis); *RPL32* 629.83 (liver), 373.75 (testis); *HSPCB* 507.99 (liver), 1213.75 (testis). Threshold value= 1.

* Expression level assessed on *L. chalumnae* ortholog.

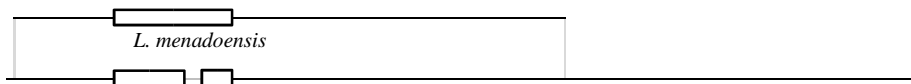
Supporting information

AMH

AMHR2^{a b}

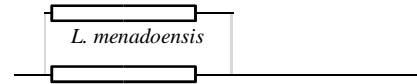


SRD5A1



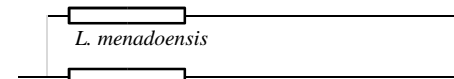
ENSLACT00000002047

SRD5A3



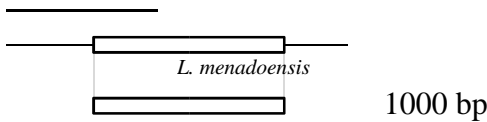
SRD5A2

ENSLACT00000014423



ENSLACT00000025936

WT1



ENSLACT00000018732

Figure S1A

Sequence pair comparison of male sex development genes.

Sequence pair comparison of male and female sex determining/differentiation transcripts in *L. menadoensis* transcriptome and *L. chalumnae* ENSEMBL predictions. Male sex development genes. Boxes represent CDS. Lines represent UTR. Dashed boxes represent a missing part in the CDS. Green lines/boxes represent an imprecise gene prediction or mismatch among *L. chalumnae* and *L. menadoensis* sequences. Scale dimension are maintained.

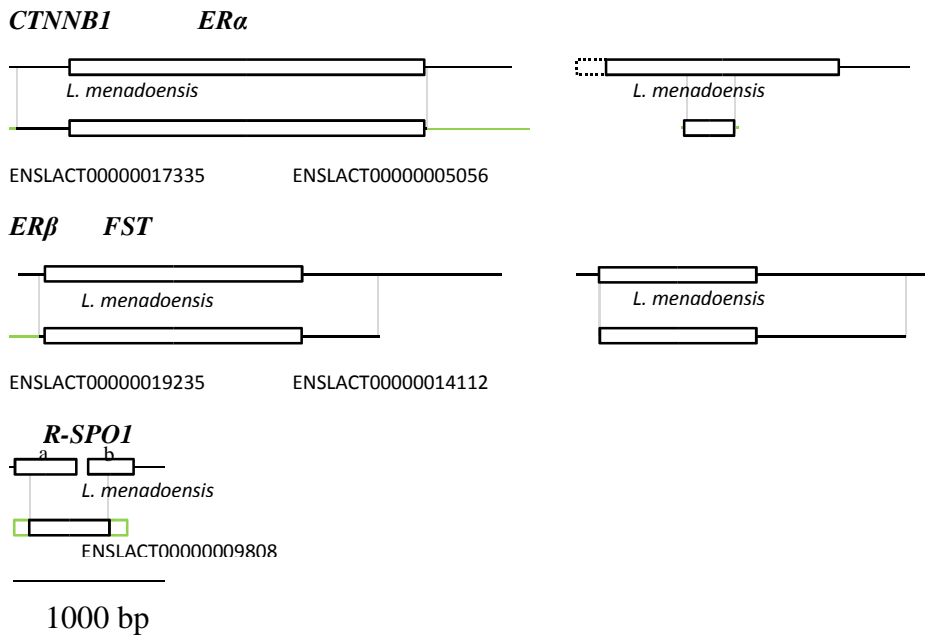


Figure S1B

Sequence pair comparison of female sex development genes.

Sequence pair comparison of male and female sex determining/differentiation transcripts in *L. menadoensis* transcriptome and *L. chalumnae* ENSEMBL predictions. Female sex development genes. Boxes represent CDS. Lines represent UTR. Dashed boxes represent a missing part in the CDS. Green lines/boxes represent an imprecise gene prediction or mismatch among *L. chalumnae* and *L. menadoensis* sequences. Scale dimension are maintained.

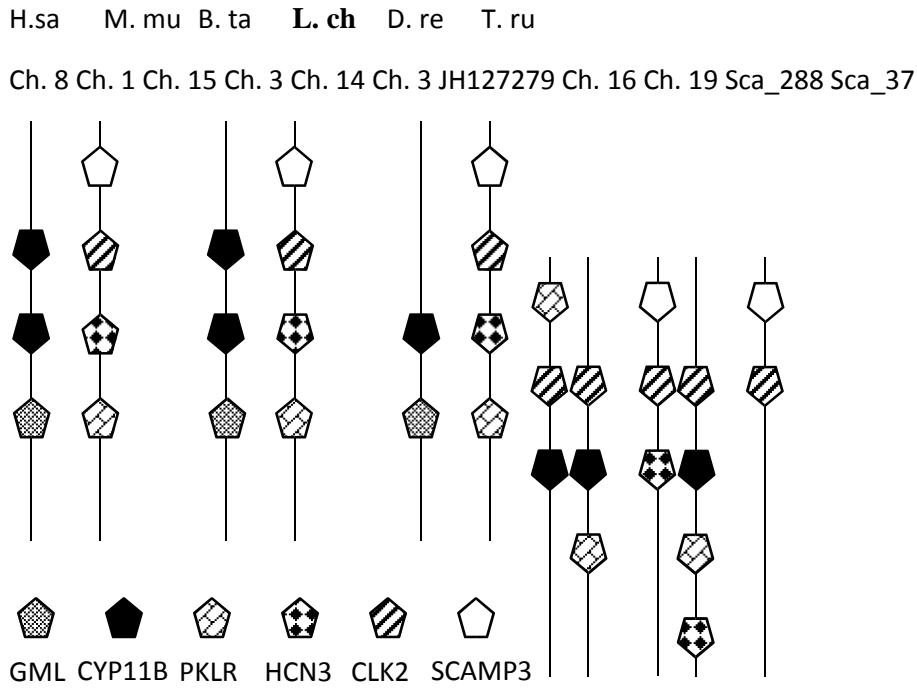


Figure S2

Micro-syntentic conservation of *CYP11B*.

Micro-syntentic conservation of genomic regions containing *CYP11B* genes. Black pentagons represent *CYP11B* genes. The tip on the line indicates the relative orientation.

Data from Ensembl: *H. sa* (*Homo sapiens*), *M. mu* (*Mus musculus*), *B. ta* (*Bos Taurus*), *L. ch* (*Latimeria chalumnae*), *D. re* (*Danio rerio*), *T. ru* (*Takifugu rubripes*).

Tables

Gene	Transcript in <i>L. menadoensis</i>			Gene Location in <i>L. chalumnae</i>				Transcript prediction in <i>L. chalumnae</i>		
	Accession	length	CDS	Scaffold	N° exons ³	Divergence ⁴	Ka/Ks	ENSEMBL accession	length	CDS
<i>AMH</i>		1312+1039 ¹	1312+670 ²	JH126742	>5	0.046	0.000	ENSLACT00000009808	1689	1689 ²
<i>AMHR2</i>		693	693 ²	<u>JH126659</u>	>9	0.289	0.343	ENSLACT00000020587	921	921 ²
<i>AR</i>		2590	2133 ²	JH126641	>8	0.165	0.000	ENSLACT00000017177	2235	1239 ²
<i>CYP11B</i>	-	-	-	JH127279	-	-	-	ENSLACT00000015536	1422	474 ²
<i>DAX1</i>		966	786	JH128268	2	0.207	0.309	ENSLACT00000007979	786	786
<i>DHH</i>		926	926 ²	JH126563	2	0.540	0.649	ENSLACT00000021749	1275	1275
<i>DMRT1</i>		2244	998 ²	JH127237	5	0.134	0.000	ENSLACT00000015034	798	798 ²
<i>DMRT3</i>	-	-	-	JH127237	-	-	-	ENSLACT00000013757	1455	1455
<i>DMRT6</i>		3121	957	JH130928	4	0.129	NA	ENSLACT00000003773	798	798 ²
<i>FGF9</i>	-	-	-	JH128123	-	-	-	Manually identified		
<i>FGF20</i>		370	370 ²	JH127134	3	0.270	NA	ENSLACT00000014939	627	627
<i>GATA-4</i>		1655	1200	JH128461	>6	0.064	NA	ENSLACT00000007000	1209	1209
<i>GSDF</i>		1258	693	JH127632 5	>3	0.826	0.470	-	-	-
<i>PDGFα</i>		968	594	JH126909	6	0.000	NA	ENSLACT00000025036	1892	594
<i>PDGFβ</i>		664	483 ²	JH128946	>4	0.000	NA	ENSLACT00000002931	630	630
<i>PDGFRα</i>		823	823 ²	JH128279	>7	0.243	NA	ENSLACT00000010417	3285	3285 ²
<i>PDGFRβ</i>		995+1055 ¹	995+1055 ²	JH126585	>17	0.195	0.954	ENSLACT00000016664	3312	3312
<i>SF-1</i>		1686	1401	JH126572 6	7	0.000	NA	ENSLACT00000021404	591	591 ²

<i>SOX8</i>		735	735 ²	JH126713	>3	0.000	NA	ENSLACT00000019016	1434	1410
<i>SOX9</i>		3306	1428	JH126581	3	0.185	0.000	ENSLACT00000021484	1908	1428
<i>SOX10</i>		1403	1353 ²	JH127309	3	0.359	0.199	ENSLACT00000005034	4571	1356
<i>SRD5A1</i>		3664	786	JH129903	5	0.164	NA	ENSLACT00000002047	6104	684
<i>SRD5A2</i>		2711	765	JH126700	5	0.112	NA	ENSLACT00000025936	2918	765
<i>SRD5A3</i>		1244	945	JH127256	5	0.000	NA	ENSLACT00000014423	2733	945
<i>WT1</i>		2260	1257	JH126652	9	0.134	0.000	ENSLACT00000018732	1260	1257

Table 1. Male sex determining/differentiation gene inventory.

¹Fragmented contig. ²Partial CDS. ³Number of exons from the alignment of *L. menadoensis* transcripts to the *L. chalumnae* genome. Where the transcript is carrying only a partial CDS, the number of exon is partial. ⁴Divergence between the two coelacanth sequences calculated as p distance x100. ⁵*SF-1* gene in *L. chalumnae* is split in scaffold JH126572 and contig AFYH01271535. ⁶*GSDF* gene on *L. chalumnae* is split in scaffold JH127632 and contig AFYH01270444.

Gene	Transcript in <i>L. menadoensis</i>			Gene Location in <i>L. chalumnae</i>				Transcript prediction in <i>L. chalumnae</i>		
	Accession	length	CDS	Scaffold	N° exons ³	Divergence ⁴	Ka/Ks	ENSEMBL accession	length	CDS
<i>Aromatase</i>	-	-	-	JH127307	-	-	-	ENSLACT00000010703	1329	1329 ₂
<i>CTNNB1</i>		3325	2346	JH127054	15	0.702	0.000	ENSLACT00000017335	3458	2346
<i>ERα</i>		2002	1541 ²	JH129227 ⁵	>8	0.352	NA	ENSLACT00000005056	396	396 ²
<i>ERβ</i>		3184	1689	JH126564	9	0.159	0.000	ENSLACT00000019235	2465	1689
<i>FOXL2</i>	-	-	-	JH127245	-	-	-	ENSLACT00000012991	915	915
<i>FST</i>		2381	1044	JH127291	6	0.221	0.000	ENSLACT00000014112	2027	1032 ₂
<i>RSPO-1</i>		474+485 ¹	425+269 ²	JH126592	>5	0.626	NA	ENSLACT00000019383	747	747 ²
<i>WNT4</i>	-	-	-	JH126950	-	-	-	ENSLACT00000017139	1068	1068

Table 2. Female sex determining/differentiation gene inventory.

¹Fragmented contig. ²Partial CDS. ³Number of exons from the alignment of *L. menadoensis* transcripts to the *L. chalumnae* genome. Where the transcript is carrying only a partial CDS, the number of exon is partial. ⁴Divergence between the two coelacanth sequences calculated as p distance x100. ⁵*ER α* gene in *L. chalumnae* genome is split in scaffold JH129227, JH129408, JH129637, and JH133026.

GO0007530 (Sex determination)	Total annotations	Matching annotations	<i>L. menadoensis</i> orthologs
<i>Danio rerio</i>	2	2	2
<i>Xenopus laevis</i>	3	3	2
<i>Gallus gallus</i>	27	27	13
<i>Canis familiaris</i>	19	4	4
<i>Sus scrofa</i>	17	19	15
<i>Bos taurus</i>	12	12	11
<i>Mus musculus</i>	21	3	3
<i>Rattus norvegicus</i>	26	25	18
<i>Homo sapiens</i>	33	31	20

Table S1

Gene Ontology analysis for “sex determination” term.

GO0007548 (Sex differentiation)	Total annotations	Matching annotations	<i>L. menadoensis</i> orthologs
<i>Danio rerio</i>	43	43	24
<i>Xenopus laevis</i>	4	4	3
<i>Gallus gallus</i>	187	187	124
<i>Canis familiaris</i>	172	33	28
<i>Sus scrofa</i>	144	143	117
<i>Bos taurus</i>	151	147	118
<i>Mus musculus</i>	203	32	30
<i>Rattus norvegicus</i>	299	291	240
<i>Homo sapiens</i>	373	363	221

Table S2

Gene Ontology analysis for “sex differentiation” term.

References

- Hayes TB. (1998) Sex determination and primary sex differentiation in amphibian: genetic and developmental mechanisms. *J Exp Zool* 281: 373-399.
- Graves JAM, Peichel CL (2010) Are homologies in vertebrate sex determination due to shared ancestry or to limited options? *Genome Biol* 11: 205.
- McClelland K, Bowles J, Koopman P (2012) Male sex determination: insights into molecular mechanisms. *Asian J Androl* 14: 164-171.
- Uhlenhaut NH, Jakob S, Anlag K, Eisenberger T, Sekido R, et al. (2009) Somatic sex reprogramming of adult ovaries to testes by FOXL2 ablation. *Cell* 139: 1130-1142.
- Matson CK, Murphy MW, Sarver AL, Griswold MD, Bardwell VJ, et al. (2011) DMRT1 prevents female reprogramming in the postnatal mammalian testis. *Nature* 476: 101-104.
- Herpin A, Schartl M (2011) Sex determination: switch and suppress. *Curr Biol* 21: R656-9.
- Clinton M, Zhao D, Nandi S, McBride D (2012) Evidence for avian cell autonomous sex identity (CASI) and implications for the sex-determination process? *Chromosome Res* 20: 177-190.
- Zarkower D (2006) Somatic sex determination (February 10, 2006), in *WormBook*. Edited by The C. elegans Research Community. doi/10.1895/wormbook.1.84.1.
- Abinawanto, Shimada K, Yoshida K, Saito N (1996) Effects of aromatase inhibitor on sex differentiation and levels of P450 (17 alpha) and P450 arom messenger ribonucleic acid of gonads in chicken embryos. *Gen Comp Endocrinol* 102: 241-246.
- Akazome Y, Mori T (1999) Evidence of sex reversal in the gonads of chicken embryos after oestrogen treatment as detected by expression of lutropin receptor. *J Reprod Fertil* 115: 9-14.
- Baker PJ, Moore HD, Burgess AM, Mittwoch U (1993) Gonadal sex differentiation in embryos and neonates of the marsupial, *Monodelphis domestica*: arrest of testis development in postterm embryos. *J Anat* 182: 267-273.
- Coveney D, Shaw G, Renfree MB (2001) Estrogen-induced gonadal sex reversal in the tammar wallaby. *Biol Reprod* 65: 613-621.
- Fadem BH (2000) Perinatal exposure to estradiol masculinizes aspects of sexually dimorphic behavior and morphology in gray short-tailed opossums (*Monodelphis domestica*). *Horm Behav* 37: 79-85.
- Guiguen Y, Baroiller JF, Ricordel MJ, Iseki K, Mcmeel OM, et al. (1999) Involvement of estrogens in the process of sex differentiation in two fish species: the rainbow trout (*Oncorhynchus mykiss*) and a tilapia (*Oreochromis niloticus*). *Mol Reprod Dev* 54: 154-162.
- Kobayashi T, Kajiura-Kobayashi H, Nagahama Y (2003) Induction of XY sex reversal by estrogen involves altered gene expression in a teleost, tilapia. *Cytogenet Genome Res* 101: 289-294.

Mackenzie CA, Berrill M, Metcalfe C, Pauli BD (2003) Gonadal differentiation in frogs exposed to estrogenic and antiestrogenic compounds. *Environ Toxicol Chem* 22: 2466-7245.

Mittwoch U (1998) Phenotypic manifestations during the development of the dominant and default gonads in mammals and birds. *J Exp Zool* 281: 466-471.

Pieau C, Dorizzi M (2004) Oestrogens and temperature-dependent sex determination in reptiles: all is in the gonads. *J Endocrinol* 181: 367-377.

Renfree MB, Coveney D, Shaw G (2001) The influence of estrogen on the developing male marsupial. *Reprod Fertil Dev* 13: 231-240.

Shaw G, Renfree MB, Short RV, O WS (1988) Experimental manipulation of sexual differentiation in wallaby pouch young treated with exogenous steroids. *Development* 104: 689-701.

Ramsey M, Crews D (2009): Steroid signaling and temperature-dependent sex determination. Reviewing the evidence for early action of estrogen during ovarian determination in turtles. *Semin Cell Dev Biol* 20: 283-292.

Schartl M (2004) Sex chromosome evolution in non-mammalian vertebrates. *Curr Opin Genet Dev* 14: 634-641.

Uller T, Helanterä H (2011) From the origin of sex-determining factors to the evolution of sex-determining systems. *Q Rev Biol* 86: 163-180.

Biason-Lauber A (2010) Control of sex development. *Best Pract Res Clin Endocrinol Metab* 24: 163-186.

Kashimada K, Koopman P (2010) Sry: the master switch in mammalian sex determination. *Development* 137: 3921-3930.

Matsuda M, Nagahama Y, Shinomiya A, Sato T, Matsuda C, et al. (2002) DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* 2002, 417: 559-563.

Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, et al. (2002) A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*. *Proc Natl Acad Sci USA* 99: 11778-11783.

Yoshimoto S, Okada E, Umemoto H, Tamura K, Uno Y, et al. (2008) A W-linked DM-domain gene, DM W, participates in primary ovary development in *Xenopus laevis*. *Proc Natl Acad Sci USA* 105: 2469-2474.

Myosho T, Otake H, Masuyama H, Matsuda M, Kuroki Y, et al. (2012) Tracing the emergence of a novel sex-determining gene in medaka, *Oryzias luzonensis*. *Genetics* 191: 163-170.

Hattori RS, Murai Y, Oura M, Masuda S, Majhi SK, et al. (2012) A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. *Proc Natl Acad Sci USA* 109: 2955-2959.

Kamiya T, Kai W, Tasumi S, Oka A, Matsunaga T, et al. (2012) A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, *Takifugu rubripes* (Fugu). *PLoS Genet* 8:e1002798.

- Angelopoulou R, Lavranos G, Manolakou P (2012) Sex determination strategies in 2012: towards a common regulatory model? *Reprod Biol Endocrinol* 10:13.
- Graves JAM (1995) The evolution of mammalian sex chromosomes and the origin of sex-determining genes. *Philos Trans R Soc Lond B Biol Sci* 350: 305-311.
- Graves JAM (2008) Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annu Rev Genet* 42: 565-586.
- Borg B (1994) Androgens in teleost fishes. *Comp Biochem Physiol Part C* 109: 219-245.
- Brunner B, Hornung U, Shan Z, Nanda I, Kondo M, et al. (2001) Genomic organization and expression of the doublesex-related gene cluster in vertebrates and detection of putative regulatory regions for DMRT1. *Genomics* 77: 8-17.
- Chardard D, Kuntz S, Chesnel A, Flament S (2003) Effects of androgens on sex differentiation of the urodele *Pleurodeles waltl*. *J Exp Zool A Comp Exp Biol* 296: 46-55.
- El-Mogharbel N, Wakefield M, Deakin JE, Tsend-Ayush E, Grützner F, et al. (2007) DMRT gene cluster analysis in the platypus: new insights into genomic organization and regulatory regions. *Genomics* 89: 10-21.
- Gautier A, Le Gac F, Lareyre JJ (2011) The *gsdf* gene locus harbors evolutionary conserved and clustered genes preferentially expressed in fish previtellogenic oocytes. *Gene* 472: 7-17.
- Gautier A, Sohm F, Joly JS, Le Gac F, Lareyre JJ (2011) The proximal promoter region of the zebrafish *gsdf* gene is sufficient to mimic the spatio-temporal expression pattern of the endogenous gene in Sertoli and granulosa cells. *Biol Reprod* 85: 1240-1251.
- Gnessi L, Emidi A, Jannini EA, Carosa E, Maroders M, et al. (1995) Testicular development involves the spatiotemporal control of PDGFs and PDGF receptors gene expression and action. *J Cell Biol* 131: 1105-1121.
- Gnessi L, Basciani S, Mariani S, Arizzi M, Spera G, et al. (2000) Leydig cell loss and spermatogenic arrest in platelet-derived growth factor (PDGF)-A-deficient mice. *J Cell Biol* 149: 1019-1025.
- Kim S, Kettlewell JR, Anderson RC, Bardwell VJ, Zarkower D (2003) Sexually dimorphic expression of multiple doublesex-related genes in the embryonic mouse gonad. *Gene Expr Patterns* 3: 77-82.
- O'Donnell L, Stanton PG, Wreford NG, Robertson DM, McLachlan RI (1996) Inhibition of 5-alpha-reductase activity impairs the testosterone-dependent restoration of spermiogenesis in adult rats. *Endocrinology* 137: 2703-2710.
- Shibata Y, Paul-Prasanth B, Suzuki A, Usami T, Nakamoto M, et al. (2010) Expression of gonadal soma derived factor (GSDF) is spatially and temporally correlated with early testicular differentiation in medaka. *Gene Expr Patterns* 10: 283-289.
- Smith CA, McClive PJ, Hudson Q, Sinclair AH (2005) Male-specific cell migration into the developing gonad is a conserved process involving PDGF signaling. *Dev Biol* 284: 337-350.
- Zaccanti F, Petrini S, Rubatta ML, Stagni AM, Giorgi PP (1994) Accelerated female differentiation of the gonad by inhibition of steroidogenesis in amphibia. *Comp Biochem Physiol A* 107: 171-179.

Amemiya CT, Alföldi J, Lee AP, Fan S, Brinkmann H, et al. (in preparation) Comparative analysis of the genome of the African coelacanth, *Latimeria chalumnae*, sheds light on tetrapod evolution. In preparation.

Pallavicini A, Canapa A, Barucca M, Alföldi J, Biscotti MA, et al. (in preparation) Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. In preparation.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29: 644-652.

Canapa A, Olmo E, Forconi M, Pallavicini A, Makapedua MD, et al. (2012) Composition and phylogenetic analysis of vitellogenin coding sequences in the Indonesian coelacanth *Latimeria menadoensis*. *J Exp Zool B Mol Dev Evol* 2012, 318: 404-416.

Makapedua DM, Barucca M, Forconi M, Antonucci N, Bizzaro D, et al. (2011) Genome size, GC percentage and 5mC level in the Indonesian coelacanth *Latimeria menadoensis*. *Mar Genomics* 2011, 4: 167-172.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, et al. (2006) KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4: 259-263.

Nei M, Gojobori T (1986): Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3: 418-426.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28: 2731-2739.

Zhang, J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* 95: 3708-3713.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) ClustalW and ClustalX version 2 (2007). *Bioinformatics* 23: 2947-2948.

Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. *Trends Genet* 19: 362-365.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294: 2310-2314.

Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in protein. *Atlas Protein Seq Struct* 5: 345-352.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8: 275-282.

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (* and other methods) version 4. Sunderland, MA: Sinauer Associates.

Itoh N, Konishi M (2007) The zebrafish *fgf* family. *Zebrafish* 4: 179-186.

- Sun YL, Zeng S, Ye K, Yang C, Li MH, et al. (2012) Involvement of FGF9/16/20 subfamily in female germ cell development of the Nile tilapia, *Oreochromis niloticus*. *Fish Physiol Biochem* DOI: 10.1007/s10695-012-9630-4.
- Herpin A, Schartl M (2011) Dmrt1 genes at the crossroads: a widespread and central class of sexual development factors in fish. *FEBS J* 278: 1010-1019.
- Matson CK, Zarkower D (2012) Sex and the singular DM domain: insights into sexual regulation, evolution and plasticity. *Nat Rev Genet* 13: 163-174.
- Chue J, Smith CA (2011) Sex determination and sexual differentiation in the avian model. *FEBS J* 278, 1027-1034.
- Raymond CS, Shamu CE, Shen MM, Seifert KJ, Hirsch B, et al. (1998) Evidence for evolutionary conservation of sex-determining genes. *Nature* 391: 691-695.
- Rhen T, Schroeder T (2010) Molecular mechanisms of sex determination in reptiles. *Sex Dev* 4: 16-28.
- Anand A, Patel M, Lalremruata A, Singh AP, Agrawal R, et al. (2008) Multiple alternative splicing of DMRT1 during gonadogenesis in Indian muggler, a species exhibiting temperature-dependent sex determination. *Gene* 425: 56-63.
- Guo Y, Cheng H, Huang X, Gao S, Yu H, et al. (2005) Gene structure, multiple alternative splicing, and expression in gonads of zebrafish DMRT1. *Biochem Biophys Res Commun* 330: 950-957.
- Zhao Y, Lu H, Yu H, Cheng H, Zhou R (2007) Multiple alternative splicing in gonads of chicken DMRT1. *Dev Genes Evol* 217: 119-126.
- Shinomiya A, Otake H, Togashi K, Hamaguchi S, Sakaizumi M (2004) Field survey of sex-reversals in the medaka, *Oryzias latipes*: genotypic sexing of wild populations. *Zoolog Sci* 21: 613-619.
- Bogart JP, Balon EK, Bruton MN (1994) The chromosomes of the living coelacanth and their remarkable similarity to those of one of the most ancient frogs. *J Hered* 85: 322-325.
- Kim Y, Kobayashi A, Sekido R, DiNapoli L, Brennan J, et al. (2006): Fgf9 and Wnt4 Act as Antagonistic Signals to Regulate Mammalian Sex Determination. *PLoS Biol.* 4:e187.
- Nakatani M, Takehara Y, Sugino H, Matsumoto M, Hashimoto O, et al. (2008) Transgenic expression of a myostatin inhibitor derived from follistatin increases skeletal muscle mass and ameliorates dystrophic pathology in mdx mice. *FASEB J* 22: 477-487.
- Higasa K, Nikaido M, Saito TL, Yoshimura J, Suzuki Y, et al. (2012) Extremely slow rate of evolution in the HOX cluster revealed by comparison between Tanzanian and Indonesian coelacanths. *Gene* 505: 324-332.
- Johnsen H, Andersen Ø (in press): Sex dimorphic expression of five dmrt genes identified in the Atlantic cod genome. The fish-specific dmrt2b diverged from dmrt2a before the fish whole-genome duplication. *Gene*, in press.
- Bratus A, Slota E (2009) Comparative cytogenetic and molecular studies of DM domain genes in pig and cattle. *Cytogenet Genome Res* 126: 180-185.

- Smith CA, Hurley TM, McClive PJ, Sinclair AH (2002) Restricted expression of DMRT3 in chicken and mouse embryos. *Gene Expr Patterns* 2:69-72.
- Winkler C, Hornung U, Kondo M, Neuner C, Duschl J, et al. (2004) Developmentally regulated and non-sex-specific expression of autosomal *dmrt* genes in embryos of the Medaka fish (*Oryzias latipes*). *Mech Dev* 121: 997-1005.
- Ooe H, Chen Q, Kon J, Sasaki K, Miyoshi H, et al. (2012) Proliferation of rat small hepatocytes requires follistatin expression. *J Cell Physiol* 227: 2363-2370.
- Rey R, Lukas-Croisier C, Lasala C, Bedecarrás P (2003) AMH/MIS: what we know already about the gene, the protein and its regulation. *Mol Cell Endocrinol* 211: 21-31.
- El-Awady MK, El-Garf W, El-Houssieny L (2004) Steroid 5alpha reductase mRNA type 1 is differentially regulated by androgens and glucocorticoids in the rat liver. *Endocr J* 51: 37-46.
- Barrionuevo F, Georg I, Scherthan H, Lécureuil C, Guillou F, et al. (2009) Testis cord differentiation after the sex determination stage is independent of Sox9 but fails in the combined absence of Sox9 and Sox8. *Dev Biol* 327: 301-312.
- Bertrand S, Camasses A, Somorjai I, Belgacem MR, Chabrol O, et al. (2011): Amphioxus FGF signaling predicts the acquisition of vertebrate morphological traits. *Proc Natl Acad Sci USA* 108: 9160-9165.

Characterization of purine catabolic pathway genes in coelacanths

Mariko Forconi¹, Maria Assunta Biscotti¹, Marco Barucca^{1*}, Francesco Buonocore², Gianluca De Moro³, Anna Maria Fausto², Marco Gerdol³, Alberto Pallavicini³, Giuseppe Scapigliati², Manfred Schartl⁴, Ettore Olmo¹, Adriana Canapa¹

Abstract

Coelacanths are an extremely precious source for understanding dynamics underlying genetic changes in the transition from aquatic to terrestrial life. One of the many interesting and biologically relevant features of the genus *Latimeria* is its ureotelic mode of eliminating waste nitrogen from the metabolism. Urea is, however, not excreted from the body but high concentrations are retained in the plasma and used as an osmoregulatory system. The purine catabolism pathway, which leads to urea production in *Latimeria*, has undergone a progressive step reduction indicating an enzyme loss during diversification of terrestrial species.

In the work presented here, analyses were performed on liver and testis transcriptomes of the Indonesian coelacanth *L. menadoensis* and on the recently fully sequenced genome of *L. chalumnae* in the framework of the ongoing coelacanth genome project. Besides *uricase*, *5-hydroxyisourate hydrolase*, *parahox neighbour B*, *allantoinase*, and *allantoicase* genes, coding respectively for the five enzymes involved in the urate degradation to urea, we report the identification of a putative second form of *5-hydroxyisourate hydrolase*, which is characteristic for the genus *Latimeria*.

Moreover the results of this work highlight the activity of this complete pathway in coelacanth liver and suggest its involvement in the maintenance of the high urea concentration in plasma.

Pontastacus leptodactylus:

I dati di sequenziamento Illumina di *Pontastacus leptodactylus* derivano da uno studio volto ad indagare i cambiamenti nei livelli d'espressione genica nell'epatopancreas di organismi privati del peduncolo oculare causati dall'applicazione dei due esomeri dell'ormone iperglicemico dei crostacei (cHH)

L'RNA è stato estratto da tessuto epato-pancreatico di 12 esemplari di femmine di *P. leptodactylus* divise in quattro gruppi: ai primi due gruppi è stato iniettato, rispettivamente, l'esomero D e L-cHH in PBS, ad un terzo gruppo è stato iniettato solo PBS e gli esemplari del quarto gruppo, a cui non è stato esportato il peduncolo oculare, sono stati mantenuti come controllo.

Inizialmente i dati trascrittomici provenienti dall'estrazione sono stati assemblati con il software CLC Genomic Workbench.

Le 296.112.296 *reads paired-end* sono state ridotte a 289.731.590 dall'eliminazione di basi a bassa qualità.

L'assemblaggio ha prodotto 42.187 *contig* con una lunghezza media di 802 paia di basi.

Gli output delle annotazioni di Gene Ontology ed Interpro sono stati elaborati con uno script in Python in modo da poter essere importati in CLC Genomics Workbench per implementare il test ipergeometrico nelle annotazioni.

Successivamente è stato effettuato il *mapping* e l'analisi statistica dei campioni provenienti dai 12 esemplari per valutare i cambiamenti nei livelli d'espressione.

Come trascritti di riferimento per il *mapping* è stato scelto di mantenere solo *contig* con alta copertura (maggiore di 25) per evitare i potenziali problemi causati dall'eventuale presenza di frammenti poco espressi non direttamente correlati all'epatopancreas.

I 6.860 *contig* scelti corrispondevano al 96% dell'espressione totale di questo organo.

Anche in questo caso il *mapping* è stato fatto usando il CLC Genomic Workbench e ha dimostrato come vi sia una notevole variazione nell'espressione genica tra i gruppi in analisi.

In particolare 214 trascritti si sono dimostrati differenzialmente espressi negli individui a cui era stato solo esportato il peduncolo oculare. E' stato possibile così valutare come l'attività metabolica dei

carboidrati è stata repressa, come ci si aspettava dall'eliminazione dell'ormone inducente un aumento della glicemia.

Una notevole variazione è stata osservata anche tra gli individui a cui è stato iniettato il D-cHH. Questi, infatti, hanno avuto una risposta molto rapida nella quale 917 trascritti hanno avuto una modificazione nei loro livelli d'espressione.

Gli organismi a cui è stato iniettato l'isomero L-chh, invece, hanno mostrato avere un effetto meno pronunciato con solo 45 geni differenzialmente espressi.

In un secondo momento da questo organismo sono stati estratti campioni di RNA proveniente dall'organo Y, organo secretore dell'ormone stimolatore di muta ecdisone, e dal vicino tessuto ipodermico.

Questi dati trascrittomici provengono da un lavoro che si propone di studiare le proteine partecipanti al processo di muta.

Le 155.534.379 *reads* così ottenute sono state come prima cosa assemblate con CLC Genomic Workbench per ottenere 76.405 *contig*.

Anche in questo le sequenze risultanti sono state annotate con Blast2Go.

Durante questo lavoro di assemblaggio si è riusciti ad ottenere l'account presso il cluster grid DIAG.

Grazie a questo account abbiamo potuto utilizzare l'*assembler* Trinity per un ulteriore assemblaggio che mettesse insieme i dati proveniente sia dal tessuto epato-pancreatico che da Y organ,.

I dati risultanti dall'assemblaggio sono stati successivamente elaborati con una versione locale di CD-HIT (Huang et al., 2010) per eliminare i *contig* ridondanti ottenuti, passando così da 110.406 a 100.737 *contig*.

Per eliminare la ridondanza e per valutare quale *assembler* svolgesse questo compito le sequenze output del CD-HIT sono state ulteriormente assemblate tra loro con diversi software di assemblaggio.

Alla fine si è scelto di utilizzare CAP3 (Huang et al., 2010) che, nonostante sia un software sviluppato da diversi anni, ha dimostrato, in questo caso, essere la scelta migliore producendo meno errori degli altri.

Dopo aver elaborato le sequenze con CAP3 siamo riusciti ad ottenere 94.482 *contig* che sono al momento alla base dei nuovi studi sul trascrittoma di *L. leptodactylus*.

Hepatopancreatic transcriptome in the crayfish *Pontastacus leptodactylus* reveals peptidase activation and glycolysis suppression following injection of D-crustacean Hyperglycemic Hormone.

Chiara Manfrin¹°, Moshe Tom²°, Gianluca De Moro¹, Marco Gerdol¹, Alessandro Mosco¹, Alberto Pallavicini¹, Piero Giulio Giulianini¹*

Abstract

The crustacean Hyperglycemic Hormone (cHH) is a neuropeptide present in many decapods. Two different chiral isomers are simultaneously present in Astacid crayfish and their specific biological functions are still poorly understood. The present study is aimed at better understanding the potentially different effect of each of the isomers on the hepatopancreatic gene expression profile in the crayfish *Pontastacus leptodactylus*, in the context of short term hyperglycemia. Hence, two different chemically synthesized cHH enantiomers, containing either L- or D-Phe³, were injected to the circulation of intermolt females following removal of their X organ-Sinus gland complex. The effects triggered by the injection of the two alternate isomers were detected after one hour through measurement of circulating glucose levels. Triggered changes of the transcriptome expression profile in the hepatopancreas were analyzed by RNA-seq analysis. A whole transcriptome shotgun sequence assembly provided the assumedly complete transcriptome of *P. leptodactylus* hepatopancreas, followed by RNA-seq analysis of changes in the expression level of many genes caused by the application of each of the hormone isomers. The hemolymph measurements revealed a much higher hyperglycemic activity in response to the D-isoform than to the L-isoform injection. Similarly, the RNA-seq analysis confirmed a stronger effect on gene expression following the administration of D-cHH, and just limited alterations were caused by the L-isomer. These findings demonstrated a more prominent short term effect of the D-cHH on the transcription profile and shed light on the effect of the D-isomer on specific functional gene groups. Another contribution of the study is the construction of a *de novo* assembly of the hepatopancreas transcriptome, consisting of 42,144 contigs, that dramatically increases the molecular information available for this species and provides an efficient tool, which enables gene expression experiments in this organ.

Keywords: *Pontastacus leptodactylus*, cHH chirality, hepatopancreas, transcriptome, RNA-seq analysis, Illumina sequencing

Introduction

The freshwater astacid *Pontastacus leptodactylus*, commonly called narrow-clawed crayfish, inhabits Western Asian and Eastern European lakes and watercourses. Its reproductive season includes mating from December to January and spawning in January. Eggs are incubated glued to the female pleopods till July [1]. Ovarian development takes place from June to November [2]. Induced molt through surgical intervention in *P. leptodactylus* maintained at 19°C, led to a premolt period of 17 days ending with ecdysis (personal observations of the present authors). Adults naturally shed their exoskeleton in summer just after hatching, even though some of them molt also in autumn [3].

Crustacean hyperglycemic hormones (cHHs) are a pleiotropic crustacean-specific neuropeptide family, functioning in a variety of physiological processes, recently reviewed by several authors [4,5,6,7,8]. The cHH family is divided into two subfamilies on the basis of their primary structure: (a) the cHH subfamily and (b) the molt-inhibiting hormone (MIH), the mandibular organ inhibiting hormone (MOIH) and the vitellogenesis/ gonad-inhibiting hormone (V/GIH) subfamily. The translated neuropeptides in the tissues as well as their isoforms derived from post-translational modifications and their modes of action were described only partially. Hence, a neurohormone name does not necessarily imply its entire range of functions. Several cHH variants are occasionally co-existing in a single species. The variability can emerge from both different primary sequence and different post-translational modifications [9,10]. Recently, chirality was observed also in a lobster VIH due to L to D alteration in the fourth N terminal amino acid, a tryptophan residue [11] demonstrating larger extent of the phenomenon in crustaceans. cHHs are produced in the neurosecretory perikarya sited in the medulla terminalis of the optic ganglion, located in the crustacean eyestalk and named X-organ. The X-organ secretes the neuropeptides into the hemal sinus gland and the entire neuroendocrine complex is abbreviated XOSG. Structurally, the cHH prepropeptide is composed of a signal peptide, a cHH precursor related peptide (CPRP) and a mature peptide of 72 amino acids. The role of the CPRP is still unknown, but CPRP structures, post-translational modifications and individual-related distribution have already been described [12,13]. The mature cHH contains six cysteine residues that form three disulfide bridges and potentially possesses an amidated C-terminus and a pyroglutamate blocked N-terminus [6]. Documented physiological processes influenced by the eyestalk ablation are vitellogenesis [14], food intake, digestion, and nutrient transport [15], molting [16], metabolism of lipids [17,18], regulation of glucose and proteins in hemolymph [17,19], hydromineral balance, regeneration and pigment regulation [20]. Several cHH members generally have an inhibitory effect, as the removal of the XOSG causes induction of both molt and reproduction. cHHs are produced also in other tissues: the

pericardial organ, the subesophageal ganglia, and the fore- and hindguts. The more comprehensively studied action of the cHH is to regulate carbohydrate metabolism. Its secretion follows a circadian rhythm, with a low concentration during the day which increases in the first hours of the night, and it is correlated to a similar daily pattern of the glycemia [21]. The injection of cHH induces a quick hyperglycemic response in treated animals. Apart from glucose metabolism, cHH mediates other metabolic functions of the hepatopancreas that represents the site of synthesis and secretion of digestive enzymes (amylases, proteases, lipases, and others) [22,23,24]. It is also involved in metabolism of proteins, lipids, and carbohydrates [25], as well as in the catabolism of organic compounds and in detoxification [26,27]. Indeed, cHH stimulates amylase secretion [28], and the release of free fatty acids and phospholipids [29] from the midgut gland. D-cHH is also involved in the control of molt, exerting its function by inhibiting the synthesis of ecdysone in the Y-organ and having an activity 10 times higher than L-cHH [30]. Contrasting activities were reported for the CHH's regulation of reproduction, probably due to species specificity [31,32,33,34].

The structure of the genes, the derived precursors and peptides have been recently reviewed by Webster and colleagues [7], while the dynamics of biosynthesis and release of cHH isoforms in *Orconectes limosus* have been clarified by Ollivaux and Soyeux [35]. cHH peptide sequences were *de novo* sequenced through a multifaceted mass spectrometry approach by Jia and colleagues [36].

Although *P. leptodactylus* is not considered a model organism, among decapods, it is one of the most studied species with about 143 published papers regarding its physiology [1,37,38,39], its resistance to different types of pollution and stress conditions [40,41,42] and concerning the cHH and its variety of functions [43,44,45,46]. Unfortunately this variety of studies does not reflect the fast progress of genomics and transcriptomics. Already in 1988 Sedlmeier [28] demonstrated that eyestalk factors have a direct effect on the digestive gland and the present study was aimed at the definition of the transcriptome of the hepatopancreas.

The effect of the two isomers of cHH on gene expression in relation to their variety of assigned functions in the relevant tissues is still lacking. In this study we evaluated the effect of the two cHHs isoforms after an hour post-injection both by measuring the circulating glucose levels and by studying the hormonal effect on gene expression in the hepatopancreas with a digestive gland whole transcriptome analysis. Beside this gene expression experiment, the construction of a *de novo* assembly of the hepatopancreas transcriptome, dramatically increased the molecular information available for this species and undoubtedly provides an useful tool for further gene expression experiments in this organ.

Materials and Methods

Crayfish maintenance and experimental design

Adult *P. leptodactylus* specimens were obtained alive from an Armenian freshwater source on July, 2011. They were kept for two weeks before the experiment in 120 L tanks provided with closed circuit filtered and thoroughly aerated tap water at ~18°C and were fed by fish pellets (Sera granular, Heisenberg, Germany) three times per week. Only females with no abdominally incubated ova were taken for the induction experiment. The females were at the end of the ova incubation period.

The chemical synthesis of the peptides and the glucose level induction protocol were accomplished according to Mosco et al. [43] with a few modifications detailed below. The two synthetic cHH hormone isomers, D-cHH and L-cHH, were injected to the circulation. Thirty-two *P. leptodactylus* females were divided into four groups each composed of 8 females. Two of the groups were injected with D- and L-cHH, respectively (0.5 µg/female in 100 µl PBS). A control group was sham-injected by the hormone carrier (S). Bilateral ablation of the XO-SG, aimed at prevention of any possible interference due to endogenous cHH was performed 48 hours before the injections, and a fourth group of naïve females (N) was added to the experiment as control just before injections. All 32 females were sacrificed one hour post-injection. The incubation period included ten minutes of anesthesia in ice before sacrifice and all efforts were made to minimize animal's suffering. Immediately prior to the hormonal injection the females were bled for the evaluation of the pre-induction hemolymphatic glucose level. A second bleeding was performed just before sacrifice. Hemocytes were pelleted from the sampled hemolymph and the serum was kept on ice for later glucose measurement, which was performed using a glucose oxidase method (glucose liquid mono reagent, Hospitex diagnostics, Italy). The statistical analyses of glucose levels recorded in the four groups were performed using R, version 2.14.1 software [47] as follows: the normality of data was checked with a Shapiro-Wilk test and homogeneity of variance across groups was checked with a Bartlett test. The null hypotheses of both tests could not be rejected, hence, differences of glucose levels among the experimental groups were tested using non-parametric statistics, Kruskal-Wallis rank sum test with post-hoc Wilcoxon rank sum test pairwise comparisons with Bonferroni correction. Box and whiskers plots were drawn with the boxplot command of R. Glucose levels are expressed throughout as mean ± standard error. Carapax length was measured from its posterior edge to the base of the eye cavity just before sacrifice and a sample of hepatopancreatic tissue of ~5x5x5 mm was dissected out and immediately snap-frozen in liquid nitrogen. The gastroliths were also taken for molt stage evaluation according to

Shechter et al. [48] using the molt mineralization index (MMI) (gastrolith width/crayfish carapace length).

RNA extraction and sequencing

Twelve females among the 32 were randomly chosen for the RNA sequencing. RNA was extracted from frozen tissues, homogenized in TriReagent RNA isolation solution (Sigma-Aldrich, Cat UN2821) following the manufacturer's instructions. The resulted RNAs were further purified using the RNeasy kit (Qiagen, manufacturer's instructions). The RNA level was quantified by spectrophotometer and its quality was examined using capillary electrophoresis (BioAnalyzer 2100, Agilent).

RNA sequencing was carried out at the Applied Genomics Institute (IGA, Udine, Italy), on an Illumina sequencer **HiSeq2000**. The hepatopancreas reference transcriptome assembly was derived from a 2X100 bp paired-end sequencing performed on a cDNA library obtained from equimolar amounts of RNA taken from all experimental females. Gene expression was evaluated using the one-sided 50 bp Illumina sequencing from the 12 distinct RNAs.

Sequences analysis

The processing and analysis of the obtained raw sequences was carried out using the CLC Genomics Workbench 4.5 software (CLC Bio, Aarhus, Denmark). Raw sequence reads were trimmed according to base calling quality. The resulting 2X100 bp sequence reads were assembled assuming a paired reads distance comprised between 100 and 600 base pairs, and setting the penalties for mismatches to 2, insertions and deletions to 3 and similarity and length fraction to 0.9 and 0.5, respectively. The minimum allowed assembled contig length was set to 200 bp. The obtained contig assembly served as a comprehensive reference for the functional genomics RNA-seq analysis [49] of the *P. leptodactylus* hepatopancreas. We determined the presumptive amount of conceptual full length transcripts by using the Full-lengther Next webtool [50] considering alignments starting before the 10th aa and with an e-value below 1e-04. Interspersed repeats and low complexity DNA sequences were detected with RepeatMasker version open-3.3.0 (default mode). Filtered contigs displaying an average coverage less than 25X of mapped 50bp reads were discarded prior to the RNA-seq analysis for creating a robust set of contigs not subjected to random expression fluctuations, corresponding to the 95th percentile of the genes expressed in the hepatopancreas.

The 50 bp sequencing reads from the 12 different samples (namely, N1, N2, N3, S1, S2, S3, L1, L2, L3, D1, D2 and D3) were individually mapped on the filtered reference set, considering a maximum number of mismatches cost of 3, a maximum of 10 hits for a read and expression values were calculated based on unique gene reads. The Baggerly's test [51] was used to identify statistically significant differential expression using S as reference group; a FDR corrected p-value <0.01 was set as threshold of significant differential expression [52] and an additional threshold of minimum fold change of 2 was also implemented. The similarity among the profiles obtained from this study was examined by hierarchical clustering (complete linkage, average linkage and single linkage) using the Pearson correlation coefficient as a distance measure. An alternative clustering was performed by Principal Component Analysis (PCA).

Transcripts annotation and their expression pattern

Differentially expressed transcripts were characterized with the Blast2Go platform [53,54]. The characterized parameters were resemblance to genes with known function using BLASTx algorithm [55] against the NCBI non-redundant protein databases with an e-value cut-off of 10^{-6} . The default Blast2Go resemblance annotations were also checked manually against the list of resemblances obtained for each contig by BLASTx, to conform to the *UniProt* nomenclature guidelines and to select the annotations in a more educated manner, based on the below described GO terms and domain characterization. Blast2Go was used also to annotate the contigs assigning Gene Ontology (GO) functional terms [56] and Interpro domains annotations [57] with default settings. The GO and Interpro annotation outputs were modified with scripts developed in-house and imported into the CLC Genomics Workbench 4.5 environment for implementing the hypergeometric test [58] on the annotations. Significantly altered GO terms and Interpro domains were detected with this test considering a p-value threshold of 0.01 and a difference between observed and expected >1.

Results and Discussion

Morphological and physiological state of the animals

The morphometric characteristics and the molt stage of the experimental females as well as their hemolymphatic glucose levels before injection and at their sacrifice are presented in **Table 1**. The glucose level of all the 3 groups of eyestalk-less females was highly significantly lower than the native ones (Wilcoxon rank sum test: $p < 0.01$; **Figure 1A**), but after an hour the injection of both cHHs was able to restore glycemia to almost normal levels (D-cHH and L-cHH injected animals vs native animals, Wilcoxon rank sum test: $p = 1$; **Figure 1B**). On the contrary, the sham-injected animals still presented a significantly lower glycemia compared to all other groups (Wilcoxon rank sum test: $p < 0.02$). The carapax length was similar among the four experimental groups, namely N, S, L and D (ANOVA and post hoc tests, $p < 0.05$). All females were at intermolt with $MMI < 0.01$ and all ovaries were immature according to Hubenova et al.[2].

De novo assembly of the P. leptodactylus hepatopancreas transcriptome

The Illumina 2X100 bp sequencing of the hepatopancreas of adult *P. leptodactylus* females, generated 296,112,296 nucleotide reads. The number was reduced to 289,731,590 after quality trimming. The *de novo* assembly by the CLC Genomic Workbench produced 42,144 contigs with an average length of 802 bp. This assembly was used as a reference transcript library for subsequent analyses. A total of 9,474 contigs longer than 1Kb was obtained, while the longest contig almost reached 15 Kb in length. **Table 2** summaries the trimming and assembly statistics. The raw Illumina reads were stored at the NCBI Sequence Read Archive (SRA: SRR650486), whereas 42,144 assembled contigs were deposited at NCBI Transcriptome Shotgun Assembly (TSAXXXX). Through the Full-Lenghter Next webtool [50] we determined the presumptive percentage of the full length transcripts present in our library. Hence, 1/3rd (32.2%) of the transcripts were putatively assembled to their full length, 1/3rd (32.6%) covered either the N-terminal or the C-terminal region, whereas the remaining 35.2% consisted of internal fragments.

From the output of Blast2Go we checked the fifteen Top-hits species sharing similarity with the *P. leptodactylus* reference library. The three most represented species were *Nematostella vectensis* (Cnidaria), *Daphnia pulex* (Crustacea) and *Tribolium castaneum* (Insecta). Interestingly, a higher similarity was observed with the coelenterate in respect to the two arthropods, reasonably explained by the high amount of nucleotide sequences belonging to *N. vectensis* (47,065) stored at NCBI, in comparison to those of *D. pulex* (7,309) and of *T. castaneum* (16,790). The first fifteen BLAST top-hits species are reported in the Supporting Information section (**Figure S 1**). To date, the only crustacean whose genome has been fully sequenced is the branchiopod *Daphnia pulex* [59]. Even though a few deep transcriptome sequencing approaches were applied to decapods [60,61], no study

has ever targeted Astacidae. An astacoid epithelial assembly from *Cherax quadricarinatus*, performed by 454 was simultaneously prepared by the present co-authors and their collaborators and will be soon released for the public (Acc. Number GADE00000000). At present, only 275 nucleotide sequences belonging to the genus *Pontastacus* are stored at NCBI, while only 67 originated from *Pontastacus leptodactylus*. Next generation sequencing greatly simplifies large-scale molecular studies of non-model organisms, and this transcriptome sequencing with the assembly of 42,144 contigs, represents the first large-scale sequencing approach in this genus, as well as a remarkable contribution to the genetic knowledge of decapods, paving the way to straightforward comparative molecular studies on non-model crustaceans.

RNA-seq analysis of cHH isomer effects

An average of 12.7 ± 5.2 million short reads were obtained from the sequencing of RNAs extracted from the hepatopancreas of each sampled female. An average of 50.8 ± 6.9 % reads were mapped to the reference transcriptome. **Table 3** reports in detail the number of reads obtained from each adult female. The RNA-seq mappings of the 12 sequenced samples (N1, N2, N3, S1, S2, S3, L1, L2, L3, D1, D2 and D3) were used for the gene expression analysis. Only a selected set of 6,860 contigs with an average coverage $>25X$ was used for the analysis, reducing the noise potentially caused by the high number of fragmented or poorly expressed transcripts, not strictly related to the hepatopancreas, as highlighted by the Full-length analysis. This selected set of transcripts accounts for about 96% of the total expression in this organ. The stringency applied to the contigs selection (FDR corrected p-value <0.01 and a weighted fold change at least of 2), identified the prominently differentially expressed transcripts in the pairwise comparisons, by using the S group as reference for all the analyses. **Table 4** summarizes the numbers of differentially expressed genes while the complete list of differentially expressed genes is available in the Supporting Information (**Table S 1**). The expression profiles of the various females, namely the list of mapped reads per each gene in each sample were clustered by two methods, hierarchical clustering (**Figure S 2A**) and a principal component analysis (**Figure S 2B**), which highlighted a large difference between D-cHH-injected animals and all the other groups, which in turn clustered together and did not display detectable differences in this preliminary analysis.

Effects of XOSG removal on the hepatopancreas gene expression

The comparison between the S and N groups was useful to depict the effects derived solely from the XOSG removal procedure. XOSG extirpation resulted in the differential expression of 214

transcripts, with up-regulation (147 sequences, 68.7%) exceeding down-regulation (67 sequences, 31.3%). It is likely that the transcripts characterized by an increase of expression in respect to the naïve state are connected to two main events, the injury and the reduced production of several hormones caused by the XOSG removal. 63% of the up-regulated and 67% of the down-regulated genes showed no resemblance to known sequences. Among the genes which demonstrated similarity we identified in particular some transcripts connected to glycolysis process, such as transcript containing enolase domain (IPR000941), probably connected to the removal of cHHs production sites.

The hypergeometric test highlighted the carbohydrate metabolism (GO:0005975) and the chitinase activity (GO:0004568) as the most relevant GO terms repressed by XOSG extirpation; and many associated Interpro domains, e.g. glycoside hydrolase, family 9 (IPR001701), six-hairpin glycosidase (IPR012341) and chitinase II (IPR011583) were significantly repressed. Many other GO terms and Interpro signatures were found to be up-regulated, even though most of them could not be easily linked to any reported physiological effect of the extirpation. The most studied process in many crustacean species triggered by XOSG extirpation is undoubtedly vitellogenesis [62,63,64], but there is no uniform vitellogenesis-related physiological response to the extirpation.

As for other genes, in pandalid shrimp, eyestalk ablation down-regulated the hepatopancreatic chitinase expression and also in *P. leptodactylus* we observed a significant repression of the IPR domain of the chitinase II (IPR011583) within the S group. This transcript pertains to the group 1 reported in Salma et al. [65] that represent chitinases produced in hepatopancreatic tissues that may function in the digestion of ingested chitin and the modification of peritrophic membrane in the intestine. Other studies [15,66] showed that XOSG extirpation caused an increase in respiratory rate and a decrease in metabolic activity with respect to intact shrimps. Similarly, in *P. leptodactylus* the repression of three transcripts associated to metabolic process (GO:0008152) was observed in the S group, differently to *Marsupenaeus japonicus* [67] in which fructose-1,6-bisphosphatase was not significantly changed after XOSG removal, we observed an up-regulation of a transcript containing the fructose-1,6-bisphosphatase domain (IPR000146). This gene is involved in gluconeogenesis and cHH could affect also this metabolic pathway. XOSG removal affects also the circulating levels of hormones other than the cHH. These include the cHH type II hormones [7], the Molt Inhibiting Hormone (MIH), the Gonad/Vitellogenesis Inhibiting Hormone (GIH/VIH) and the Mandibular Organ-Inhibiting Hormone (MOIH). Hence, pathways controlled by these hormones may potentially be affected by the extirpation, an effect that would be difficult to discriminate due to the pleiotropic nature of the cHH members, affecting multiple processes [68].

Differential effects of the two cHH-isomers on gene expression.

The injection of the two cHHs, distinguished only by chirality, showed different effects on gene expression. The comparison with the S group revealed only 45 differentially expressed transcripts in response to the administration of the L-isomer, whereas 917 transcripts were responsive to the D-isomer. Among the transcripts differentially up-regulated after L-cHH injection we depicted an increase of aminopeptidase n-like and trypsin 4, both correlated to proteolytic activity (GO:0006508), while all the down-regulated ones were without similarity. Moreover no functional or structural feature (GO term or domain) could be significantly linked to L-cHH injection by the hypergeometric test.

To examine a possible overlap in the effects exerted by L- and D-cHH we identified differentially expressed transcripts shared by the two treatments. Six transcripts elucidated differential expression cause by both treatments, but the majority of them had different trend of expression. No transcripts shared up-regulation caused by both treatments, while 2 non-annotated transcripts were mutually down-regulated (Pastle_hepa_c8804 and Pastle_hepa_c22565). Three transcripts were up-regulated after L-cHH, but down-regulated following D-cHH injection; among them a trypsin 4, a transcript containing chitinase II domain (IPR011583) and a non-annotated one. Only one non annotated transcript resulted to be up-regulated in D and down-regulated in L group.

In the light of the above results, the analysis was mainly focused on the variations between the S and the D groups.

Table 5 presents the Interpro domains and the GO terms altered in the in D group in comparison to the S group as revealed by the hypergeometric test.

As previously shown by the PCA analysis, the D-cHH administration caused pronounced alterations in comparison to the sham injected females, dominated by down-regulated genes (857 transcripts, accounting for 93.5%). The hypergeometric test elucidated a few D-cHH up-regulated functions, and in particular several GO terms related to peptidases, including cysteine-type peptidase (GO:0008234) and endopeptidase activity (GO:0004197), the papain domain of peptidase C1A, (IPR013128) and the cathepsin propeptide domain of proteinase inhibitor I29 (IPR013201), which is usually found in the N-terminus of many proteinases.

On the contrary, glycolysis (GO:0006096) appeared to be the GO term most significantly repressed in the D-cHH-injected female crayfish, which may be related to the return of the glycemic level to

the native state, and suggests a modulation activity of the D-cHH on the glucose and energy metabolism in the hepatopancreas.

Given the absence of the source of endogenous cHH, it was of interest to assess whether the injection of exogenous hormone was able to restore, even partially, the native conditions. Obviously, XOSG-ablation does not only remove the primary organ of cHH synthesis (cHH type I), but also prevents the secretion of other cHH superfamily hormones (cHH type II), possibly causing a series of gene expression changes not solely related to the deprivation of cHH. Therefore, the administration of synthetic cHH was only expected to partially restore the gene expression profile observed prior to XOSG ablation. In fact, about 1/3th (50 out of 147) of the transcripts up-regulated in the S group were restored at lower level following the D-cHH injection (**Figure 2A**). Among them we identified transcripts related to several different cellular processes and molecular functions, including cellular organization (i. e. actin and tubulin beta 2), transcripts related to RNA processing and protein synthesis, (i.e, mRNA turnover protein 4-like protein), a transcript containing Fibrillar domain (IPR000692) that plays a role in ribosomal RNA processing and lysyl-tRNA synthetase. Two cuticle proteins were also detected, playing a role in calcium ion binding; an armet-like protein precursor containing EF-Hand 1, calcium-binding site domain (IPR018247) and crustin 1 antimicrobial peptide were also restored by the D-cHH injection. Figure 2A shows two transcripts with an expression fold-change higher than the others, one encoding a C-reactive protein (CRP) containing a pentaxin domain (IPR001759) and the other transcript coding for a cuticle protein which contains the Reberse and Riddiford arthropod chitin binding domain (IPR000618). CPR is a fundamental player in the inflammatory response, in particular binding to the phosphocholine expressed on the surface of dead or dying cells in order to activate the complement system via the C1Q complex. Eyestalk ablation did not cause an increase in the level of CPR mRNA in *M. japonicus* [69]. In the present study and in a different tissue an up-regulation of about 9.5-fold was depicted following the XOSG-ablation, subsequently restored by the D-cHH administration. But most notably, the expression of four transcripts crucially involved in glycolysis, which were increased in the S group, returned to lower levels after the D-cHH administration. Namely, phosphoglycerate mutase 1, two highly similar enolases catalyzing the eighth and ninth steps of glycolysis respectively, and fructose-1,6-bisphosphatase, an allosteric regulator of pyruvate kinase, the final step of glycolysis, all followed the same expression trends. This supports the hypothesis that D-cHH could have mainly an inhibitory effect on transcripts up-regulated following XOSG ablation and a minor effect on those silenced by the ablation, that probably required other molecules and centers of control located in the removed eyestalk. Hence, only 5% of the repressed genes upon ablation (3 out of 60) changed their trend of expression after D-cHH administration (**Figure 2B**). One of them resembles the 3-hydroxybutyrate

dehydrogenase, related to oxydoreductase and metabolic processes and the two remaining transcripts do not have an annotated function.

The L-configuration provides the peptide with a strictly hyperglycemic activity whereas the D-cHH may exhibit, in addition to a strong hyperglycemic potency. The finding of several transcripts containing a sodium symporter domain (IPR002657) suggests that D-cHH may also regulate glucose absorption by the hepatopancreas, a process that occurs through a Na⁺/D-glucose co-transport [70].

These data all together suggest that, despite being apparently able to restore glycemia to almost normal levels in eye-ablated animals one hour post injection (see **Table 1**), L-cHH does not significantly influence gene expression in the hepatopancreas. Therefore the apparent positive effect of the two neuropeptides on glycemia, at least after a short period, is either provided by a specific action of different tissues, or more likely still on the hepatopancreas, but through different mechanisms, one involving gene expression and the other one involves post transcriptional enzyme activation. It is also possible that the effect of L-cHH on gene expression would be only visible at longer periods from its injection, after more than one hour. An additional experiment, similar to the one described here, but implementing a prolonged time-course and sampling of additional tissues has already been performed in our laboratory, and is currently being analyzed to provide better insight to the presented alternative hypotheses.

Conclusions

The effects of two cHH enantiomers on a multi-gene expression profile of *P. leptodactylus* were examined here. The presence of both cHH isomers has been demonstrated only in Astacoidea [71]. The scarcity of *P. leptodactylus* sequenced transcripts required the construction of a comprehensive

transcriptomic library of its hepatopancreas of 42,144 contigs obtained through Illumina sequencing. This library was used as a reference gene assembly in a RNA-seq experiment. A short term effect of two enantiomers of the peptidic hyperglycemic hormone (D-cHH and L-cHH) of *P. leptodactylus* on hepatopancreatic gene expression was examined at two levels, the specific gene level and the functional group level.

The overall effect of the different treatments was demonstrated by the clustering of the 12 individual transcript profiles, which resulted in the independent clusterization of D-cHH treated females in respect with all the other samples.

The XOSG synthesizes and secretes multiple peptidic hormones, including D- and L-cHH. The effect of its removal on gene expression profiles revealed the differential expression of 214 transcripts, 147 up-regulated and 67 down-regulated. The carbohydrate metabolic activity was repressed as expected from the elimination of the glycemia inducing hormone.

The L-cHH injected individuals revealed an expression profile similar to sham injected individuals and to the native non- treated ones. Moreover, D-cHH caused a considerable short term change in the hepatopancreatic gene expression profile. 917 transcripts were responsive to this isomer: 857 out of them were down regulated and only 60 were up-regulated. The functions mainly affected were the up-regulated proteolytic activity and down-regulated glycolysis. L-cHH caused much less pronounced effect, with only 45 differentially expressed genes (30 up-regulated and 15 down-regulated), without the detectable alteration of any specific function.

Hypothesized glycemia-related mechanisms of L- and D- cHH may involve protein activation with no changes in gene expression for both isomers and an additional D-cHH mechanism which involves changes in gene expression, which may be related or not to the increase of glucose levels in the circulation.

D-cHH undoubtedly acted on the molecular patterns of gene expression of the digestive gland, as clearly highlighted by the number of differentially expressed transcripts triggered after the hormone injection and by the marked functional alterations identified by the hypergeometric test. This latter test, distinguished an opposite trend related to the glycolysis in the XOSG ablated group (S) characterized by the up regulation of the enolase (IPR000941) that resulted to be significantly repressed after the D-cHH administration. Moreover, this short-time study showed that D-cHH was apparently able to restore the expression of a certain number of transcripts to the level of the naïve state, and it was also able to affect molecular pathways not altered by the XOSG ablation.

Acknowledgements

The authors acknowledge the computer resources and technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga, Spain

Figures

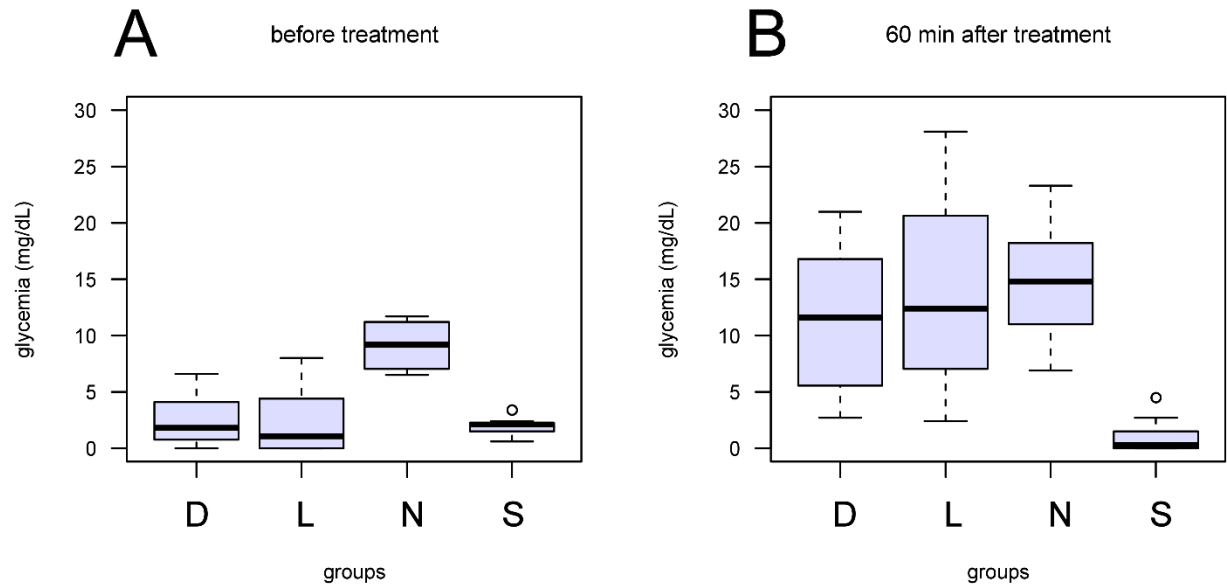


Figure 1: Box plot of the glycemia values obtained in the four groups S – sham injected, N – intact, L – L-cHH-injected, D- D-cHH-injected at the beginning of the experiment and one hour post-injection.

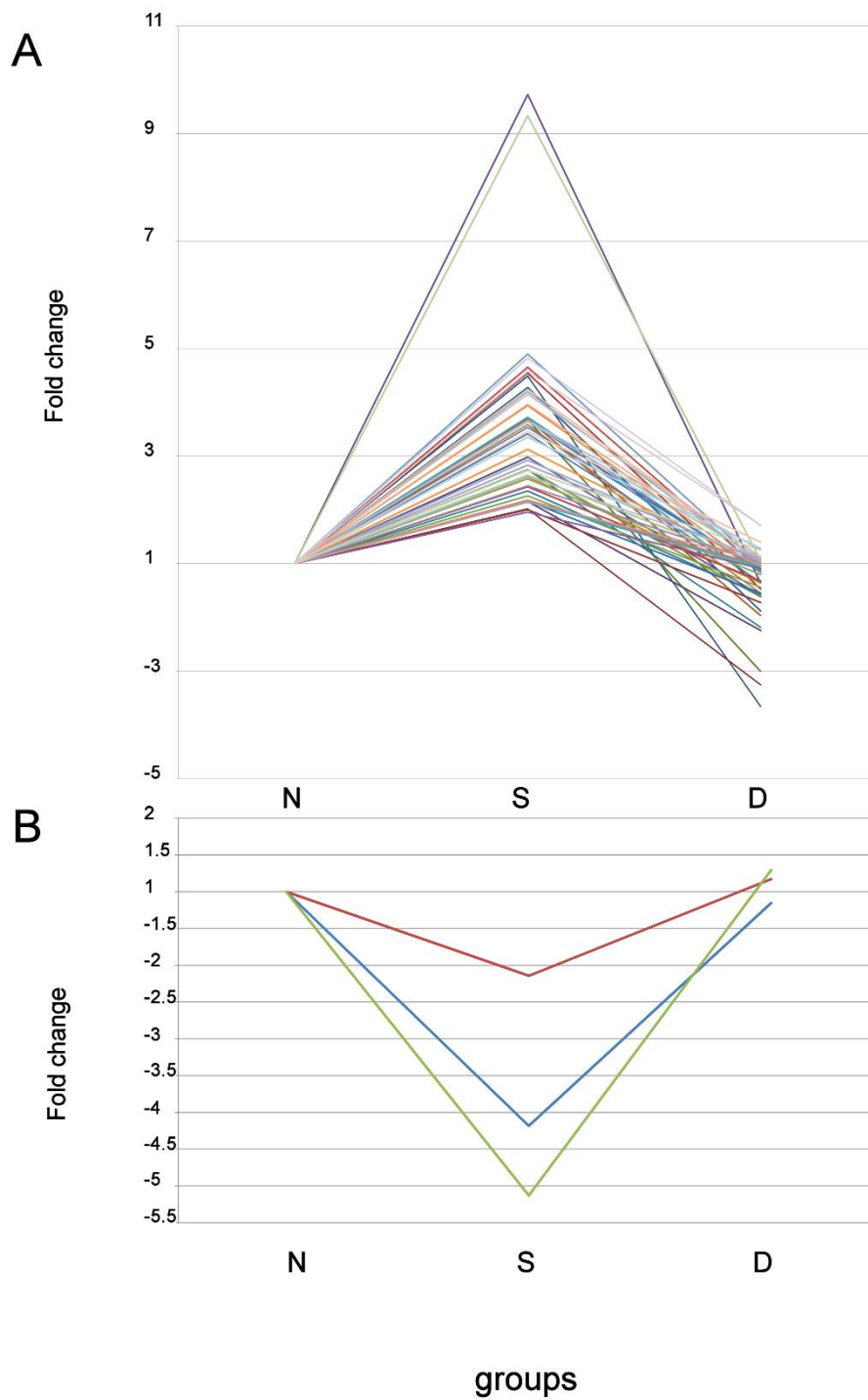


Figure 2: A Expression trend of the transcripts which were up-regulated after XOSG ablation and restored their original level following D-cHH injection.

B Expression trend of the transcripts which were down-regulated after XOSG ablation and up-regulated following D-cHH injection.

Supporting Information Figure captions

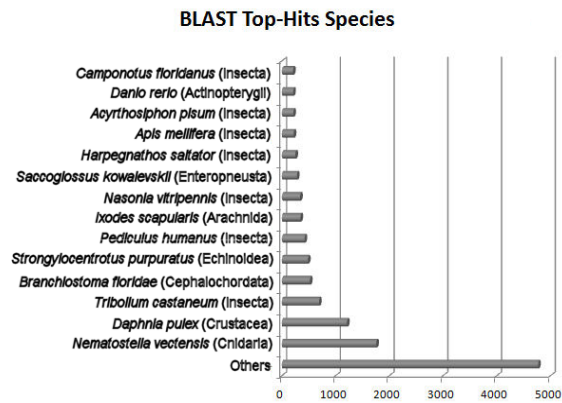


Figure S 1: The distribution of BLAST 15 Top Hit Species matching *P. leptodactylus* contigs.

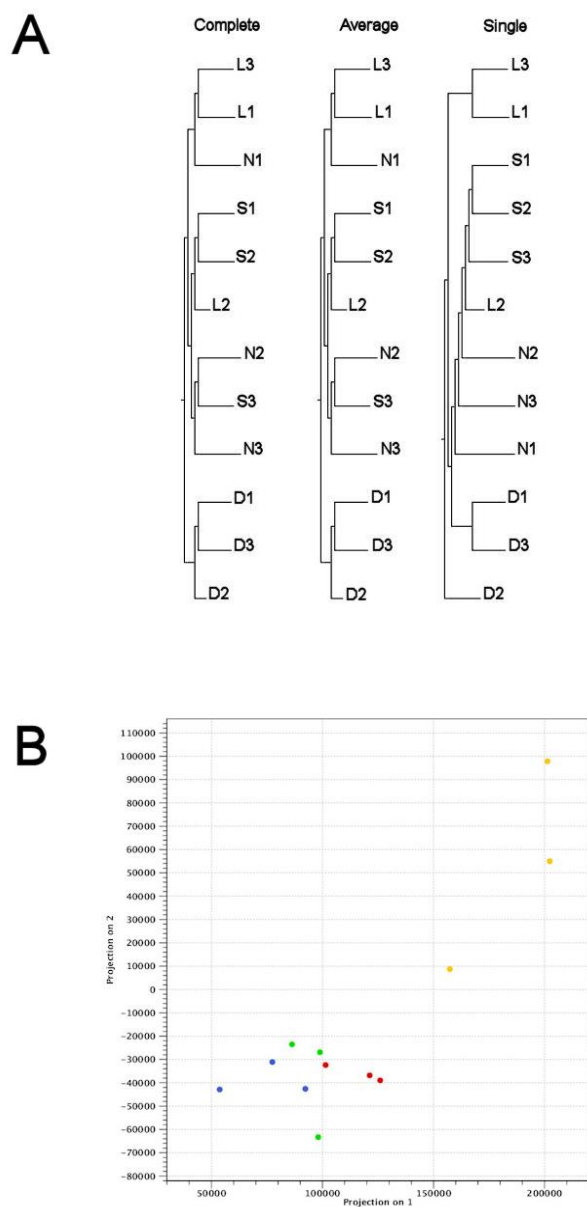


Figure S 2 A: Hierarchical clustering of the 12 mapping profiles. S – sham injected, N – intact, L – L-cHH-injected, D- D-cHH-injected. **B:** principal component analysis of the 12 mapping profiles. Green – native, Red – sham, Blue – L-cHH-injected and Yellow – D-cHH-injected.

Tables

Experimental group [n=8 in each group]	Carapax length [mm]	*MMI - Gastrolith width/Carapax length [mm/mm]	Glucose [mg/dL] levels	
			0min.	60min
Native	1.2±39.4	0.0048±0.0067	0.7±9.14 5	1.93±14.78
Sham-injected	2.4±39.4	0.0086±0.0094	0.3±1.94 3	0.59±1.01
L-cHH-injected	2.0±41.5	0.0065±0.0081	1.1±2.36 0	3.13±13.84
D-cHH-injected	1.7±39.9	0.0001±0.0025	0.8±2.49 3	2.32±11.45

Table 1: Morphometric characteristics, molt stage of the experimental females, their hemolymphatic glucose levels before injection and at their sacrifice. *MMI – molt mineralization index.

Trimming statistics	
Number of reads before trimming	296,112,296
Number of reads after trimming	289,731,590
Paired-end reads after trimming	283,857,480
Single reads after trimming	5,874,110
Sequences discarded during trimming	6,380,706
Average length before trimming	101.0
Average length after trimming	97.2
Assembly statistics	
Assembly size [bps]	33,838,039
Total number of contigs	42,187
N50	1208
N80	454
N90	318
Mean contig length	802
Median contig length	449
Longest contig	14,918
Number of contigs longer than 1 Kb	9,474
GC content	%40.44

Table 2: Trimming statistics and assembly summary of the Reference dataset. N50, 80 and 90 are the quantiles corresponding to the 50, 80 and 90 percentiles, respectively.

Sample	Total number of reads	Mapped reads (%)
N1	23,784,636	13,086,802 (55.0%)
N2	10,985,623	5,738,820 (52.2%)
N3	17,933,507	9,093,453 (50.7%)
S1	8,589,134	4,239,139 (49.4%)
S2	12,303,830	5,904,274 (48.0%)
S3	14,403,132	6,722,862 (46.7%)
L1	10,579,155	6,740,537 (63.7%)
L2	13,929,531	7,154,589 (51.4%)
L3	11,168,774	6,111,228 (54.7%)
D1	6,239,023	2,596,219 (41.6%)
D2	12,781,748	4,431,787 (34.7%)
D3	7,379,961	3,173,378 (43.0%)

Table 3: Total number of sequencing reads obtained per animal and the number of the reads mapped and the corresponding percentage, N= native group, S= sham-injected, L= L-cHH injected and D= D-cHH injected

Group	Number of differentially expressed transcripts	Up regulated	Down regulated
S vs N	214	147	67
L vs S	45	30	15
D vs S	917	60	857

Table 4: Number of differentially expressed transcripts in each of the experimental groups. FDR corrected p-value <0.01, minimum difference in fold change = 2. S= sham-injected, L= L-cHH injected and D= D-cHH injected.

Group	Trend	Annotation	p-value	Description
S vs N	+	IPR006195	0.000117	Aminoacyl-tRNA synthetase, class II
S vs N	+	IPR002930	0.000979	Glycine cleavage H-protein
S vs N	+	IPR004365	0.000979	Nucleic acid binding, OB-fold, tRNA/helicase-type
S vs N	+	IPR020809	0.000979	Enolase, conserved site
S vs N	+	IPR020810	0.000979	Enolase, C-terminal
S vs N	+	IPR000537	0.00288	UbiA prenyltransferase family
S vs N	+	IPR000941	0.00288	Enolase
S vs N	+	IPR004364	0.00288	Aminoacyl-tRNA synthetase, class II (D/K/N)
S vs N	+	IPR003604	0.00288	Zinc finger, U1-type
S vs N	+	IPR018150	0.00288	Aminoacyl-tRNA synthetase, class II (D/K/N)-like
S vs N	+	IPR011053	0.00288	Single hybrid motif
S vs N	+	IPR001790	0.00564	Ribosomal protein L10/acidic P0
S vs N	+	IPR016027	0.00681	Nucleic acid-binding, OB-fold-like
S vs N	+	IPR012340	0.00681	Nucleic acid-binding, OB-fold
S vs N	+	IPR012675	0.00921	Beta-grasp domain
S vs N	+	GO:0005960	0.000979	Glycine cleavage complex
S vs N	+	GO:0004659	0.000979	Prenyltransferase activity
S vs N	+	GO:0030414	0.00288	Peptidase inhibitor activity
S vs N	+	GO:0000287	0.00847	Magnesium ion binding
S vs N	+	GO:0004634	0.00921	Phosphopyruvate hydratase activity
S vs N	+	GO:0004867	0.00921	Serine-type endopeptidase inhibitor activity
S vs N	-	IPR001701	1.36E-05	Glycoside hydrolase, family 9
S vs N	-	IPR012341	1.89E-05	Six-hairpin glycosidase
S vs N	-	IPR008928	1.89E-05	Six-hairpin glycosidase-like
S vs N	-	IPR017853	0.000453	Glycoside hydrolase, superfamily

S vs N	-	IPR011583	0.0023	Chitinase II
S vs N	-	IPR001223	0.0023	Glycoside hydrolase, family 18, catalytic domain
S vs N	-	IPR013781	0.00314	Glycoside hydrolase, catalytic domain
S vs N	-	GO:0005975	0.000496	Carbohydrate metabolic process
S vs N	-	GO:0043169	0.000584	Cation binding
S vs N	-	GO:0003824	0.000858	Catalytic activity
S vs N	-	GO:0004568	0.0023	Chitinase activity
vs S D	+	IPR013128	0.0000548	Peptidase C1A, papain
vs S D	+	IPR013201	0.000777	Proteinase inhibitor I29, cathepsin propeptide
vs S D	+	IPR000668	0.0024	Peptidase C1A, papain C-terminal
vs S D	+	GO:0008234	0.000113	Cysteine-type peptidase activity
vs S D	+	GO:0004197	0.00544	Cysteine-type endopeptidase activity
vs S D	-	IPR000941	0.0058	Enolase
D vs S	-	IPR011042	0.0058	Six-bladed beta-propeller, TolB-like
vs S D	-	GO:0006096	0.000818	Glycolysis
vs S D	-	GO:0030234	0.00444	Enzyme regulator activity

Table 5: List of the Interpro domains and GO terms differentially regulated by XOSG removal (S) or by D-cHH-injection (D), detected by the Hypergeometric test. + designates up-regulation and – down-regulation.

References

- Balik I, Çubuk H, Özkök R, Uysal R (2005) Some biological characteristics of crayfish (*Astacus leptodactylus*) Eschscholtz, 1823) in Lake Eğirdir. *Turkish Journal of Zoology* 29: 295-300.
- Hubenova TA, Vasileva PL, Zaikov AN (2009) Histological analysis of ovary development in narrow-clawed crayfish *Astacus leptodactylus* Esch. 1823 (crustaceae, decapoda, astacidae), reared in ponds in South Bulgaria. *AACL Bioflux* 2: 261-270.
- Van Herp F, Bellon-Humbert C (1978) Setal development and molt prediction in the larvae and adults of the crayfish, *Astacus leptodactylus* (Nordmann, 1842). *Aquaculture* 14: 289-301.
- Hopkins PM (2012) The eyes have it: A brief history of crustacean neuroendocrinology. *General and Comparative Endocrinology* 175: 357-366.
- Fanjul-Moles ML (2006) Biochemical and functional aspects of crustacean hyperglycemic hormone in decapod crustaceans: Review and update. *Comparative Biochemistry and Physiology - C Toxicology and Pharmacology* 142: 390-400.
- Giulianini P, Edomi, P (2006) Neuropeptides controlling reproduction and growth in Crustacea: a molecular approach. In: Satake H, editor. *Invertebrate Neuropeptides and Hormones: Basic Knowledge and Recent Advances Kerala, India: Research Signpost*. pp. 225-252.
- Webster SG, Keller R, Dirksen H (2012) The CHH-superfamily of multifunctional peptide hormones controlling crustacean metabolism, osmoregulation, moulting, and reproduction. *General and Comparative Endocrinology* 175: 217-233.
- Chung JS, Zmora N, Katayama H, Tsutsui N (2010) Crustacean hyperglycemic hormone (CHH) neuropeptides family: Functions, titer, and binding to target tissues. *General and Comparative Endocrinology* 166: 447-454.
- Hsu YWA, Weller JR, Christie AE, de la Iglesia HO (2008) Molecular cloning of four cDNAs encoding prepro-crustacean hyperglycemic hormone (CHH) from the eyestalk of the red rock crab *Cancer productus*: Identification of two genetically encoded CHH isoforms and two putative post-translationally derived CHH variants. *General and Comparative Endocrinology* 155: 517-525.
- Lee KJ, Doran RM, Mykles DL (2007) Crustacean hyperglycemic hormone from the tropical land crab, *Gecarcinus lateralis*: Cloning, isoforms, and tissue expression. *General and Comparative Endocrinology* 154: 174-183.

- Ollivaux C, Gallois D, Amiche M, Boscaméric M, Soyez D (2009) Molecular and cellular specificity of post-translational aminoacyl isomerization in the crustacean hyperglycaemic hormone family. *FEBS Journal* 276: 4790-4802.
- Fu Q, Goy MF, Li L (2005) Identification of neuropeptides from the decapod crustacean sinus glands using nanoscale liquid chromatography tandem mass spectrometry. *Biochemical and Biophysical Research Communications* 337: 765-778.
- Stemmler EA, Hsu YWA, Cashman CR, Messinger DI, de la Iglesia HO, et al. (2007) Direct tissue MALDI-FTMS profiling of individual Cancer productus sinus glands reveals that one of three distinct combinations of crustacean hyperglycemic hormone precursor-related peptide (CPRP) isoforms are present in individual crabs. *General and Comparative Endocrinology* 154: 184-192.
- Palacios E, Carreño D, Rodríguez-Jaramillo MC, Racotta IS (1999) Effect of eyestalk ablation on maturation, larval performance, and biochemistry of white pacific shrimp, *Penaeus vannamei*, Broodstock. *Journal of Applied Aquaculture* 9: 1-23.
- Rosas C, Fernandez I, Brito R, Diaz-Iglesia E (1993) The effect of eyestalk ablation on the energy balance of the pink shrimp, *Penaeus notialis*. *Comparative Biochemistry and Physiology - A Physiology* 104: 183-187.
- Chang ES (1985) Hormonal control of molting in decapod crustacea. *Integrative and Comparative Biology* 25: 179-185.
- Teshima SI, Kanazawa A, Koshio S, Horinouchi K (1989) Lipid metabolism of the prawn *Penaeus japonicus* during maturation: Variation in lipid profiles of the ovary and hepatopancreas. *Comparative Biochemistry and Physiology -- Part B: Biochemistry and Physiology* 92: 45-49.
- Santos EA, Nery LEM, Keller R, Gonçalves AA (1997) Evidence for the involvement of the crustacean hyperglycemic hormone in the regulation of lipid metabolism. *Physiological Zoology* 70: 415-420.
- Santos EA, Keller R (1993) Regulation of circulating levels of the crustacean hyperglycemic hormone: evidence for a dual feedback control system. *Journal of Comparative Physiology B* 163: 374-379.
- Keller R, Sedlmeier, D (1988) A metabolic hormone in crustaceans: the hyperglycemic hormone. In: Laufer H, Downer, FGH, editor. *Endocrinology of selected invertebrates*. New York: Alan R Liss Inc. pp. 315-326.

- Kallen JL AS, Van Herp F (1990) Circadian rhythmicity of the crustacean hyperglycemic hormone (CHH) in the hemolymph of the crayfish. *Biol Bull* 179.
- Rivera-Pérez C, Navarrete del Toro MdlÁ, García-Carreño FL (2010) Digestive lipase activity through development and after fasting and re-feeding in the whiteleg shrimp *Penaeus vannamei*. *Aquaculture* 300: 163-168.
- Vogt G, Stfcker, W, Storeh, V, Zwilling, R (1989) Biosynthesis of *Astacus* protease, a digestive enzyme from crayfish. *Histochemistry* 91: 373-381.
- Castro PF, Freitas Jr ACV, Santana WM, Costa HMS, Carvalho Jr LB, et al. (2012) Comparative study of amylases from the midgut gland of three species of penaeid shrimp. *Journal of Crustacean Biology* 32: 607-613.
- Ceccaldi H (1989) Anatomy and physiology of digestive tract of Crustaceans Decapods reared in aquaculture. *Advances in tropical aquaculture. Tahiti*. pp. 243-259.
- James M (1989) Cytochrome P450 monooxygenases in crustaceans. *Xenobiotica* 19: 1063-1076.
- Ahearn GA, Mandal, P K, Mandal, A (2004) Mechanisms of heavy-metal sequestration and detoxification in crustaceans: a review. *Journal of Comparative Physiology B* 174: 439-452.
- Sedlmeier D (1988) The crustacean hyperglycemic hormone (CHH) releases amylase from the crayfish midgut gland. *Regul Peptides* 20: 91-98.
- Santos E, Nery, LEM, Keller, R, Gonçalves, AA (1997) Evidence for the involvement of the crustacean hyperglycemic hormone in the regulation of lipid metabolism. *Physiol Zool* 70: 415-420.
- Yasuda A, Yasuda, Y, Fujita, T, Naya, Y (1994) Characterization of crustacean hyperglycemic hormone from the crayfish (*Procambarus clarkii*): multiplicity of molecular forms by stereoinversion and diverse function. *Gen Comp Endocrinol* 95: 387-398.
- Khayat M, Yang, W, Aida, K, Nagasawa, H, Tietz, A, Funkenstein, B, Lubzens, E (1998) Hyperglycemic hormones inhibit protein and mRNA synthesis in in vitro-incubated ovarian fragments of the marine shrimp *Penaeus semisulcatus*. *Gen Comp Endocrinol* 110: 307-318.
- Avarre JC KM, Michelis R, Nagasawa H, Tietz A, Lubzens E (2001) Inhibition of de novo synthesis of a jelly layer precursor protein by crustacean hyperglycemic hormone family peptides and posttranscriptional regulation by sinus gland extracts in *Penaeus semisulcatus* ovaries. *Gen Comp Endocrinol* 124: 257-268.

- Tsutsui N KH, Ohira T, Nagasawa H, Wilder MN, Aida K (2005) The effects of crustacean hyperglycemic hormone-family peptides on vitellogenin gene expression in the kuruma prawn, *Marsupenaeus japonicus*. *Gen Comp Endocrinol* 144: 232-239.
- de Kleijn D, Janssen, KPC, Waddy, SL, Hegeman, R, Lai, WY, Martens, GJM, Van Herp, F (1998) Expression of the crustacean hyperglycaemic hormones and the gonad-inhibiting hormone during the reproductive cycle of the female American lobster *Homarus americanus*. *Journal of Endocrinology* 156: 291–298.
- Ollivaux C, Soyeux D (2000) Dynamics of biosynthesis and release of crustacean hyperglycemic hormone isoforms in the X-organ-sinus gland complex of the crayfish *Orconectes limosus*. *European Journal of Biochemistry* 267: 5106-5114.
- Jia C, Hui L, Cao W, Lietz CB, Jiang X, et al. (2012) High-definition de novo sequencing of Crustacean Hyperglycemic Hormone (CHH)-family neuropeptides. *Molecular and Cellular Proteomics* 11: 1951-1964.
- Susanto G, Charmantier, G (2001) Crayfish freshwater adaptation starts in eggs: ontogeny of osmoregulation in embryos of *Astacus leptodactylus*. *J Exp Zool* 289: 433-440.
- Harlioglu A, Aydin, S, Yilmaz, O (2012) Fatty acid, cholesterol and fat-soluble vitamin composition of wild and captive freshwater crayfish (*Astacus leptodactylus*). *Food Sci Technol Int* 18: 93-100.
- Giulianini P, Bierti, M, Lorenzon, S, Battistella, S, Ferrero, EA (2007) Ultrastructural and functional characterization of circulating hemocytes from the freshwater crayfish *Astacus leptodactylus*: cell types and their role after in vivo artificial non-self challenge. *Micron* 38: 49-57.
- Tunca E UE, Ozkan AD, Ulger ZE, Tekinay T (2013) Tissue Distribution and Correlation Profiles of Heavy-Metal Accumulation in the Freshwater Crayfish *Astacus leptodactylus*. *Arch Environ Contam Toxicol* Epub ahead of print.
- Barim O, Karatepe, M (2010) The effects of pollution on the vitamins A, E, C, beta-carotene contents and oxidative stress of the freshwater crayfish, *Astacus leptodactylus*. *Ecotoxicol Environ Saf* 73: 138-142.
- Malev O, Srut, M, Maguire, I, Stambuk, A, Ferrero, EA, Lorenzon, S, Klobucar, GI (2010) Genotoxic, physiological and immunological effects caused by temperature increase, air exposure or food deprivation in freshwater crayfish *Astacus leptodactylus*. *Comp Biochem Physiol C Toxicol Pharmacol* 152: 433-443.

Mosco A, Zlatev V, Guarnaccia C, Pongor S, Campanella A, et al. (2012) Novel protocol for the chemical synthesis of crustacean hyperglycemic hormone analogues - an efficient experimental tool for studying their functions. *PLoS ONE* 7.

Mosco A, Edomi, P, Guarnaccia, C, Lorenzon, S, Pongor, S, Ferrero, EA, Giulianini, PG (2008) Functional aspects of cHH C-terminal amidation in crayfish species. *Regul Pept* 147: 88-95.

Lebaupain F, Boscameric, M, Pilet, E, Soyez, D, Kamech, N (2012) Natural and synthetic chiral isoforms of crustacean hyperglycemic hormone from the crayfish *Astacus leptodactylus*: hyperglycemic activity and hemolymphatic clearance. *Peptides* 34: 65-73.

Serrano L, Grousset, E, Charmantier, G, Spanings-Pierrot, C (2004) Occurrence of L- and D- crustacean hyperglycemic hormone isoforms in the eyestalk X-organ/sinus gland complex during the ontogeny of the crayfish *Astacus leptodactylus*. *J Histochem Cytochem* 52: 1120-1140.

Team RDC (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Shechter A, Tom M, Yudkovski Y, Weil S, Chang SA, et al. (2007) Search for hepatopancreatic ecdysteroid-responsive genes during the crayfish molt cycle: From a single gene to mutagenicity. *Journal of Experimental Biology* 210: 3525-3537.

Wang ZY, Gerstein, M., Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 57-63.

Fernández-Pozo N, Guerrero-Fernández, D., Bautista, R., Gonzalo Claros, M. Full-LengtherNext: A tool for fine-tuning de novo assembled transcriptomes of non-model organisms. Paper in preparation, personal communication.

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003) Differential expression in SAGE: Accounting for normal between-library variation. *Bioinformatics* 19: 1477-1483.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc* 57: 289-300.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* 36: 3420-3435.

- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25: 3389-3402.
- Ashburner M, Lewis S, Subramaniam, Noble, Kanehisa (2002) On ontologies for biologists: The gene ontology - Untangling the web. pp. 66-83.
- Zdobnov EM, Apweiler R (2001) InterProScan - An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847-848.
- Falcon S, Gentleman, R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* 23: 257.
- Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, et al. (2011) The ecoresponsive genome of *Daphnia pulex*. *Science* 331: 555-561.
- Kawahara-Miki R, Wada K, Azuma N, Chiba S (2011) Expression profiling without genome sequence information in a non-model species, pandalid shrimp (*Pandalus latirostris*), by next-generation sequencing. *PLoS ONE* 6.
- Jung H, Lyons RE, Dinh H, Hurwood DA, McWilliam S, et al. (2011) Transcriptomics of a giant freshwater prawn (*Macrobrachium rosenbergii*): De Novo assembly, annotation and marker discovery. *PLoS ONE* 6.
- Uawisetwathana U, Leelatanawit R, Klanchui A, Prommoon J, Klinbunga S, et al. (2011) Insights into eyestalk ablation mechanism to induce ovarian maturation in the black tiger shrimp. *PLoS ONE* 6.
- Chaves AR (2000) Effect of x-organ sinus gland extract on [³⁵S] methionine incorporation to the ovary of the red swamp crawfish *Procambarus clarkii*. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology* 126: 407-413.
- Sudha K, Anilkumar G (2007) Elevated ecdysteroid titer and precocious molt and vitellogenesis induced by eyestalk ablation in the estuarine crab, *Metopograpsus messor* (Brachyura: Decapoda). *Journal of Crustacean Biology* 27: 304-308.

- Salma U, Uddowla MH, Kim M, Kim JM, Kim BK, et al. (2012) Five hepatopancreatic and one epidermal chitinases from a pandalid shrimp (*Pandalopsis japonica*): Cloning and effects of eyestalk ablation on gene expression. *Comparative Biochemistry and Physiology - B Biochemistry and Molecular Biology* 161: 197-207.
- Venkitraman PR, Jayalakshmy KV (2008) Effect of eyestalk ablation on the metabolic activities of two penaeid prawns: *Metapenaeus monoceros* and *M. dobsoni* (De Man). *Italian Journal of Zoology* 75: 345-360.
- Nagai C, Nagata S, Nagasawa H (2011) Effects of crustacean hyperglycemic hormone (CHH) on the transcript expression of carbohydrate metabolism-related enzyme genes in the kuruma prawn, *Marsupenaeus japonicus*. *General and Comparative Endocrinology* 172: 293-304.
- Zmora N, Trant J, Zohar Y, Chung JS (2009) Molt-inhibiting hormone stimulates vitellogenesis at advanced ovarian developmental stages in the female blue crab, *Callinectes sapidus* 1: An ovarian stage dependent involvement. *Saline Systems* 5.
- Okumura T, Kim, YK, Kawazoe, I, Yamano, K, Tsutsui, N, Aida, K (2006) Expression of vitellogenin and cortical rod proteins during induced ovarian development by eyestalk ablation in the kuruma prawn, *Marsupenaeus japonicus*. *Comp Biochem Physiol A Mol Integr Physiol* 143: 246-253.
- Vilella S, Zilli, L, Ingrosso, L, Schiavone, R, Zonno, V, Verri, T, Storelli, C (2003) Differential expression of Na⁺/D-glucose cotransport in isolated cells of *Marsupenaeus japonicus* hepatopancreas. *J Comp Physiol B* 173: 679-686.
- Soyez D (2003) Recent data on the crustacean hyperglycemic hormone family; R FMN, editor. Enfield (NH), USA, Plymouth, UK: Science.

Altri organismi:

Nel corso di questo dottorato ho avuto la possibilità di lavorare con dati trascrittomici anche di altre specie.

I manoscritti derivanti dall'analisi di questi dati, però, non sono ancora in fase sufficientemente avanzata per la pubblicazione in questa tesi.

Assieme a *M. galloprovincialis* nei primi anni della tesi ho avuto modo di elaborare dati di sequenziamento trascrittomico 454 di *Ruditapes philippinarum* che ha fornito 1.288.514 sequenze.

A causa delle caratteristiche intrinseche di questo tipo di sequenze, che le differenziano dalle sequenze Illumina essendo più lunghe e con una possibilità di avere al loro interno un maggior numero di errori di sequenziamento, si è reso necessario modificare la *pipeline* di analisi.

In questo caso in alternativa al CLC Genomic Workbench, non particolarmente adatto per l'elaborazione di dati 454, è stato utilizzato Newbler, un software prodotto dalla 454 Life Sciences appositamente sviluppato per la gestione di dati proveniente da sequenziamento 454.

Questo assemblaggio ha prodotto 81.410 *contig*.

L'aver assemblato anche il trascrittoma di *R. philippinarum* ci ha permesso, tra le altre cose, di effettuare uno studio comparativo dei trascritti codificanti peptidi antimicrobici tra questo trascrittoma e il trascrittoma di *Mytilus galloprovincialis*.

Un altro organismo di cui abbiamo assemblato e analizzato il trascrittoma è il gambero della Luisiana *Procambarus clarkii*.

Dal peduncolo oculare di 4 esemplari di sesso misto è stato possibile ottenere 83.170.732 *reads paired-end* mediante sequenziamento Illumina (Hiseq 2000).

L'assemblaggio è stato elaborato con Trinity e i *contig* risultanti sono stati processati nello stesso modo dei *contig* di *Pontastacus leptodactylus*, ossia con una *pipeline* comprendente CD-HIT e successivamente CAP3.

Da questo assemblaggio è stato possibile ottenere 81.231 *contig* con una lunghezza media di ben 1036 pb e un N50 di 1860pb.

I dati ottenuti al momento sono in fase di analisi ed è in scrittura un articolo descrittivo del trascrittoma di questo organismo.

Bibliografía

- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* *11*, R106.
- Benton, M.J., and Donoghue, P.C.J. (2007). Paleontological Evidence to Date the Tree of Life. *Mol Biol Evol* *24*, 26–53.
- Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* *21*, 3674–3676.
- Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics*.
- Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics* *28*, 3169–3177.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* *29*, 644–652.
- Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* *26*, 680–682.
- Kumar, S., and Blaxter, M.L. (2010). Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* *11*, 571.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* *9*, 357–359.
- Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* *74*, 5463–5467.
- Schadt, E.E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics* *19*, R227–R240.

- Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* *26*, 1135–1145.
- Venier, P., De Pittà, C., Bernante, F., Varotto, L., De Nardi, B., Bovo, G., Roch, P., Novoa, B., Figueras, A., Pallavicini, A., et al. (2009). MytiBase: a knowledgebase of mussel (*M. galloprovincialis*) transcribed sequences. *BMC Genomics* *10*, 72.
- Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* *131*, 281–285.
- Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* *26*, 136–138.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* *10*, 57–63.
- Yao, J.Q., and Yu, F. (2011). DEB: A web interface for RNA-seq digital gene expression analysis. *Bioinformatics* *7*, 44–45.