ISJ 18: 19-32, 2020 ISSN 1824-307X

RESEARCH REPORT

Evolutionary insights on a novel mussel-specific foot protein-3α gene family

E Bortoletto[1], P Venier[1], A Figueras[2], B Novoa[2], U Rosani[1,3]*

[1]Department of Biology, University of Padua, Padua, Italy
[2]Institute of Marine Research (IIM), Spanish National Research Council (CSIC). Vigo, Spain
[3]Coastal Ecology Section, AWI - Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, List auf Sylt, Germany

 Accepted December 18, 2020

Abstract

Silky byssus threads enable a number of marine and freshwater bivalve mollusks to attach themselves to hard substrates. Byssus production is an energy-costly process, which accompany the switch from planktonic to sessile life. Pointing the attention to a small foot protein (fp-3α) first identified in *Perna viridis* and abundantly secreted during the bissogenesis, we report the presence of a fp-3α gene family in species of the *Mytilus* complex, byssogenic bivalve mollusks mostly inhabiting marine waters. In the genome of *Mytilus galloprovincialis* we identified twelve fp-3α genes showing differences in exon-intron organization and suggesting that, as in the case of arthropod and mollusk defensins, exon shuffling could have played an important role in the evolution of this gene family. Also, the different tissue expression patterns of these mussel genes support their functional diversification. All predicted fp-3α proteins curiously possess a Csαβ three-dimensional motif based on 10 highly conserved cysteines and exhibit structural similarity to invertebrate defensins. The role of these small cysteine-rich proteins in supporting the byssus-mediated mussel adhesion or their action as host defence peptides remains to be established with further study.

Key Words: byssogenesis; DOPA-containing proteins; foot proteins; fp-3α; *Mytilida; Mytilus galloprovincialis*

Introduction

The class of *Bivalvia* comprises many thousands of marine and fresh water mollusk species (WoRMS - World Register of Marine Species, 2020), including relevant aquaculture bivalves such as mussels, oysters and clams (Wijsman *et al*., 2019). Bivalves are found in a variety of aquatic habitats throughout the world, from cold-water seas to freshwater basins and even in deep anoxic vents (Darrigran and Damborenea, 2011; Karatayev *et al*., 2015; Bielen *et al*., 2016). Many bivalves live attached to hard substrates thanks to the byssus, a bunch of silky threads ending with a tiny adhesion plaque (Waite, 2017). The byssus threads with their terminal plaques are secreted by composite glandular tissues, possess an outer cuticle and span 2-6 cm in length. Such a byssus-mediated anchoring allows the aggregation of mussels in dense intertidal beds and also prevents the physical damage consequent to wave motion (Bennett and Sherratt, 2019).

The byssus was postulated to appear during the Devonian period (350-400 Mya), following a modification of the Gastropoda pedal glands (Waite, 1992). Although the anchoring process is essential to both post-larval dispersal and settlement (Sigurdsson *et al*., 1976; De Blok and Tan-Maas, 1977), the byssus structure is retained throughout life in several species (Figure 1). The byssus production is a very common life strategy in bivalves, excepting subclasses of *Protobranchia.* Actually, byssus is detectable in all orders of *Pteriomorphia* with the exception of *Ostreida*, in *Unionida* (a monophyletic order of freshwater mussels), and in all orders of *Heterodonta*, although this bivalve subclass mostly includes species adapted to a burrowing life style (Canapa *et al*., 2001).

Corresponding author:
Umberto Rosani
Department of Biology
University of Padova
Viale Giuseppe Colombo 3, 35131 Padova, Italy
E-mail: umberto.rosani@unipd.it

List of abbreviations:
amino acid, aa; expressed sequence tag, EST; open reading frame, ORF; RNA sequencing, RNA-seq; transcriptome shotgun assembly, TSA; untranslated transcript region, UTR

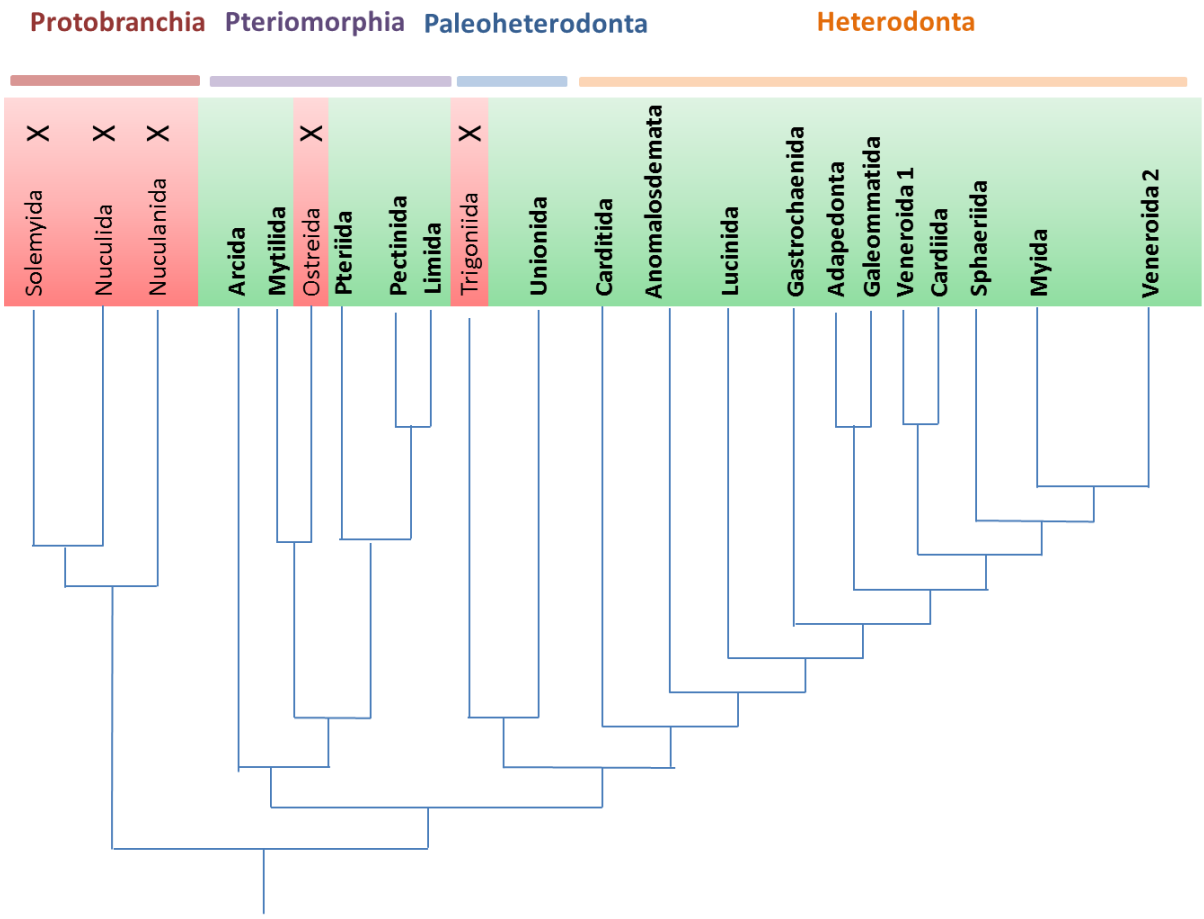

Fig. 1 Phylogenetic tree representing main orders of Bivalvia (based on González *et al*., 2015; Combosch *et al*., 2017; Lee *et al*., 2019). A cross and a red background indicate orders whose known species do not produce byssus. The orders written in bold against a green background include at least one byssus-producing species.

The byssus formation has been widely studied in *Mytilus spp.* and more than 20 proteins known as mussel foot proteins (mfps) have been identified (Waite, 2017). This process begins with the mussel foot searching for a suitable substrate and the subsequent creation of a watertight vacuum between foot and substrate (Ponder *et al*., 2020). The three types of glandular tissue involved in the byssogenesis are named core, cuticle and plaque glands in agreement to the final destination of the secreted proteins. The latter are released in secretion vesicles and move along the ventral foot grove to the different parts of each byssal thread (Harrington *et al*., 2018). It is worth noticing that the plaque gland was formerly called phenol gland, owing to its L-3,4-dihydroxyphenylalanine (DOPA)-rich proteins.

The core of byssus threads, covered by a protective cuticle, is constituted by pre-polymerized collagen (PreCOL) assembled with matrix proteins (TMPs) (Sagert and Waite, 2009) whereas the secondary structure of the core vesicle proteins, as resolved by Raman spectroscopy, resembles that of type I collagen (Priemel *et al*., 2017). In the case of KCl-induced byssogenesis, mfp1-containing vesicles released by the cuticle gland progressively aggregate in the ventral foot grove and, finally, the discharged proteins form a thin layer on the assembled thread core (Priemel *et al*., 2017). DOPA residues present in the primary structure of mfp1 mediate the formation of metal coordination complexes, which confer hardness and stiffness to the cuticle (Harrington *et al*., 2010).

The adhesive plaque is composed by at least 6 different mfps. Its formation occurs as a temporally regulated process in which the proteins involved in the adhesion are secreted first, namely mfp3 and mfp5, followed by the other proteins (Petrone *et al*., 2015). The adhesive plaque proteins are rich in DOPA residues, which easily autoxidize thus impairing the DOPA adhesion potential. This risk is prevented by the formation of disulphide bridges and cysteinyl–DOPA bonds involving cysteine-rich proteins such as mfp6 (Yu *et al*., 2011) and, in the case of *Perna viridis*, the foot protein-3α (fp-3α) protein (Petrone *et al*., 2015). Amazingly, the byssus threads are initially gel-like but quickly coagulate after the contact with the water (Ponder *et al*., 2020). Notably, byssogenesis and under-water adhesion inspired many studies aimed at the

development of novel adhesive and extensible materials (Lee *et al*., 2007; Lee *et al*., 2011; Zhong *et al*., 2014; Harrington *et al*., 2018).

In this study, we focused the attention on a small protein (fp-3α), firstly identified by amino acid sequencing in the Asian green mussel *P. viridis* (Guerette *et al*., 2013; Petrone *et al*., 2015). The considerable amount of this protein in *P. viridis* byssus threads and the fact that it was never reported in other species, initially indicated fp-3α as a *Perna*-exclusive gene product. Following genomic and transcriptomic data analyses, we present fp-3α sequences identified in some bivalve species of the order *Mytilida*. Accordingly, we report the taxonomic distribution fp-3α proteins and, in particular, we discuss the fp-3α gene family detected in *Mytilus galloprovincialis* in the light of evolution.

Table 1 Fp-3α genes retrieved from the *M. galloprovincialis* genome. For each of the eight fp-3α gene types, the transcript ID, species of origin, presence (Y) or absence (N) of the related genes in the *M. galloprovincialis* genome (with gene ID, length, and exon-intron number) are reported.

fp-3α type	Transcript ID	*Species*	MG gene [Y/N]	Gene ID	Length	exons/introns
fp3α -1	Pvir_29208	*P. viridis*	Y	MGAL10A052660, MGAL10A002276	2.98 4.64	4 exons/ 3 introns 3 exons/2 introns
fp3α -2	Medu_01889 Medu_32159 Mcal_11206 Mtro_29948 Mtro_29950 Medu_62051 Mtro_86325 Mtro_86326 Medu_62050 Medu_81265 Mgal_12191 Medu_62048	*M. edulis* *M. edulis* *M. californianus* *M. trossulus* *M. trossulus* *M. edulis* *M. trossulus* *M. trossulus* *M. edulis* *M. edulis* *M. galloprovincialis* *M. edulis*	Y	MGAL10ncA009345, MGAL10A009044	3.79 1.71	3 exons/ 2 introns 3 exons/ 2 introns
fp3α -3	Mgal_36283a Mgal_36283b Mgal_36283c Mtro_41977 Mtro_41976 Mtro_66797	*M. galloprovincialis* *M. galloprovincialis* *M. galloprovincialis* *M. trossulus* *M. trossulus* *M. trossulus*	Y	MGAL10ncA069861, MGAL10A001553, MGAL10ncA009347 MGAL10ncA073074	0.45 3.44 0.26 1.60	1 exon/0 introns 3 exons/2 introns exon/0 introns 3 exons/2 introns
fp3α -4	Medu_44249 Mgal_6748 Mcor_91378 Mcal_60322 Thir_34375 Svir_10851	*M. edulis* *M. galloprovincialis* *M. coruscus* *M. trossulus* *T. hirsuta* *M. virgata*	Y	MGAL10ncA031470	3.46	4 exons/3 introns
fp3α -5	Medu_59989 Mgal_91363 Mgal_52154 Medu_60453 Mtro_48177 Mtro_48176	*M. edulis* *M. galloprovincialis* *M. galloprovincialis* *M. edulis* *M. trossulus* *M. trossulus*	Y	MGAL10A048479	3.55	3 exons/2 introns
fp3α -6	Mcor_37770 Medu_63946 Mcor_6509 Medu_12581	*M. coruscus* *M. edulis* *M. coruscus* *M. edulis*	Y	MGAL10A031169 MGAL10A001023	3.21 4.34	3 exons/2 introns 4 exons/3 introns
fp3α -7	Pvir_01678 Pvir_37518 Pvir_02094	*P. viridis* *P. viridis* *P. viridis*	N			
fp3α -8	Pvir_18700 Pvir_03227 Pvir_22782 Pvir_29362	*P. viridis* *P. viridis* *P. viridis* *P. viridis*	N			

Material and Methods

Retrieval and preliminary analysis of bivalve sequences

The available bivalve genomes and the corresponding gene models were retrieved from public repositories. Basically, we considered 14 bivalve genomes sequenced up to now, including *Argopecten irradians* (Liu *et al*., 2020), *Bathymodiolus platifrons* (Sun *et al*., 2017), *Crassostrea gigas* (Zhang *et al*., 2012), *Crassostrea virginica* (Gómez-Chiarri *et al*., 2015), *Dreissena polymorpha* (McCartney *et al.*, 2019), *Limnoperna fortunei* (Uliano-Silva *et al*., 2018), *Mizuhopecten yessoensis* (Wang *et al*., 2017), *Modiolus philippinarum* (Sun *et al*., 2017), *Pecten maximus* (Kenny *et al*., 2020), *Pinctada fucata* (Du *et al*., 2017), *Ruditapes philippinarum* (Yan *et al*., 2019), *Saccostrea glomerata* (Powell *et al*., 2018), *Sinonovacula constricta* (Ran *et al*., 2019), *Venustaconcha ellipsiformis* (Renaut *et al*., 2018). Moreover, unassembled RNA-seq reads referring to 20 bivalve species were retrieved from the NCBI SRA archive (*Atrina pectinate, Bathymodiolus platifrons, Dreissena rostriformis, Limnoperna fortunei, Loripes orbiculatus, Mytilisepta virgata, Mytilus californianus, Mytilus coruscus, Mytilus edulis, Mytilus galloprovincialis, Mytilus trossulus, Paphia undulata, Perna viridis, Pinctada fucata, Pinna nobilis, Pyganodon grandis, Tegillarca granosa, Trichomya irsuta, Uniomerus tetralasmus, Villosa lienosa*). Raw RNA-seq data in .sra format were converted into .fastq using the fastq-dump utility of the NCBI SRA toolkit (Staff, 2011) and the reads were trimmed for the presence of adaptor sequences and for quality using TrimGalore! (Babraham Bioinformatics-Trim Galore, 2019), allowing a maximum of two ambiguous bases and a quality threshold of PHRED20. High quality reads were *de-novo* assembled with the CLC assembler (CLC Genomic Workbench v.20, Qiagen, Germany) with word and bubble size parameters set to 'automatic' and a minimal contig length of 200 bp. The obtained contigs were subjected to open reading frame (ORF) prediction using *Transdecoder* tool (Trinity suite *(Grabherr et al*., 2011)), modifying the minimal ORF length to 150 nucleotides (-m 50 parameter).

Identification of fp-3α homologs and phylogenetic analysis

Putative fp-3α sequences were extracted from the above mentioned bivalve datasets using *phmmer* (Eddy, 2011), with *P. viridis* fp-3α (AGZ84285.1) as initial query. Since no recognizable Pfam domains are present on this protein, we built a HMM profile (*hmmbuild*) based on the alignment of all the identified fp-3α proteins, then using it to scan against the protein dataset with *hmmsearch* (Eddy, 2011) and applying an E-value cut-off of 0.05. All the resulting protein sequences were aligned with MUSCLE (Edgar, 2004). ModelTest-NG v0.1.2 (Posada and Crandall, 1998) was used to assess the best-fitting model of molecular evolution for the multiple sequence alignment (identified as the WAG+I+G). Bayesian phylogenetic analysis was performed using MrBayes v3.2.6 (Ronquist *et al*., 2012). In detail, Markov Chain Monte Carlo analysis were run until reaching the convergence cutoff of 0.01, with a sampling frequency of 1,000 and a burn-in removal of 50% of the sampled trees. The convergence of parallel runs was estimated by reaching an average standard deviation of split frequency <0.05 and of a potential scale reduction factor equal to 1. Adequate posterior sampling was evaluated by reaching an effective sample size > 200 for each of the estimated parameters using Tracer v1.6 (Drummond *et al*., 2012). The final majority-rule consensus tree was visualized and edited using FigTree v1.4.3 (Rambaut- Fig Tree, 2012).

Identification of M. galloprovincialis fp-3α *genes*

In order to identify the possible fp-3α genes, we searched the fp-3α HMM profile in the *M. galloprovincialis* genomic contigs translated into the six lecture frames. According to the mussel genome annotation, we retrieved all annotated fp-3α proteins and we compared them with the transcriptome-derived ones, correcting the former in case of errors. To predict intron-exon boundaries, fp-3α transcript sequences have been back mapped on the *M. galloprovincialis* genomic contigs, using the CLC splice-aware aligner (CLC *large gap mapping tool*) and applying 0.8 of similarity on 0.2 of transcript length.

Mussel tissue sampling, RNA extraction and high-throughput sequencing

Adult mussels (*M. galloprovincialis*, shell length 5-7 cm) were collected near the Chioggia lagoon outlet (45° 14.227'N, 12° 16.706'E) during spring-summer 2019 in order to assess levels and trends of fp-3α gene transcripts. At the arrival to the laboratory, mussels were cleaned from epibionts and let to acclimatize in reconstituted seawater at 35 ‰ (Instant Ocean, Aquarium systems, Mentor, USA; 2 L/mussel) with constant aeration and feeding them once a day (Coral Food Xpro, Aquatic Nature, Oostnieuwkerkesteenweg, Belgium) at 18 °C for two days.

Mussel hemolymph was individually withdrawn from the adductor muscle with a 23G disposable needle and, following dissection, two gill samples and two hemolymph samples were selected, placed in Trizol (1 mL/50 mg tissue, 0.5 mL/pellet from 1 mL hemolymph) ThermoFisher, Bremen, Germany) and stored at −80 °C. Total RNA was extracted with RNeasy Micro kit (Qiagen, Hilden, Germany). RNA quantity and quality were checked using a Nanodrop instrument (ThermoFisher) and an Agilent RNA6000 Nanochip (Agilent Technologies, Santa Clara, CA, USA), respectively. One microgram of high-quality RNA per sample was used to construct libraries using the TruSeq Stranded mRNA kit (Illumina, California, USA) according to the kit instruction. Amplified libraries were checked for size and quality and run with a 2 × 150 read layout in a NextSeq500 instrument.

Gene expression analysis based on RNA-seq data

We also used available RNA-seq datasets of *M. galloprovincialis* to compute the expression profiles of fp-3α genes in different tissues, for a total of 30

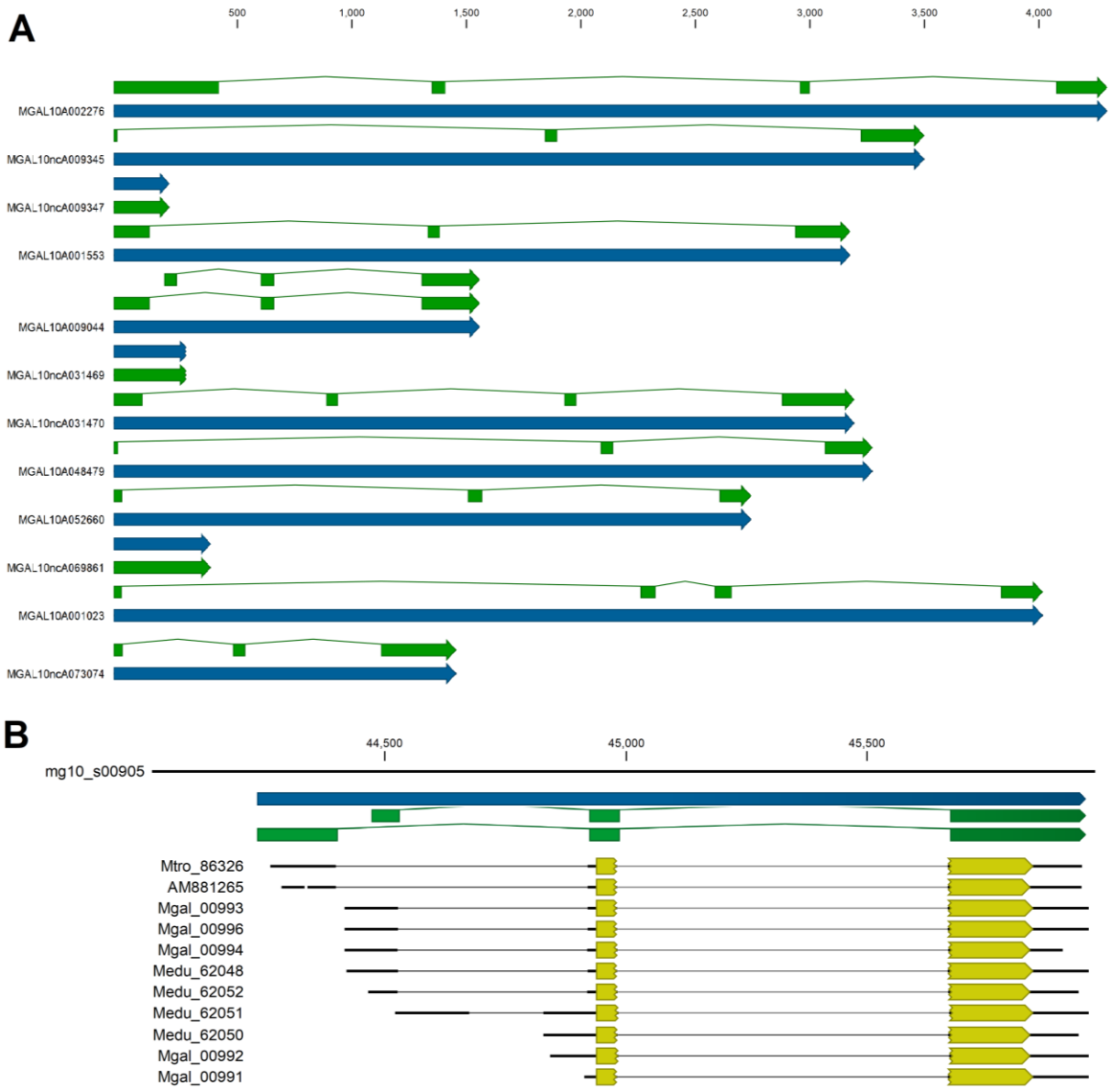

Fig. 2 Mussel fp-3α genes. A. The gene structure of the 12 *M. galloprovincialis* fp-3α genes (continuous and interrupted arrows describe the complete genes and gene exons, respectively). Notice that MGAL10A009044 has two splicing variants. B. Detail of MGAL10A009044 (arrows next to contig ID represent the ORFs)

available samples (PRJNA525609, PRJNA295512, PRJNA230138, PRJNA88481, PRJNA484309) plus the 4 samples produced for this work. Clean RNA-seq reads were mapped on the *M. galloprovincialis* genome annotated with the predicted gene models, using the CLC read mapper, setting length and similarity fractions to 0.8 and 0.8, respectively, whereas mismatch/insertion/deletion penalties were set to 3/3/3. The number of unique mapped reads of each dataset were counted and used to calculate digital expression values as Transcripts Per Million (TPM), to ensure the comparability of different datasets (Wagner *et al*., 2013).

FP-3α protein structure prediction

Molecular weight and isoelectric point were calculated for each fp-3α sequence with the ExPASy Compute pI/Mw tool (Gasteiger *et al*., 2005), whereas average hydrophobicity and hydrophilicity (GRAVY), aliphatic and instability index were calculated with ExPASy ProtParam (Gasteiger *et al*., 2005). The secondary structure of FP-3α proteins was predicted with PSSPred (Yan *et al*., 2013). Then, the protein sequence was aligned to Protein Data Bank (PDB) (Berman *et al*., 2000) templates using Muster (Wu and Zhang, 2008) and the alignment was refined with CysBar (Shafee *et al*., 2016). Conserved cysteines were imposed as

distance constraints for I-Tasser modeling (Roy *et al.*, 2010). The cysteine connectivity was predicted from I-Tasser models and imposed before refinement based on molecular dynamics (MD) as previously described (Franzoi *et al*., 2017).

Results

Identification of fp-3α *transcripts and genes*

In order to investigate the distribution of fp-3α genes among bivalves, we used the only available fp-3α sequence record (*P. viridis*, AGZ84285.1) to search similar sequences in the different NCBI databases (nucleotide, EST, protein and TSA). As a result, we retrieved few hits, namely two ESTs from *M. galloprovincialis* and *M. edulis* (FL496987.1 and AM881265.1) and two TSA datasets (belonging to *P. viridis*). To expand the dataset, we assembled RNA-seq data of 20 byssogenic bivalve species belonging to *Mytilinae*, *Unionoida*, *Bathymodiolinae* and *Pteriidae* families. As a result, we retrieved 43 fp-3α transcripts belonging to *M. galloprovincialis* (7), *M. edulis* (11), *M. trossulus* (9), *M. californianus* (3), *M. coruscus* (3), *Trichomya hirsuta* (one incomplete sequence), *Mytilisepta virgata* (1) and *P. viridis* (8). Among the 14 bivalve genomes already sequenced, we could identify fp-3α genes only in *M. galloprovincialis*. In detail, we searched for fp-3α genes in the *M. galloprovincialis* genome translated into six reading frames by using HMMer. This allowed the identification of 12 fp-3α gene regions, of which 7 and 5 were annotated as coding and non-coding genes, respectively (Table 1).

In order to reconstruct exon-intron boundaries of the mussel fp-3a genes, we mapped all fp-3α transcripts on the *M. galloprovincialis* genome using a splice-aware aligner. As a result, we could map 20 out of 43 sequences, including hits of *M. galloprovincialis, M. edulis* and *M. trossulus*: most of the transcripts mapped to the MGAL10A009044 (12) and MGAL10ncA009345 fp-3α genes (3). The twelve fp-3α genic sequences identified in *M. galloprovincialis* generally displayed a small size (< 4 kb) and a structure based on three exons. In detail, the first exon contains the 5' UTR, the second one encodes for the remaining part of 5'-UTR and the main part of the signal peptide, whereas the third exon contains the rest of the coding sequence, including a short C-terminal region. However, some fp-3α genes displayed a single exon or an additional exon between those encoding for the signal peptide and the mature peptide (Figure 2A).

0	0	0	0	0	0	MGAL10A052660
0	0	0	4	81	0	MGAL10A002276
1	3	42	1	2	8	MGAL10ncA031470
5	27	0	20	1	1	MGAL10A031169
6	2	1	27	6	1	MGAL10A001023
0	15	0	0	0	1	MGAL10A048479
2	2	0	0	0	0	MGAL10ncA069861
2	2	0	0	0	0	MGAL10A001553
1	5	3	0	1	1	MGAL10ncA009347
1	37	0	0	0	1	MGAL10ncA009345
5	52	0	0	4	3	MGAL10A009044
0	8	0	0	0	0	MGAL10ncA073074
Inner mantle	Mantle edge	Hemolymph	Gill	Digestive Gland	Muscle	

Fig. 3 Tissue-specific expression levels in transcripts per million (TPM) for twelve genes encoding fp-3α proteins in *M. galloprovincialis*. The heatmap reports the average of the expression values per tissue

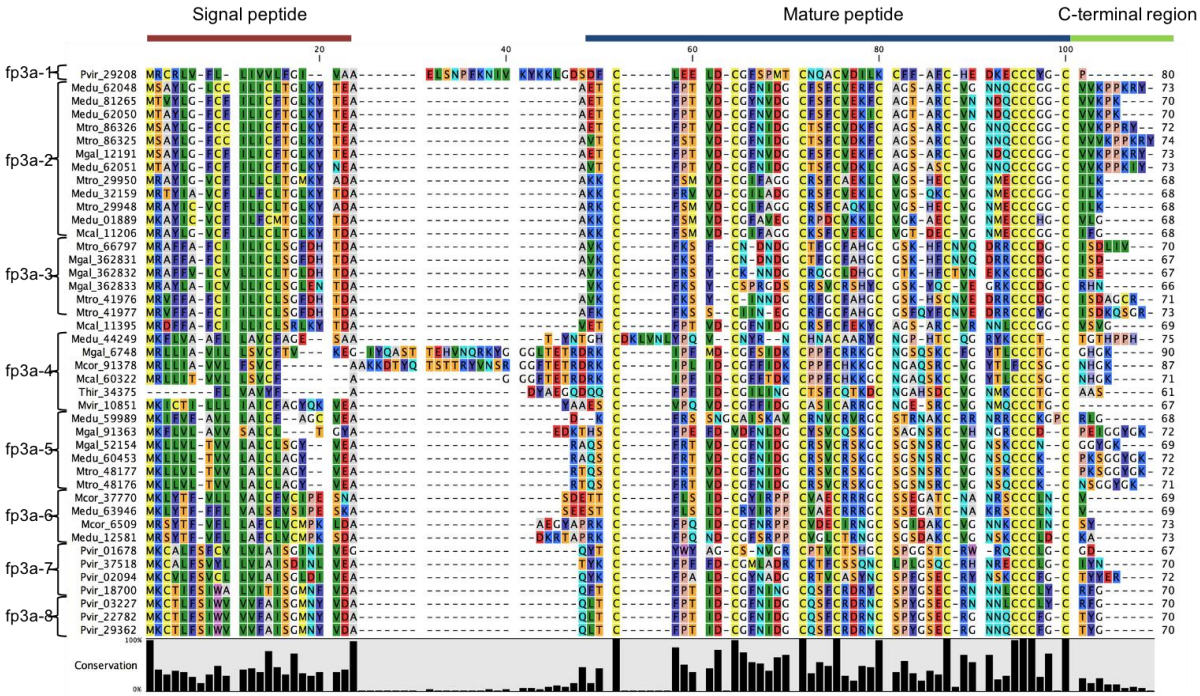

Fig. 4 Alignment of 43 predicted fp-3α proteins from *Mytilida* RNA-seq data. The segmented line over the multiple alignment indicates the signal peptide, mature peptide and the C-terminal region. The contig names are joined with a curly bracket per each fp-3a isoform. The conservation percentage at each amino acid position is also reported under the multiple alignment of the fp-3α protein sequences

Moreover, we identified two splicing isoforms for MGAL10A009044, with the first exon located slightly upstream in *M. edulis* and *M. trossulus*, compared to *M. galloprovincialis*, and thus resulting in a longer intron (519 bp compared to 392 bp). While *M. trossulus* and *M. galloprovincialis* possess only one MGAL10A009044 isoform, *M. edulis* has both the splicing isoforms (Figure 2B). *M. edulis* has two different splicing isoforms also for MGAL10ncA009345.

Expression levels of mussel fp-3α *genes based on RNA-seq data*

To investigate the expression patterns of the 12 *M. galloprovincialis* fp-3α genes, we used 30 carefully selected RNA-seq datasets generated from different tissues of naïve mussels, plus the RNA-seq data from two gill samples (128 milions reads) and from two hemolymph samples (74 million reads) produced for this study from *M. galloprovincialis*. Among the six fp-3a genes expressed in mantle, MGAL10A009044 showed the highest average expression, namely 52 Transcripts Per Kilobase Million (TPM) (range: 20-118 TPMs) whereas MGAL10A031169 was also expressed in gills with 27 and 20 TPM, respectively (Figure 3). Conversely, MGAL10A002276, MGAL10ncA031470 and MGAL10A001023 were selectively expressed in digestive gland, hemolymph and gills (average expression values: 81, 42 and 27 TPMs, respectively). The expression levels of the three remaining fp-3α genes was considered negligible (< 5 TPMs).

Inferences on mussel fp-3α proteins

The virtual translation of the fp-3α ORFs resulted in proteins of 58-90 aa residues, including a signal peptide (15-23 aa in length) and no recognizable InterPro or PFAM domains (Figure 4). The mature peptide region (length: 40-44 aa) included 10 highly conserved cysteines organized as follows: C(n)C(n)C(n)C(n)C(n)C(n)CCCxxC. Some of these mature peptide sequences displayed an N terminal extension long up to 25 aa. Even if very short, not more than 8 aa, some C-terminal region included peculiar motifs, like GGYGK or KPPKRY.

Fp-3α phylogenesis and classification

The Bayesian phylogenetic tree resulting from the multiple alignment clearly divided *P. viridis* fp-3α sequences from the others (upper branch in Figure 5). The remaining sequences split into two main clades. Clade 1 included two different subclades and one outlier *M. californianus* sequence, whereas Clade 2 grouped three different subclades: one subclade included sequences from different bivalve species (*Mytilus sp., Mytilisepta virgata, Trichomya hirsuta*) whereas the other two subclades included only sequences from *Mytilus* spp.

Based on the primary protein sequences and phylogenetic analysis, we could distinguish eight different fp-3α types in *Mytilida*: fp3α-1, -7, -8 typical of *P. viridis* and fp3α-2, -3, -4, -5, -6 occurring in other species (Table 1). Six of these fp-3α types are represented by one or more genes in the *M. galloprovincialis* genome (see Table 1). The most

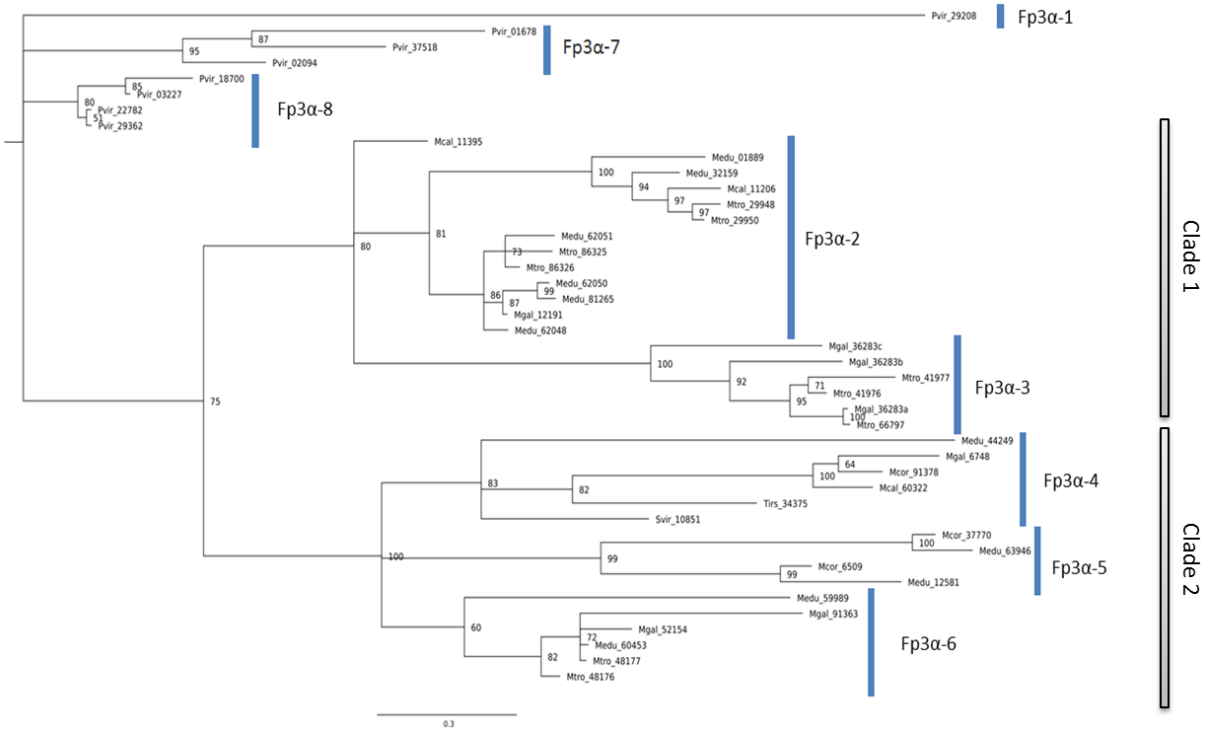

Fig. 5 Bayesian phylogenetic tree of 43 virtually translated fp-3α transcript sequences. The Bayesian phylogenesis was run for 3 million generations with a sampling frequency of 1,000 and a burn-in removal of 50% of the sampled trees. In the sequence names, the uppercase letter indicates the genus, the lower case letter indicates the species whereas numbers identify the contigs. The node labels report the bootstrap values as probability percentage. The scale bar at the bottom indicates the branch length representing 0.3 as amount of genetic change

abundant protein type in the above described Bayesian tree was fp3α-2 (12 sequences). Since *M. galloprovincialis* possesses two fp3α-1 genes, this protein type is not exclusive of *P. viridis* and could be reminiscent of the common *Mytilida* ancestor molecule.

Structural model of the M. galloprovincialis fp3α-2 protein

The 43 fp-3α proteins derived from the RNA-seq data are characterized by a low molecular weight, commonly between 4-6 KDa. The predicted isoelectric point varied from 4.17 to 10.2, whereas the GRAVY score, indicative of average hydrophobicity and hydrophilicity, ranged from -0.75 to +0.67. The stability index of these fp-3α proteins varied from -1.9 to 69.4.

The predicted secondary structure computed for Mgfp3α-2 (Mgal_12191) indicated the presence of three β-strands and one α-helix motif between the 1st and the 2nd β-strands. Within the cysteine array typically found in the fp-3α proteins, a stretch of three cysteine residues at the C-term of the mature peptide region is curiously located in one predicted β-strand.

In the absence of PDB protein homologs, we generated a fp3α-2 draft structure and then refined it by imposing the disulfide bonds. As regards the mature peptide, disulfides 9-31, 16-37 and 20-39 resulted to be highly conserved among the MUSTER hits (Figure 6A) and were imposed for model generation with I-Tasser. The resulting structures supported cysteine bonds also between residues 25-42 and 3-38, which were then imposed for the further refinement of the protein model. Hence, the final fp3α-2 model was characterized by three β-strands between aa 5-9, 30-33 and 37-40 (including an antiparallel β-strand) and a α-helix motif between aa 17-22 (Figure 6B). The quality of the fp3α-2 model was supported by Ramachandran plot (Lovell *et al*., 2003), since all the residues dihedrals were in the allowed regions. With a ProSA II score of -5.16, the final model fell in the range of native conformations (Wiederstein & Sippl, 2007). During the DM simulation, the C-terminal region showed a considerable displacement, indicative of an intrinsic flexibility and consistent with the PrDOS analysis (Ishida and Kinoshita, 2007; Richardson *et al*., 2009). DALI structural alignment revealed a high similarity with several different toxins and defensins (Table 2).

Discussion

The byssus production has been studied in both marine and freshwater mussels and more than 20 different proteins have been involved in this process (Lee *et al*., 2011). Crucially, the adhesive

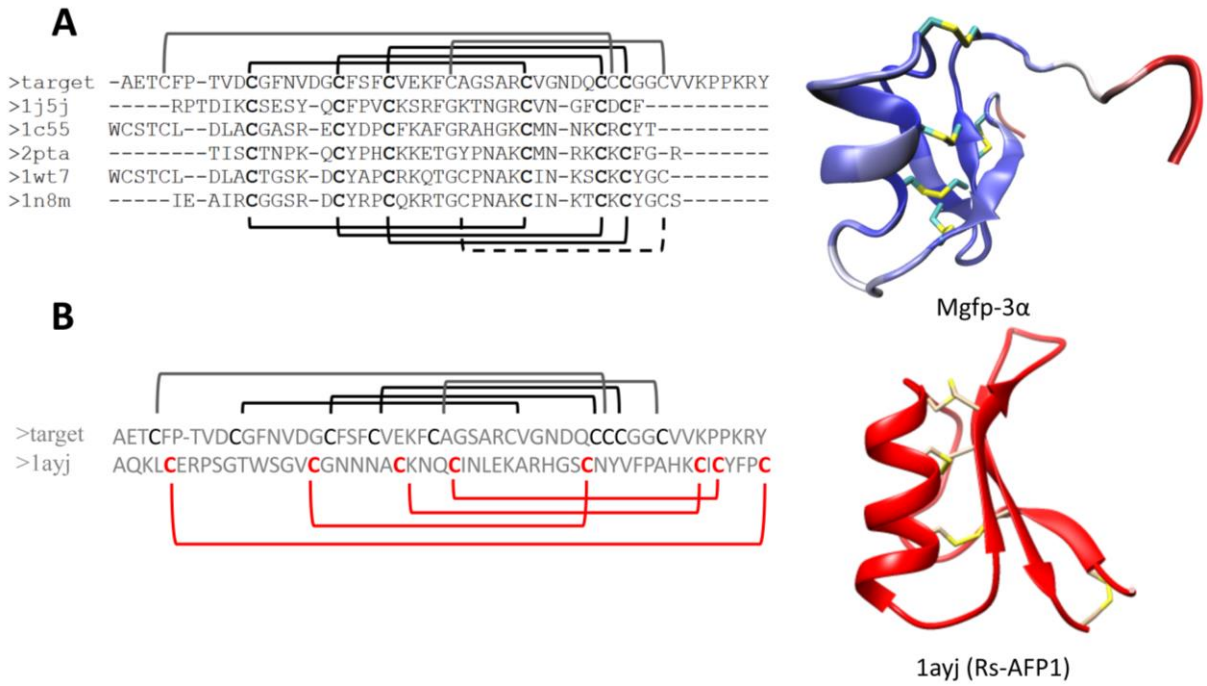

Fig. 6 Structural modelling of *M. galloprovincialis* foot protein 3α (Mgfp-3α) related to the plant defensin *Raphanus sativus* antifungal protein 1 (Rs-AFP1) which showed the highest Z score among other defensin proteins (see Table 2). A. Multiple alignment of Mgfp-3α with the best fitting hits from Protein Data Bank (PDB); The Mgfp-3α ribbon model is coloured according to the values of Root Mean Square Fluctuation index (RMSF) (blue and red indicate low and high RMSF, respectively). B. The Mgfp-3α and Rs-AFP1 alignment makes evident a different pattern of disulphide bonds, although the Rs-AFP1 and Mgfp-3α ribbon models are rather similar

plaque formation depends on a time-regulated secretion of different foot-associated proteins (Guerette *et al*., 2013) as the secretion of DOPA-containing proteins is immediately followed by that of other cysteine-rich proteins, which likely prevent the oxidation of DOPA residues through redox interactions. In *Mytilus spp.*, the cysteine-rich foot protein 6 (mfp6) has been considered largely responsible for the reducing activity within the terminal plaque (Yu *et al*., 2011; Waite, 2017) and a similar role has been outlined for the foot protein 3 alpha (fp-3α) of *P. viridis* (Petrone *et al*., 2015).

Searching into genomic and transcriptomic data of byssogenic bivalve species, we identified 43 fp-3α-like transcript sequences. In addition to *P. viridis,* we could trace these sequences in *M. virgata*, *T. hirsuta*, and *Mytilus sp. (Brachidotinae Septiferinae and Mytilinae* subfamilies, respectively*)*. The absence of fp-3α sequences in the subfamilies *Bathymodiolinae*, *Modiolinae*, and *Limnoperninae* would locate a primordial fp-3α gene after the division of the *Mytilidae* family in its two main clades (Lee *et al*., 2019). Notably, we could describe 12 fp-3α genes in the genome of *M. galloprovincialis,* which is the only available genome for a species having fp-3α genes. The tentative reconstruction of the evolutionary history of fp-3α genes suggests the duplication of an ancestral gene similar to fp3α-1 before the divergence of the Asian green mussels from the Mediterranean mussels.

The fp-3α genes exhibit a very diversified gene organization, based on 1, 3 or 4 exons. Such a situation could be explained by exon-shuffling events involving the exon encoding for the mature peptide (Froy and Gurevitz, 2003). Exon shuffling consists in an intron-mediated recombination of exons from existing genes *(*Long *et al*., 2003). To corroborate such an hypothesis for the fp-3α gene family, the mature peptide region has to be encoded by a single exon, with phase I flanking introns, and it should display folding authonomy (Patthy, 1999; Froy and Gurevitz, 2003). And this appears to be the case. Possibly favoured by exon shuffling, we also reported two different *M. edulis* splicing isoforms for the MGAL10A009044 and MGAL10ncA009345 genes. Three different fp-3α mussel genes possess 4 exons and 3 introns, with the third exon encoding for the N-terminal extension. Irrespective of the exon number, the nine remaining fp-3α mussel genes are devoid of the exon with the N-terminal extension, which could possibly have been lost via exon shuffling and perhaps not essential to the protein function. A similar phenomenon has been reported for *Crassostrea gigas* β-defensins, which likely originated from an ancestral big defensin by loss of

Table 2 Best hits from DALI Server for FP-3α. Protein Data Bank (PDB) code, name, class, organism of origin and function are reposted (Z, confidence index; rmsd, root mean square deviation between fp-3α and the structural homolog; lali, length of the alignment, % id, percentage of common residues)

PDB code	Name	Class	Organism	Function	Z	rmsd	lali	%id	Reference
2I61	LqhIT2	Toxin	*Leiurus quinquestriatus hebraeus*	Sodium channel activator	3.6	3.1	46	20	(Karbat *et al*., 2007)
1AGT	alpha-KTx 3.2	Toxin	*Leiurus quinquestriatus hebraeus*	Potassium channel inhibitor	3.3	1.9	36	19	(Krezel *et al*., 1995)
1FH3	Lqh3	Toxin	*Leiurus quinquestriatus hebraeus*	Sodium channel activator	3.2	3.3	45	18	(Krimm *et al*., 1999)
1NPI	Ts1	Toxin	*tityus serrulatus*	Sodium channel activator	3.2	3.7	46	20	(Pinheiro *et al*., 2003)
2UVS	alpha-KTx 3.1	Toxin	*Androctonus mauretanicus*	Potassium channel inhibitor	3.1	2.5	37	22	(Pentelute *et al*., 2009)
2K4U	alpha-KTx 3.6	Toxin	*Mesobuthus martensii*	Potassium channel inhibitor	3.1	2.5	36	19	(Yin *et al*., 2008)
1AYJ	Rs-afp1	Defensin	*Raphanus sativus*	Fungicide	3	2.5	40	23	(Fant *et al*., 1998)
4HE7	Brazzein	Defensin	*Pentadiplandra brazzeana*	Sweet-taste receptor binding peptide	3	3.4	40	23	(Nagata *et al*., 2013)
2LJ7	Lc-def	Defensin	*Lens culinaris*	Antimicrobial	2.9	2.3	40	23	(Shenkarev *et al*., 2014)
3PSM	Spe-10	Defensin	*Pachyrhizus erosus*	Antimicrobial	2.9	2.5	39	21	(Song *et al*., 2011)

the N-terminal domain via exon shuffling (Loth *et al*., 2019). Conversely, the fp-3α genes having a single-exon (MGAL10ncA069861, MGAL10ncA009347) could be explained by a retrotransposition mechanism, which basically consists in the reintegration of a mRNA back into genomic DNA (Navarro and Galante, 2015).

As regards the genomic location of mussel fp-3α genes, the *M. galloprovincialis genome* is assembled to scaffold level and, therefore, the distance between these genes can be established with certainty only if they fall in the same scaffold. Actually, only MGAL10A001023 and MGAL10A031169, both encoding for fp3α-6 type proteins and expressed at similar levels in the mussel gills, are in close vicinity in the same scaffold, thus supporting a recent duplication event.

From a functional point of view, only 9 of the 12 fp-3α genes resulted to be expressed at and variable, not negligible, level in the analysed transcriptomic datasets derived from multiple tissues and populations of *M. galloprovincialis*. Five of them were uniquely expressed in the mantle, while the remaining four were expressed in gills or digestive gland or hemolymph (Figure 3). Regarding the foot tissue, no transcriptome data were available for *M. galloprovincialis* but RNA-seq data specifically obtained from the three foot-associated glands of *M. californianus* (DeMartini *et al*., 2017) indicated an almost exclusive expression of MGAL10ncA069861, MGAL10A009044 and MGAL10ncA009347 in the accessory (cuticle) gland.

Mapping the bivalve transcript sequences (listed in Materials and methods) on the mussel genome, we could associate the fp3α-2 and fp3α-3 protein types (those composing clade 1 in the Bayesian tree) to the fp-3α genes expressed either in the accessory (cuticle) gland or in the mantle edge, tissues which contain mfp1, a DOPA-rich protein (Waite, 2017), and periostracin, another DOPA-rich protein (Waite *et al*., 1979), respectively. Reasonably, these two fp-3α protein types could have a DOPA-protective role, preventing its spontaneous oxidation and supporting DOPA-mediated cohesion processes (Nicklisch and Waite, 2012). Conversely, the fp3α-4, -5 and -6 (those composing clade 2 in the Bayesian tree) are mainly expressed in other mussel tissues. Curiously, these fp-3α protein types possess a signal peptide and a conserved CSαβ folding (including a Cys-x-Gly pattern in the γ-core) similarly to CSαβ antimicrobial peptides (Yeaman and Yount, 2007). With no doubt, their possible host defence role requires a dedicated investigation. If this was the case, it would corroborate the hypothesis of an ancestor mussel fp-3α gene evolved by gene duplication and functionally diversified through a neofunctionalization or a subfunctionalization mechanism (Innan and Kondrashov, 2010).

As reported in *P. viridis* and *Mytilus spp.*, native foot proteins start from a disordered state and assume elongated conformations still including disordered sequence tracts (Olivieri *et al*., 1997; Hwang and Waite, 2012; Mirshafian *et al*., 2016). Similar to defensins and scorpion toxins, the five disulphide bridges and the CSαβ motif of mussel fp-3α proteins (here modeled for fp3 α-2) are expected to improve the resistance to protease degradation and protein stability at non-physiological values of temperature and pH (Tam *et al.*, 2015; Koehbach, 2017). Moreover, the tyrosine residue present in the

flexible C-terminal of the mature peptide region in fp3α-2 transcripts (subterminal in some other fp-3α types) could be modified to DOPA, as demonstrated for the *P. viridis* homolog (Petrone *et al*., 2015). Overall, these findings suggest that at least the C-terminal of fp3α-2 proteins can be involved in binding activity, for instance with other proteins involved in the byssogenesis.

In conclusion, the comparison between 43 transcript-derived- and 12 genome-derived- fp-3α protein sequences demonstrated that *M. galloprovincialis* possesses at least one gene for each of the six fp-3α types detected in the *Mytilus* sp. complex, including fp3α-1 which is expressed in *P. viridis.* Likely expanded by several duplication events, the twelve members of the mussel fp-3α gene family show peculiar tissue expression patterns indicative of functional divergence at least between the two main gene clades recognized by Bayesian phylogenesis. While confirming the role of some fp-3α proteins in the maintenance of DOPA properties, the antimicrobial activity reported for canonical defensin might be demonstrated in the future at least for some of these genes. Actually, the defensin-like character of mussel fp-3α proteins is supported by the highly conserved cysteine residues, exon-intron boundaries and protein structure modeling.

Overall, the present study provides a first insight on the evolution of this highly peculiar mussel gene family and future studies are expected to expand our understanding on the evolution and functional roles of the fp-3α.

Acknowledgments

We are grateful to Marco Franzoi for helping us in the modelling of the fp3α structure. We thank Cristina Breggion and Andrea Sambo for their support in sampling *M. galloprovincialis* mussels.

Author contribution

EB and UR prepared the samples for sequencing, EB and UR performed bioinformatic analysis, EB and PV wrote the manuscript, AF e BN provided *M*. *galloprovincialis* genome, AF, BN and UR commented, contributed and approved the final version of the paper

Funding

This work was supported by the integrated budget for DIBIO-UNIPD research 2019 (BIRD 2019).

References

Babraham Bioinformatics - Trim Galore. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, 2019.

Bennett JJR, Sherratt JA. Large scale patterns in mussel beds: Stripes or spots? J. Math. Biol. 78: 815-835, 2019.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al*. The Protein Data Bank. Nucleic Acids Res. 28: 235-242, 2000.

Bielen A, Bošnjak I, Sepčić K, Jaklič M, Cvitanić M, Lušić J, *et al*. Differences in tolerance to anthropogenic stress between invasive and native bivalves. Sci. Total Environ, 543: 449-459, 2016.

Canapa A, Barucca M, Marinelli A, Olmo E. A Molecular Phylogeny of Heterodonta (Bivalvia) Based on Small Ribosomal Subunit RNA Sequences. Mol. Phylogenetics Evol. 21: 156-161, 2001.

Combosch DJ, Collins TM, Glover EA, Graf DL, Harper EM, Healy JM, *et al*. A family-level Tree of Life for bivalves based on a Sanger-sequencing approach. Mol. Phylogenetics Evol. 107: 191-208, 2017.

Darrigran G, Damborenea C. Ecosystem Engineering Impact of *Limnoperna fortunei* in South America. Zool. Sci., 28(1): 1-7, 2011.

De Blok JW, Tan-Maas M. Function of byssus threads in young postlarval *Mytilus*. Nature 267: 558-558, 1977.

DeMartini DG, Errico JM, Sjoestroem S, Fenster A, Waite JH. A cohort of new adhesive proteins identified from transcriptomic analysis of mussel foot glands. J. R. Soc. Interface 14: 20170151, 2017.

Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29: 1969-1973, 2012.

Du X, Fan G, Jiao Y, Zhang H, Guo X, Huang R, *et al*. The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. Gigascience 6: 1-12, 2017.

Eddy SR. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7: e1002195, 2011.

Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792-1797, 2004.

Fant F, Vranken W, Broekaert W, Borremans F. Determination of the three-dimensional solution structure of *Raphanus sativus* antifungal protein 1 by 1H NMR. J. Mol. Biol. 279: 257-270, 1998.

Franzoi M, Sturlese M, Bellanda M, Mammi S. A molecular dynamics strategy for CSαβ peptides disulfide-assisted model refinement. J. Biomol. Struct. Dyn. 35: 2736-2744, 2017.

Froy O, Gurevitz M. Arthropod and mollusk defensins – evolution by exon-shuffling. Trends Genet. 19. 684–687, 2003.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, *et al*. Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Walker (Ed.), The Proteomics Protocols Handbook (pp. 571–607). Humana Press, 2005.

Gómez-Chiarri M, Warren WC, Guo X, Proestou D. Developing tools for the study of molluscan immunity: The sequencing of the genome of the eastern oyster, *Crassostrea virginica*. Fish Shellfish Immunol. 46: 2-4, 2015.

González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, *et al*. A phylogenetic backbone for Bivalvia: An RNA-seq approach. P. Roy. Soc. B-Biol. Sci. 282: 20142332, 2015.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29: 644-652, 2011.

Guerette PA, Hoon S, Seow Y, Raida M, Masic A, Wong FT, *et al*. Accelerating the design of biomimetic materials by integrating RNA-seq with proteomics and materials science. Nat. Biotechnol. 31: 908-915, 2013.

Harrington MJ, Jehle F, Priemel T. Mussel Byssus Structure-Function and Fabrication as Inspiration for Biotechnological Production of Advanced Materials. Biotechnol. J. 13: 1800133, 2018.

Harrington MJ, Masic A, Holten-Andersen N, Waite JH, Fratzl P. Iron-Clad Fibers: A Metal-Based Biological Strategy for Hard Flexible Coatings. Science 328: 216-220, 2010.

Herbert Waite J, Saleuddin ASM, Andersen SO. Periostracin - A soluble precursor of sclerotized periostracum in *Mytilus edulis* L. J. Comp. Physiol. 130: 301-307, 1979.

Hwang DS, Waite JH. Three intrinsically unstructured mussel adhesive proteins, mfp-1, mfp-2, and mfp-3: Analysis by circular dichroism. Proteins 21: 1689-1695, 2012.

Innan H, Kondrashov F. The evolution of gene duplications: Classifying and distinguishing between models. Nat. Rev. Genet. 11: 97-108, 2010.

Ishida T, Kinoshita K. PrDOS: Prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res. 35 (suppl 2): W460-W464, 2007.

Karatayev AY, Burlakova LE, Mastitsky SE, Padilla DK. Predicting the spread of aquatic invaders: Insight from 200 years of invasion by zebra mussels. Ecol. Appl. 25: 430-440, 2015.

Karbat I, Turkov M, Cohen L, Kahn R, Gordon D, Gurevitz M, *et al*. X-ray Structure and Mutagenesis of the Scorpion Depressant Toxin LqhIT2 Reveals Key Determinants Crucial for Activity and Anti-Insect Selectivity. J. Mol. Biol. 366: 586-601, 2007.

Kenny NJ, McCarthy SA, Dudchenko O, James K, Betteridge E, Corton C, *et al*. The gene-rich genome of the scallop *Pecten maximus*. Gigascience 9: https://doi.org/10.1093/gigascience/giaa037, 2020.

Koehbach J. Structure-Activity Relationships of Insect Defensins. Front. Chem. 5: https://doi.org/10.3389/fchem.2017.00045, 2017.

Krezel AM, Kasibhatla C, Hidalgo P, MacKinnon R, Wagner G. Solution structure of the potassium channel inhibitor agitoxin 2: Caliper for probing channel geometry. Proteins 4: 1478-1489, 1995.

Krimm I, Gilles N, Sautière P, Stankiewicz M, Pelhate M, Gordon D, *et al*. NMR structures and activity of a novel alpha-like toxin from the scorpion *Leiurus quinquestriatus hebraeus*. J. Mol. Biol. 285: 1749-1763, 1999.

Lee BP, Messersmith PB, Israelachvili JN, Waite JH. Mussel-Inspired Adhesives and Coatings. Annu. Rev. Mater. Res. 41: 99-132, 2011.

Lee H, Lee BP, Messersmith PB. A reversible wet/dry adhesive inspired by mussels and geckos. Nature 448: 338-341, 2007.

Lee Y, Kwak H, Shin J, Kim S-C, Kim T, Park J-K. A mitochondrial genome phylogeny of Mytilidae (Bivalvia: Mytilida). Mol. Phylogenetics Evol. 139: 106533, 2019.

Liu X, Li C, Chen M, Liu B, Yan X, Ning J, *et al*. Draft genomes of two Atlantic bay scallop subspecies *Argopecten irradians irradians* and *A. i. concentricus*. Sci. Data 7: 1-8, 2020.

Long M, Betrán E, Thornton K, Wang W. The origin of new genes: Glimpses from the young and old. Nat. Rev. Genet. 4: 865-875, 2003.

Loth K, Vergnes A, Barreto C, Voisin SN, Meudal H, Da Silva J, *et al*. The Ancestral N-Terminal Domain of Big Defensins Drives Bacterially Triggered Assembly into Antimicrobial Nanonets. MBio 10: e01821-19, 2019.

Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, *et al*. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. Proteins 50: 437-450, 2003.

McCartney MA, Auch B, Kono T, Mallez S, Zhang Y, Obille A, *et al*. The Genome of the Zebra Mussel, *Dreissena polymorpha*: A Resource for Invasive Species Research. BioRxiv 696732, 2019.

Mirshafian R, Wei W, Israelachvili JN, Waite JH. α,β-Dehydro-Dopa: A Hidden Participant in Mussel Adhesion. Biochemistry 55: 743-750, 2016.

Nagata K, Hongo N, Kameda Y, Yamamura A, Sasaki H, Lee WC, *et al*. The structure of brazzein, a sweet-tasting protein from the wild African plant *Pentadiplandra brazzeana*. Acta Cryst. D, Biol. Crystallogr. 69: 642-647, 2013.

Navarro FCP, Galante PAF. A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biol. Evol. 7: 2265-2275, 2015.

Nicklisch SCT, Waite JH. Mini-review: The role of redox in Dopa-mediated marine adhesion. Biofouling 28: 865-877, 2012.

Olivieri MP, Wollman RM, Alderfer JL. Nuclear magnetic resonance spectroscopy of mussel adhesive protein repeating peptide segment. J. Pept. Res. 50: 436-442, 1997.

Patthy L. Genome evolution and the evolution of exon-shuffling- A review. Gene 238: 103-114, 1999.

Pentelute BL, Mandal K, Gates ZP, Sawaya MR, Yeates TO, Kent SBH. Total chemical synthesis and X-ray structure of kaliotoxin by racemic protein crystallography. Chem. Comm. 46: 8174-8176, 2009.

Petrone L, Kumar A, Sutanto CN, Patil NJ, Kannan S, Palaniappan A, *et al*. Mussel adhesion is dictated by time-regulated secretion and molecular conformation of mussel adhesive proteins. Nature Comm. 6: 8737, 2015.

Pinheiro CB, Marangoni S, Toyama MH, Polikarpov I. Structural analysis of *Tityus serrulatus* Ts1 neurotoxin at atomic resolution: Insights into interactions with Na+ channels. Acta Cryst. D, Biol. Crystallogr. 59: 405-415, 2003.

Ponder WF, Lindberg DR, Ponder JM. Biology and Evolution of the Mollusca, Volume 2. CRC Press, Boca Raton, Florida 2020 pp 125-127, 2020.

Posada D, Crandall KA. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14: 817-818, 1998.

Powell D, Subramanian S, Suwansa-Ard S, Zhao M, O'Connor W, Raftos D, *et al*. The genome of the oyster *Saccostrea* offers insight into the environmental resilience of bivalves. DNA Res. 25: 655-665, 2018.

Priemel T, Degtyar E, Dean MN, Harrington MJ. Rapid self-assembly of complex biomolecular architectures during mussel byssus biofabrication. Nature Comm. 8: 1-12, 2017.

Ran Z, Li Z, Yan X, Liao K, Kong F, Zhang L, *et al*. Chromosome-level genome assembly of the razor clam *Sinonovacula constricta* (Lamarck, 1818). Mol. Ecol. Resour. 19: 1647-1658, 2019.

Rambaut A. Fig Tree. http://tree.bio.ed.ac.uk/software/figtree/, 2012.

Renaut S, Guerra D, Hoeh WR, Stewart DT, Bogan AE, Ghiselli F, *et al*. Genome Survey of the Freshwater Mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida) Using a Hybrid De Novo Assembly Approach. Genome Biol. Evol. 10: 1637-1646 2018.

Richardson JM, Colloms SD, Finnegan DJ, Walkinshaw MD. Molecular architecture of the Mos1 paired-end complex: The structural basis of DNA transposition in a eukaryote. Cell 138: 1096-1108, 2009.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, *et al*. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol., 61: 539-542, 2012.

Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. Nat. Protoc. 5: 725-738, 2010.

Sagert J, Waite JH. Hyperunstable matrix proteins in the byssus of *Mytilus galloprovincialis*. J. Exp. Biol., 212: 2224-2236, 2009.

Shafee TMA, Robinson AJ, Weerden N, Anderson MA. Structural homology guided alignment of cysteine rich proteins. SpringerPlus 5: 1-7, 2016.

Shenkarev ZO, Gizatullina AK, Finkina EI, Alekseeva EA, Balandin SV, Mineev KS, *et al*. Heterologous expression and solution structure of defensin from lentil *Lens culinaris*. Biochem. Biophys. Res. Commun. 451: 252-257, 2014.

Sigurdsson JB, Titman CW, Davies PA. The dispersal of young post-larval bivalve molluscs by byssus threads. Nature 262: 386-387, 1976.

Song X, Zhang M, Zhou Z, Gong W. Ultra-high resolution crystal structure of a dimeric defensin SPE10. FEBS Lett. 585: 300-306, 2011.

Staff SRAS. Using the SRA Toolkit to convert .sra files into other formats. In: SRA Knowledge Base [Internet]. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK158900/, 2011.

Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, *et al*. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat. Ecol. Evol. 1: 1-7, 2017.

Tam JP, Wang S, Wong KH, Tan WL. Antimicrobial Peptides from Plants. Pharmaceuticals 8: 711-757, 2015.

Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, Americo JA, *et al*. A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, *Limnoperna fortunei*. Gigascience 7: https://doi.org/10.1093/gigascience/gix128, 2018.

Wagner GP, Kin K, Lynch VJ. A model based criterion for gene expression calls using RNA-seq data. Theory Biosci. 132: 159-164, 2013.

Waite JH. The formation of mussel byssus: Anatomy of a natural manifacturing process. In: Case ST (ed.) Structure, Cellular Syntesis and Assembly of Biopolimers, Springer-Veriag, Berlin, Germany, pp 27-54, 1992.

Waite JH. Mussel adhesion – essential footwork. J. Exp. Biol. 220: 517-530, 2017.

Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, *et al*. Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat. Ecol. Evol. 1: 1-12, 2017.

Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35: W407-W410, 2007.

Wijsman JWM, Troost K, Fang J, Roncarati A. Global Production of Marine Bivalves. Trends and Challenges. In: Smaal AC, Ferreira JG, Grant J, Petersen JK, Strand Ø (eds.), Goods and Services of Marine Bivalves, Springer International Publishing, pp. 7–26, 2019.

WoRMS - World Register of Marine Species. http://www.marinespecies.org/aphia.php?p=stats, 2020.

Wu S, Zhang Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72: 547-556, 2008.

Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci. Rep. 3: 1-9, 2013.

Yan X, Nie H, Huo Z, Ding J, Li Z, Yan L, *et al*. Clam Genome Sequence Clarifies the Molecular Basis of Its Benthic Adaptation and Extraordinary Shell Color Diversity. IScience 19: 1225-1237, 2019.

Yeaman MR, Yount NY. Unifying themes in host defence effector polypeptides. Nat. Rev. Microbiol. 5: 727-740, 2007.

Yin S-J, Jiang L, Yi H, Han S, Yang D-W, Liu M-L, *et al*. Different residues in channel turret determining the selectivity of ADWX-1 inhibitor peptide between Kv1.1 and Kv1.3 channels. J. Proteome Res. 7: 4890-4897, 2008.

Yu J, Wei W, Danner E, Ashley RK, Israelachvili JN, Waite JH. Mussel protein adhesion depends on interprotein thiol-mediated redox modulation. Nat. Chem. Biol. 7: 588-590, 2011.

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 490: 49-54, 2012.

Zhong C, Gurry T, Cheng AA, Downey J, Deng Z, Stultz CM, *et al*. Strong underwater adhesives made by self-assembling multi-protein nanofibres. Nat. Nanotechnol. 9: 858-866, 2014.