Aquaculture Research

Aquaculture Research, 2013, 1–13 doi: 10.1111/are.12273

SNP discovery and gene annotation in the surf clam *Mesodesma donacium*

Cristian Gallardo-Escárate[1,2], Valentina Valenzuela-Muñoz[1,2], Gustavo Núñez-Acuña[1,2] & Pilar Haye[2,3]

[1]Laboratory of Biotechnology and Aquatic Genomics, University of Concepción, Concepción, Chile

[2]Interdisciplinary Center for Aquaculture Research (INCAR), University of Concepción, Concepción, Chile

[3]Laboratorio de Diversidad Molecular, Departamento de Biología Marina, Facultad de Ciencias del Mar, Universidad Católica del Norte, Coquimbo, Chile

Correspondence: Dr. C Gallardo-Escárate, Universidad de Concepción, P.O. Box 160-C, Concepción, Chile.
E-mail: crisgallardo@udec.cl

Abstract

The main objective of this research was to identify single-nucleotide polymorphisms (SNPs) from an Expressed Sequences Tags (EST) data set generated by 454 pyrosequencing in the soft clam *Mesodesma donacium*. A total of 180 159 ESTs were yielded from a *M. donacium* cDNA library. *De novo* assembly was performed using stringent calling parameters, producing 10 178 contigs and 41 765 singletons. Here, a total of 2594 SNPs were discovered related to 613 consensus sequences, achieving a frequency of 1 SNPs per 260 bp. SNP variants showed that A/G, A/T and C/T were the most abundant among the identified polymorphisms. We validated a total of 12 SNPs loci by HRMA for annotated genes such as heat shock protein-70 and the translation elongation factor 1-alpha. The Gene Ontology analysis regarding molecular function level revealed that sequences with SNPs were mainly classified to protein and nucleotide binding, as well hydrolase activity, ion binding and oxidoreductase activity. Further, biological processes like cellular and metabolic process, biogenesis, localization and biological regulation were highly annotated. The most expressed genes were related to the mitochondrial electron transport chain, senescence-associated protein, ubiquitin and actin. Interestingly, some relevant genes related to immune response and biomineralization showed a high abundance, such as tumor necrosis factor (TNF)-alpha-receptor-like protein, serine protease inhibitor, heat shock protein, aragonite-binding protein and ferritin. This study contributes to relevant genes associated with functional polymorphisms and gives an overview for future genetic investigations.

Keywords: *Mesodesma donacium*, pyrosequencing, transcriptome, SNP

Introduction

Generating EST collections is a popular approach for initiating genomic research of model and non-model species because it provides information about the genome that is actually expressed (Wilhelm & Landry 2009; Bhasker & Hardiman 2010; Cerda, Douglas & Reith 2010; Venier, Varotto, Rosani, Millino, Celegato, Bernante, Lanfranchi, Novoa, Roch, Figueras & Pallavicini 2011). EST data have been reported for a few commercially important bivalves, such as oysters, mussels and scallops. Until November 2012, the NCBI EST database browser retrieved 369 093 entries for Bivalvia. Here, 221 293 for Ostreidae, 71 328 for Mytilidae, 13 124 for Veneridae and 21 101 for Pectinidae have been mainly reported. All other taxa for Bivalvia comprise a total of 42 247 ESTs. These collections have provided a substantial number of sequences that are similar to genes already known, but a large proportion of ESTs show no hit sequences after database searches (Saavedra & Bachère 2006; Tanguy, Bierne, Saavedra, Pina, Bachere, Kube, Bazin, Bonhomme, Boudry, Boulo, Boutet, Cancela, Dossat, Favrel, Huvet, Jarque, Jollivet, Klages, Lapegue, Leite, Moal, Moraga, Reinhardt, Samain, Zouros & Canario 2008). Moreover, ESTs can serve as a source of molecular DNA

markers and provide novel insights for studies in molecular ecology, evolution and biotechnology (Bouck & Vision 2007; Wheat 2010; Johansen, Karlsen, Furmanek, Andreassen, Jorgensen, Bizuayehu, Breines, Emblem, Kettunen, Luukko, Edvardsen, Nordeide, Coucheron & Moum 2011; Rice, Rudh, Ellegren & Qvarnstrom 2011).

Without doubt the major challenge to the transcriptomics of nonmodel organisms is to increase rapidness and accuracy in the search for new genes and metabolic pathways, as well as determining how gene transcription variations are regulated by specific DNA polymorphisms. Next-generation sequencing (NGS) offers the opportunity to generate genome-wide sequence data sets of nonmodel organisms at a reasonable cost (Hudson 2008; Vera, Wheat, Fescemyer, Frilander & Crawford 2008; Marguerat & Bahler 2010). Although these powerful and rapidly evolving technologies have only been available for a few years, they are already making substantial contributions to our understanding of genome expression and regulation. A popular target for NGS is the generation of species transcriptome, which offers direct access to the coding sequences of many genes and information on their relative expression levels (Bellin, Ferrarini, Chimento, Kaiser, Levenkova, Bouffard & Delledonne 2009; Wilhelm & Landry 2009; Neira-Oviedo, Tsyganov-Bodounov, Lycett, Kokoza, Raikhel & Krzywinski 2011). Next-generation sequencing transcriptome data analysis is therefore a promising source of genetic markers (e.g. SNPs and EST-SSRs) that are potentially applicable to closely related taxa and eventually to any taxonomic group (Cardenas, Sanchez, Gomez, Fuenzalida, Gallardo-Escarate & Tanguy 2011; Everett, Grau & Seeb 2011; Seeb, Carvalho, Hauser, Naish, Roberts & Seeb 2011; Aguilar-Espinoza, Guzmán-Riffo, Haye & Gallardo-Escárate 2012).

The surf clam, *Mesodesma donacium*, is an endemic sand-dwelling marine bivalve, found from 5°S to 43°S along the south-eastern Pacific Coast. It plays an important socio-economic role in small-scale fishing in Chile, where 4056 tons were extracted in 2010. The species has not been evaluated to assign a conservation category even though it is highly exploited and some local populations have disappeared. Despite its importance and precariousness, there have been few genetic studies of *M. donacium* (Marins & Levy 1999; Amar, Rojas, von Brand & Jara-Seguel 2008). The latest was based solely on the mitochondrial marker cytochrome oxidase I (COI) to estimate population genetic structure (Peralta 2008). The study showed a very sharp discontinuity around the southern border (ca. 33–34S) of a known biogeographical transition zone of the Chilean coast. A recent study by our research group using simple sequence repeats found a high level of genetic differentiation among surf clam populations (Aguilar-Espinoza *et al.* 2012), which was consistent with the results of the study employing the COI marker. However, functional polymorphisms and candidate gene expression have never been analysed for *M. donacium* because of the lack of specific markers. Understanding variation in neutral and selectively affected loci is pivotal to obtaining an in-depth understanding of local putative adaptations that can ultimately serve to prevent loss of genomic biodiversity in *M. donacium* along the Chilean coast. In this study, through 454 pyrosequencing, we conducted extensive RNA sequencing of *M. donacium* to provide an overview of relevant functional genes associated with SNPs.

Material and methods

RNA isolation and pyrosequencing

Fifteen adult individuals of *M. donacium* were collected from the natural bank in Caleta San Pedro, Coquimbo, Chile (29°54′S–71°13′W), in February 2011. Immediately after sampled, 100 mg of gill and muscle tissue was extracted from the clams and conserved in 1 mL of RNAlater stabilization solution (Ambion®, Life Technologies, Carlsbad, CA, USA) and stored at −80°C until total RNA extraction. In preliminary analysis, these tissues provided good quality RNA and hence they are the choice for the SNP discovery through transcriptomic analysis. Total RNA was extracted from 100 mg of gills and muscle from each individual using Trizol reagent (Invitrogen, Life Technologies) in accordance with the supplier's instructions. Total RNA of gills and muscle was pooled by tissues, and the concentration and purity were measured with a ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE, USA) and the integrity was visualized with electrophoresis in MOPS/formaldehyde agarose gels at 1.2% staining with ethidium bromide at 0.001%. RNA was also checked for quality on a 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). RNA extracts that presented 260/280 and

260/230 purity indices equal to or greater than 2.0, and with integral RNA in electrophoresis and bioanalyser measurements (RIN > 8), were selected and pooled. Subsequently, mRNA pools were precipitated overnight with 2× volume of absolute ethanol and 0.1× volume of sodium acetate 0.3 M at −80°C. Double-stranded cDNA was synthesized from the extracted RNA, and 1/8 plate was pyrosequenced in the 454 GS FLX titanium platform in Macrogen Inc. (Seoul, Korea).

Assembly and annotation

Sequence assembly was carried out using CLC Genomics Workbench software version 5.5.1 (CLC Bio, Aarhus, Denmark). Here, *de novo* assembly was applied with overlap criteria of 70% and similarity of 0.9 to exclude paralogous sequence variants (PSVs) (Renaut *et al.* 2010). Used settings were mismatch cost = 2, deletion cost = 3, insert cost = 3, minimum contig length = 200 bp and trimming quality score = 0.05. After the assembly process, singletons were retained in the data set because they could be fragments of low-expression transcripts. However, their sequence redundancy was removed using the application Duplicate Finder incorporated in Geneious version 5.1.7 software (Biomatters, Auckland, New Zealand).

Candidate SNPs were also identified using CLC software from *de novo* assembly. The parameters were window length = 11, maximum gap and mismatch count = 2, minimum average quality of surrounding bases = 15, minimum quality of central base = 20, maximum coverage = 100, minimum coverage = 8, minimum variant frequency (%) = 35.0, maximum expected variations (ploidy) = 2. In addition, 454-homopolymer indels filter was applied.

To gauge the number of transcripts and gene function, BLASTx was carried out on consensus sequences annotated with SNPs. Consensus sequences with a SNP variant from *M. donacium* transcriptome were annotated to the UniProtKB/Swiss-Prot database (http://uniprot.org) to determine putative gene descriptions, and also, associated Gene Ontology (GO) terms were acceded by Blast2Go software (Conesa, Gotz, Garcia-Gomez, Terol, Talon & Robles 2005) with a cut-off E-value of 1E-05. Moreover, to evaluate the transcription expression of sequences annotated with SNP variants, we performed an RNA-Seq analysis, where the ESTs generated by pyrosequencing were mapped against the polymorphic consensus contigs. The RNA-seq settings were minimum length fraction = 0.6 and minimum similarity fraction (long reads) = 0.5. The expression value was set in reads per kilobase of exon model value (RPKM).

SNP validation by high-resolution melting analysis

To further characterize a subset of SNPs identified from sequences analysis, 40 contigs with putative SNPs were selected. Here, SNP variants were analysed by high-resolution melting analysis (HRMA) in 10 individuals of soft clam *M. donacium*. High-resolution melting analysis primers were designed using Primer3 included in Geneious Pro 5.1.7 software (Biomatters, New Zealand). The PCR was carried out in 10 μL reaction with 13 ng template DNA using Fast EvaGreen® qPCR Master Mix (Biotum, Hayward, CA, USA). For HRMA, thermal cycling was performed with an ECO Real-Time PCR System (Ilumina Inc, San Diego, CA, USA) as follows: 2 min for enzyme activation, 40 cycles: 95°C for 5 s, 56°C for 5 s, 60°C for 25 s. High-resolution melting analysis data were collected between 60 and 95°C with a temperature interval of 0.3%. The genotyping was analysed for the presence of discrete melting curve using the software Eco Real-Time System (Illumina Inc).

Results

De novo assembly

We obtained a total of 181 159 sequences, ranging from 41.3 to 758.4 bp and the quantity of 65.8 Mb. The mean length sequenced was 581 bp and the CG content 34.63%. All sequence data generated from the cDNA library of *M. donacium* are available for download at the Dryad Digital Depository (http://datadryad.org/) under the access http://dx.doi.org/10.5061/dryad.8jd18.

After quality trimming and excluding very short (length <35 bp) and poor quality reads, a comprehensive ESTs data set was assembled using *de novo* assembly. The assembly of reads resulted in 10 178 contigs and 41 765 singletons, with average lengths of 581 and 361 bp, respectively (Table 1). Seventy-seven percentage of the reads were assembled using the *de novo* approach. There was an average of 95 assembled reads per consensus sequence with a mean coverage of 14×. Figure 1 shows the contigs length sequence distribution from *de novo* assembly and the GC

contents in the *M. donacium* EST database. *De novo* assembly yielded a higher frequency of longer contigs (>500 bp) with a 35% of CG content.

SNP discovery and gene annotation

The calling SNP criteria proposed for this study identified a total of 2594 putative SNPs in 613 contigs. The average consensus sequence length was 1013 bp, showing an SNP frequency of 1/260 bp estimated for sequences that were annotated with SNPs variants. Further, the frequency calculated for all data set was 1/2279 (Table 1). The putative SNPs (53%) were mainly identified as transitions (A/G or C/T), while A/C, G/T and C/G transversions were found in less than 12%, except for the A/T transversions substitution, reached to 21% of the detected SNPs (Fig. 2).

Table 1 Summary of pyrosequencing from cDNA of surf clam *Mesodesma donacium*

	***De novo* assembly**
ESTs	
Reads	180 159
Average length (bp)	365
Matched	138 394 (77%)
Contigs	10 178
Average length (bp)	581
Singletons	41 765
Average length (bp)	361
Number nucleotides (Mb)	65.8
SNPs	
Number of SNPs	2594
Contigs with SNPs	613
Average sequence length with SNPs (bp)	1013
Frequency in contigs with SNPs (SNP/bp)	1/260
Frequency in all data set (SNP/bp)	1/2279

According to GO terms, 11% of the sequences resulting *de novo* assembly are associated with biological processes, 9% with cellular components and 13% with molecular functions. Sixty-seven percentage of the sequences showed no significant hits to the UniProtKB/Swiss-Prot database. However, BLAST analysis carried out on contigs annotated with SNP variants, revealed that from 613 sequences, 533 showed significant hits to known transcripts (see Data S1). Here, we performed a GO analysis considering the 613 sequences annotated with SNPs variants. The GO analysis regarding molecular function level revealed that sequences with SNPs were mainly classified to protein and nucleotide binding, as well hydrolase activity, ion binding and oxidoreductase activity. Further, biological processes like cellular and metabolic process, biogenesis, localization and

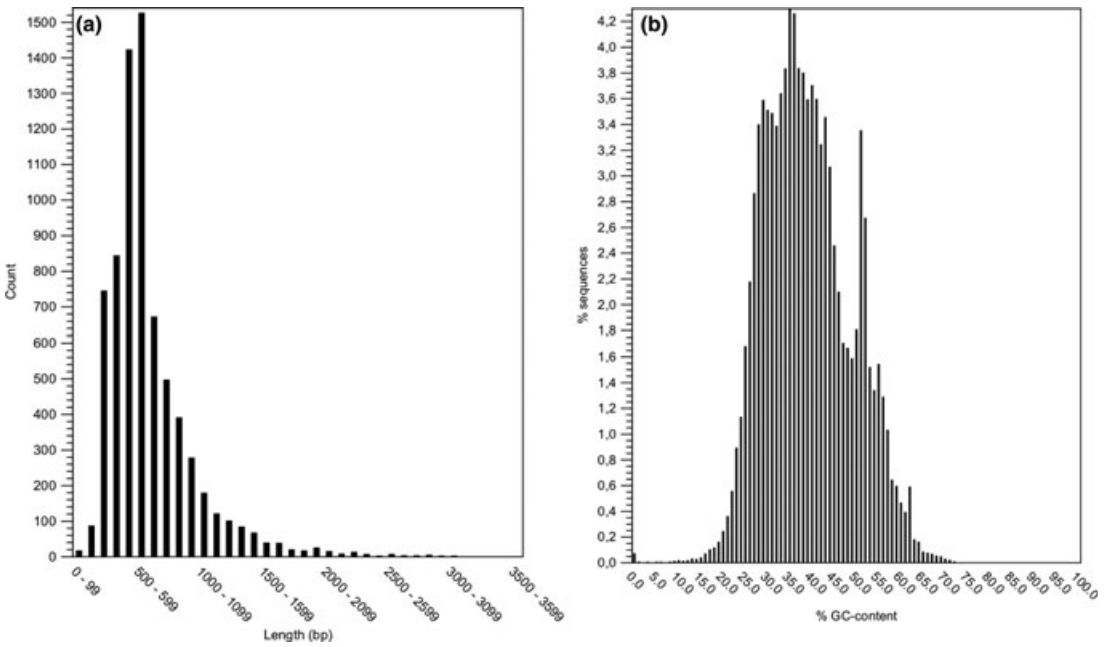

Figure 1 Contig length distribution from *de novo* assembly (a) and GC contents (b) in the *Mesodesma donacium* EST-database.

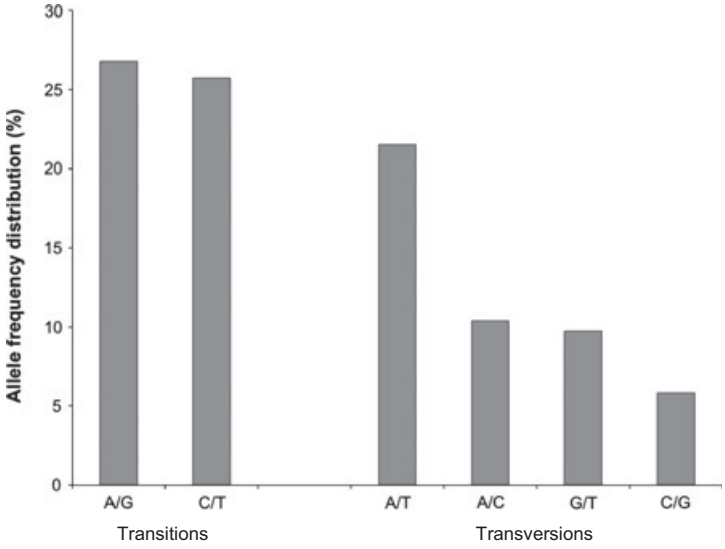

Figure 2 Distribution of transitions and transversions of single-nucleotide polymorphism alleles from *Mesodesma donacium* transcriptome.

biological regulation were highly annotated. Additional process involved in developmental process, response to stimulus, immune system, growth and biological adhesion was less achieved. With respect to cellular component level, intracellular, intracellular parts and intracellular organelle part were mainly identified (Fig. 3). Details of the relevant genes are presented in Table 2. For instance, genes related to immune response, such as putative C1q, mytimacin-5 and lysozyme; response to stimulus, such as heat shock protein 70 and 90, cathepsin L-like cysteine protease 2 and cAMP-responsive element binding protein; and biogenesis such as 60S ribosomal protein L11 and putative ribosome biogenesis protein RLP24 were associated with DNA polymorphisms. Other pivotal biological processes in sequences with putative SNPs were also classified in reproduction, nutrition, growth, glycolysis, homeostasis, regulation of transcription and proteolysis (see Data S2 for more details).

Gene transcription of contigs annotated with SNP variants

Gene expression analysis was performed using RNA-seq data extracted from the number of reads that map uniquely to each contigs or gene from the corresponding alignment. These count data will serve as a proxy for the magnitude of gene expression of transcripts that were annotated with putative SNPs from *M. donacium* transcriptome. Here, the Figure 4 shows the top-hit expression of these contigs, revealing that the most expressed transcripts are related to the mitochondrial electron transport chain such as NADH dehydrogenase subunit 4, cytochrome c oxidase subunit I, ATP synthase F0 subunit 6 and cytochrome b. Further genes were also highly expressed such as senescence-associated protein, ubiquitin and actin. Interestingly, some relevant genes related to immune response and biomineralization showed a high abundance, such as TNF-alpha-receptor-like protein, serine protease inhibitor, heat shock protein 70, 90, aragonite-binding protein and ferritin.

SNP validation

A total of 2594 putative SNPs were identified from 613 contigs from *M. donacium* transcriptome. To validate the putative SNPs, 40 primer pairs were designed and synthesized for their evaluation by PCR. As shown in Figure 5, the polymorphisms were assessed by HRMA. After the standardization procedure, a total of 12 SNP loci evidenced unique PCR products, expected size and reproducibility (see Data S3). The SNPs panel was annotated to heat shock protein 70, translation elongation factor 1-alpha, serine/threonine protein kinase, among others.

Discussion

In marine invertebrates and especially in bivalves, transcriptomes studies have been carried out to gain insights into relevant biological process such as innate immune response, reproduction, growth and

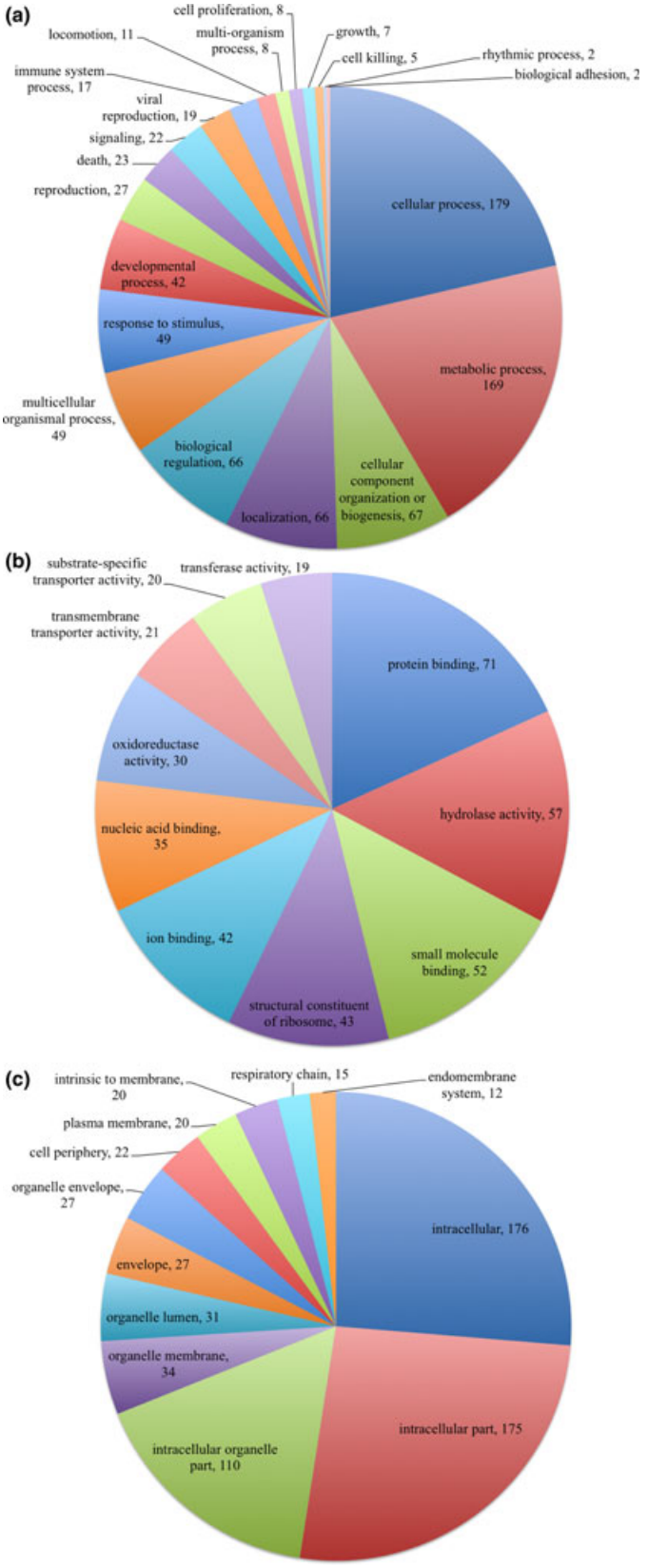

Figure 3 Gene Ontology analysis performed in contigs sequences annotated with single-nucleotide polymorphism variants in *Mesodesma donacium* transcriptome: biological process (a), molecular function (b) and cellular component (c).

Table 2 Relevant genes summary annotated with SNP variants and classification according to GO terms

Seq. Name	Seq. Description	Seq. Length	Min. eValue
Immune response			
Mdcontig9681	Mytimacin-5, partial [Mytilus galloprovincialis]	624	5.44E-16
Mdcontig2997	Hypothetical protein BRAFLDRAFT_81702 [Branchiostoma floridae]	1020	3.16E-04
Mdcontig2653	40S ribosomal protein S6 [Aplysia californica]	1040	1.42E-112
Mdcontig2547	Cathepsin L-like cysteine protease 2 [Plautia stali]	1451	1.86E-80
Mdcontig2353	Dolichyl-diphosphooligosaccharide–protein glycosyltransferase 48 kDa subunit-like [Crotalus adamanteus]	1558	7.77E-92
Mdcontig3019	Putative C1q domain-containing protein MgC1q67 [Mytilus galloprovincialis]	657	3.88E-07
Mdcontig2690	Lysozyme [Ruditapes philippinarum]	1179	5.27E-51
Reproduction			
Mdcontig2785	Hypothetical protein LOC100714763 [Cavia porcellus]	863	2.58E-60
Mdcontig2254	Zinc finger protein 84-like [Oreochromis niloticus]	1879	1.29E-03
Mdcontig2201	Integral membrane protein 2A [Crassostrea gigas]	2398	1.02E-20
Metabolic process			
Mdcontig9502	Aspartate aminotransferase, mitochondrial [Crassostrea gigas]	1159	0.0
Mdcontig9420	Cytochrome b-c1 complex subunit 8-like [Loxodonta africana]	820	1.05E-14
Mdcontig9404	Predicted protein [Trichoplax adhaerens]	917	6.97E-07
Mdcontig9389	Adenosylhomocysteinase A [Crassostrea gigas]	1153	1.39E-72
Mdcontig9330	DBH-like monooxygenase protein 1 [Crassostrea gigas]	2265	1.08E-75
Mdcontig779	Putative NAD-like isoform 1 [Bombus terrestris]	885	5.64E-59
Mdcontig7736	Extracellular copper/zinc superoxide dismutase [Hyriopsis cumingii]	851	2.95E-36
Mdcontig7650	EGF and pentraxin domain-containing protein 1 [Crassostrea gigas]	753	8.53E-30
Mdcontig7389	Hypothetical protein CGI_10011402 [Crassostrea gigas]	637	3.43E-46
Mdcontig7128	70-kilodalton heat shock protein, partial [Oncorhynchus mykiss gairdneri]	427	8.74E-72
Mdcontig7082	ATP synthase beta subunit [Pinctada fucata]	893	1.43E-120
Mdcontig7040	Hypothetical protein CPAR2_500330 [Candida parapsilosis]	914	6.78E-09
Mdcontig7011	Cartilage matrix protein [Crassostrea gigas]	1338	2.95E-40
Mdcontig6998	Beta-hexosaminidase [Crassostrea gigas]	1953	6.11E-12
Mdcontig6984	Phosphoenolpyruvate carboxykinase, cytosolic [GTP] [Crassostrea gigas]	2751	0.0
Mdcontig6926	Putative tyrosinase-like protein tyr-3 [Crassostrea gigas]	2155	1.38E-72
Mdcontig6913	rCG56483, isoform CRA_a [Rattus norvegicus]	1420	1.65E-161
Mdcontig6901	Malate dehydrogenase, cytoplasmic, partial [Crassostrea gigas]	1518	1.39E-153
Mdcontig6892	D-lactate dehydrogenase [Octopus vulgaris]	2044	5.98E-114
Mdcontig6881	Putative tyrosinase-like protein tyr-3 [Crassostrea gigas]	2326	1.45E-41
Mdcontig6875	EGF and pentraxin domain-containing protein 1 [Crassostrea gigas]	1632	2.34E-81
Mdcontig6859	Neutral and basic amino acid transport protein rBAT [Crassostrea gigas]	2468	3.35E-131
Mdcontig5307	Bifunctional heparan sulphate N-deacetylase/N-sulphotransferase 4 [Gallus gallus]	617	3.85E-26
Mdcontig5303	Cat eye syndrome critical region protein 5 [Crassostrea gigas]	671	4.96E-72
Mdcontig5268	Ribosomal protein L14-like [Saccoglossus kowalevskii]	607	1.91E-40
Mdcontig5239	Bifunctional protein NCOAT [Crassostrea gigas]	1156	5.48E-40
Mdcontig5224	NADH-ubiquinone oxidoreductase 75 kDa subunit, [Crassostrea gigas]	2216	0.0
Mdcontig4095	Solute carrier family 25 member 38-like isoform 1 [Nasonia vitripennis]	1132	5.97E-25
Mdcontig3840	Predicted protein [Nematostella vectensis]	550	2.78E-12
Mdcontig3130	Amine oxidase [Streptosporangium roseum DSM 43021]	938	6.69E-34
Mdcontig2987	Succinate dehydrogenase [ubiquinone] iron-sulphur subunit [Oreochromis niloticus]	416	5.79E-62
Mdcontig2695	Predicted protein [Trichoplax adhaerens]	1024	3.84E-08
Mdcontig2651	Pancreatic lipase-related protein 1 [Crassostrea gigas]	982	1.13E-77
Mdcontig2611	Chitinase-3 [Hyriopsis cumingii]	722	7.91E-64
Mdcontig2585	S-adenosylhomocysteine hydrolase [Crassostrea ariakensis]	917	0.0
Mdcontig2561	Microsomal glutathione S-transferase [Ruditapes philippinarum]	1397	6.74E-64
Mdcontig2430	Glutamine synthetase [Tegillarca granosa]	1286	0.0
Mdcontig2395	Triosephosphate isomerase [Crassostrea ariakensis]	745	3.30E-107
Mdcontig2282	Citrate synthase, mitochondrial [Crassostrea gigas]	1891	0.0
Mdcontig2255	Putative inhibitor of apoptosis [Crassostrea gigas]	1150	2.46E-42
Mdcontig2243	AF218064_1 malate dehydrogenase precursor [Nucella lapillus]	1887	1.62E-166
Mdcontig2203	Mitochondrial H+ ATPase a subunit [Pinctada fucata]	1931	0.0
Mdcontig2196	Mitochondrial-processing subunit betalike 2 [Strongylocentrotus purpuratus]	1429	3.05E-145

Table 2 (continued)

Seq. Name	Seq. Description	Seq. Length	Min. eValue
Mdcontig10167	Ubiquitin B, isoform CRA_e [Homo sapiens]	482	5.85E-41
Growth			
Mdcontig9398	Hypothetical protein CGI_10017178 [Crassostrea gigas]	2860	0.0
Mdcontig9287	Chaperone protein [Geodia cydonium]	1670	4.28E-63
Mdcontig9127	Ubiquitin, partial [Homarus americanus]	574	2.59E-26
Mdcontig7013	Thrombospondin-3 [Crassostrea gigas]	1202	5.86E-120
Mdcontig6881	Putative tyrosinase-like protein tyr-3 [Crassostrea gigas]	2326	1.45E-41
Mdcontig6461	Polyubiquitin [Ictalurus punctatus]	719	1.96E-31
Mdcontig5239	Bifunctional protein NCOAT [Crassostrea gigas]	1156	5.48E-40
Mdcontig5229	GJ20779 [Drosophila virilis]	1108	9.61E-50
Mdcontig3837	Hypothetical protein MYCTH_2033977, partial [Myceliophthora thermophila]	675	7.39E-15
Mdcontig3346	Intermediate filament protein Ov71 (fragment) [Brugia malayi]	583	2.42E+00
Mdcontig2765	Vascular endothelial growth factor D [Crassostrea gigas]	1122	1.03E-30
Mdcontig2580	Receptor for activated protein kinase C [Scrobicularia plana]	1361	0.0
Mdcontig2294	Putative Na+/K + -ATPase alpha subunit [Paroctopus digueti]	1802	1.14E-132
Mdcontig10167	Ubiquitin B, isoform CRA_e [Homo sapiens]	482	5.85E-41
Biogenesis			
Mdcontig9393	60S ribosomal protein L11 [Danio rerio]	908	3.02E-103
Mdcontig9316	Hoip-prov protein isoform A [Lysiphlebus testaceipes]	855	3.94E-55
Mdcontig7015	60S ribosomal L7a-like protein [Phragmatopoma lapidosa]	951	2.85E-126
Mdcontig7008	Similar to ribosomal protein L5 [Ciona intestinalis]	1024	6.52E-145
Mdcontig6997	60S acidic ribosomal protein P0 [Crassostrea gigas]	1318	1.26E-142
Mdcontig6931	Brix domain-containing protein 2 [Crassostrea gigas]	951	1.81E-118
Mdcontig5358	Ribosomal protein S7 [Argopecten irradians]	635	2.21E-97
Mdcontig5268	Ribosomal protein L14-like [Saccoglossus kowalevskii]	607	1.91E-40
Mdcontig3220	Putative ribosomal protein S7 [Sipunculus nudus]	936	2.87E-19
Mdcontig2594	Hypothetical protein BRAFLDRAFT_114927 [Branchiostoma floridae]	481	5.93E-80
Mdcontig2534	Putative ribosome biogenesis protein RLP24 [Crassostrea gigas]	969	3.01E-76
Response to stimulus			
Mdcontig9350	Major vault protein [Crassostrea gigas]	1047	1.53E-40
Mdcontig7693	Heat shock protein 70 [Meretrix meretrix]	575	9.93E-38
Mdcontig7128	70-kilodalton heat shock protein, partial [Oncorhynchus mykiss gairdneri]	427	8.74E-72
Mdcontig6984	Phosphoenolpyruvate carboxykinase, cytosolic [GTP] [Crassostrea gigas]	2751	0.0
Mdcontig6913	rCG56483, isoform CRA_a [Rattus norvegicus]	1420	1.65E-161
Mdcontig6884	cAMP-responsive element binding protein [Crassostrea ariakensis]	1437	6.21E-10
Mdcontig6861	Hypothetical protein BRAFLDRAFT_115434 [Branchiostoma floridae]	1797	6.96E-158
Mdcontig5296	Heat shock protein 70 [Cristaria plicata]	834	8.38E-92
Mdcontig5239	Bifunctional protein NCOAT [Crassostrea gigas]	1156	5.48E-40
Mdcontig3776	Heat shock protein 22 isoform 1 [Ruditapes philippinarum]	1159	3.55E-10
Mdcontig3019	Putative C1q domain-containing protein MgC1q67 [Mytilus galloprovincialis]	657	3.88E-07
Mdcontig2997	Hypothetical protein BRAFLDRAFT_81702 [Branchiostoma floridae]	1020	3.16E-04
Mdcontig2935	Similar to lethal (3) neo18 CG9762-PA [Tribolium castaneum]	631	4.82E-14
Mdcontig2816	Hypothetical protein BRAFLDRAFT_125323 [Branchiostoma floridae]	694	2.58E-09
Mdcontig2721	HSP70 [Danaus plexippus]	837	7.33E-07
Mdcontig2690	Lysozyme [Ruditapes philippinarum]	1179	5.27E-51
Mdcontig2561	Microsomal glutathione S-transferase [Ruditapes philippinarum]	1397	6.74E-64
Mdcontig2547	Cathepsin L-like cysteine protease 2 [Plautia stali]	1451	1.86E-80
Mdcontig2538	Heat shock protein 90 [Laternula elliptica]	2784	0.0
Mdcontig2274	Serine/threonine protein kinase PINK1, mitochondrial-like [Nasonia vitripennis]	1932	2.75E-86

response to environmental stress. For instance, an interactive catalogue of 7112 transcripts of *M. galloprovincialis*, called Mytibase, offers the opportunity to look for relevant gene sequences. In particular, innate immunity-related genes such as antimicrobial peptides, complement C1q, C-type lectins, fibrinogen-like transcripts and many carbohydrate-binding proteins have emerged as the most abundant genes in host defence systems (Venier *et al.* 2011). Next-generation sequencing studies have characterized genes associated with reproduction and analysed their expression patterns and

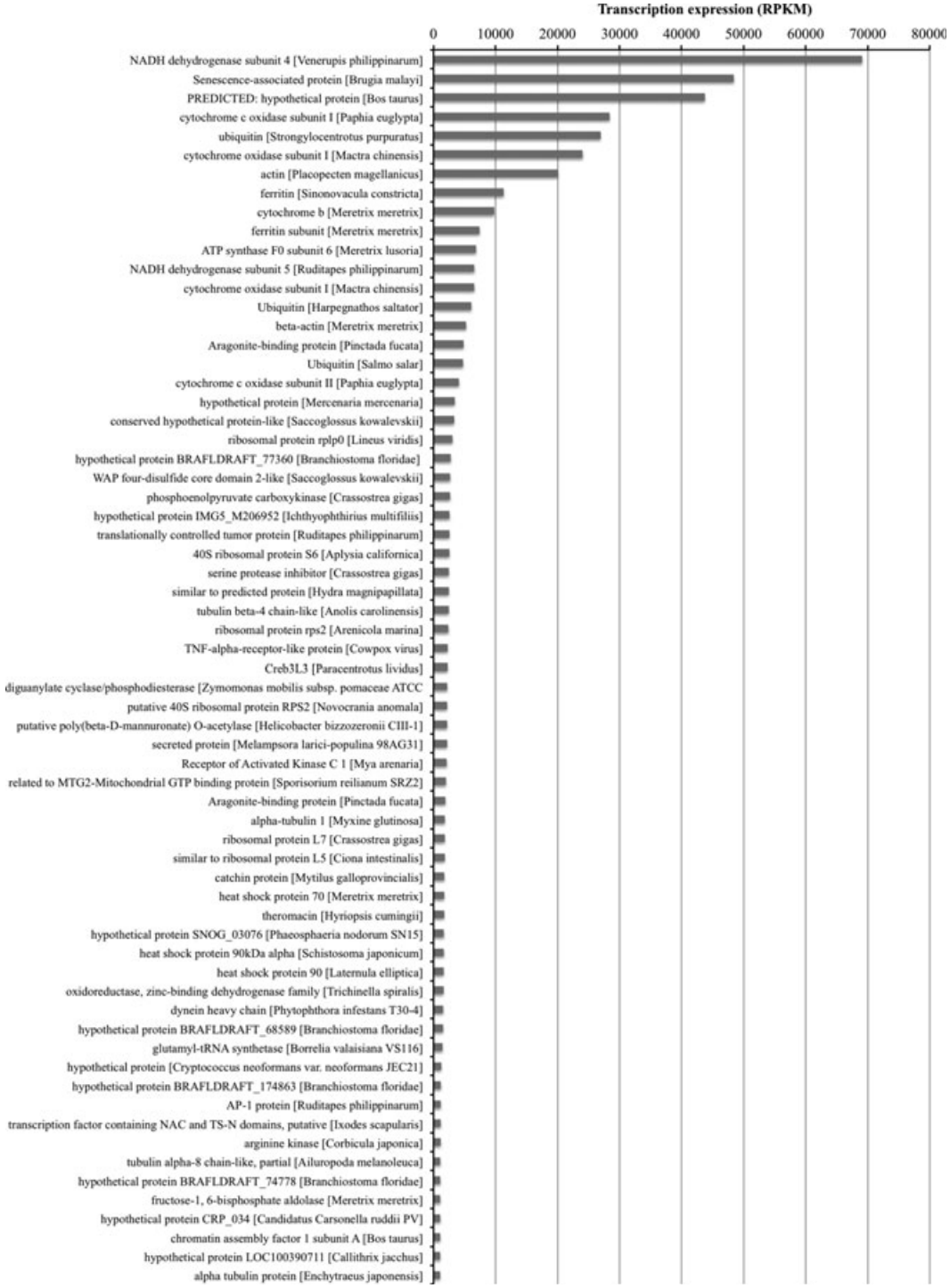

Figure 4 Top-hit transcript expression for contigs annotated with single-nucleotide polymorphism variants. Cut-off values were set at 1000 RPKM.

polymorphism, providing insights into the molecular mechanisms regulating sex determination. Ghiselli, Milani, Chang, Hedgecock, Davis, Nuzhdin and Passamonti (2012) compared the transcriptomes of male and females in the manila clam *Ruditapes philippinarum* and identified 1575 genes with strong

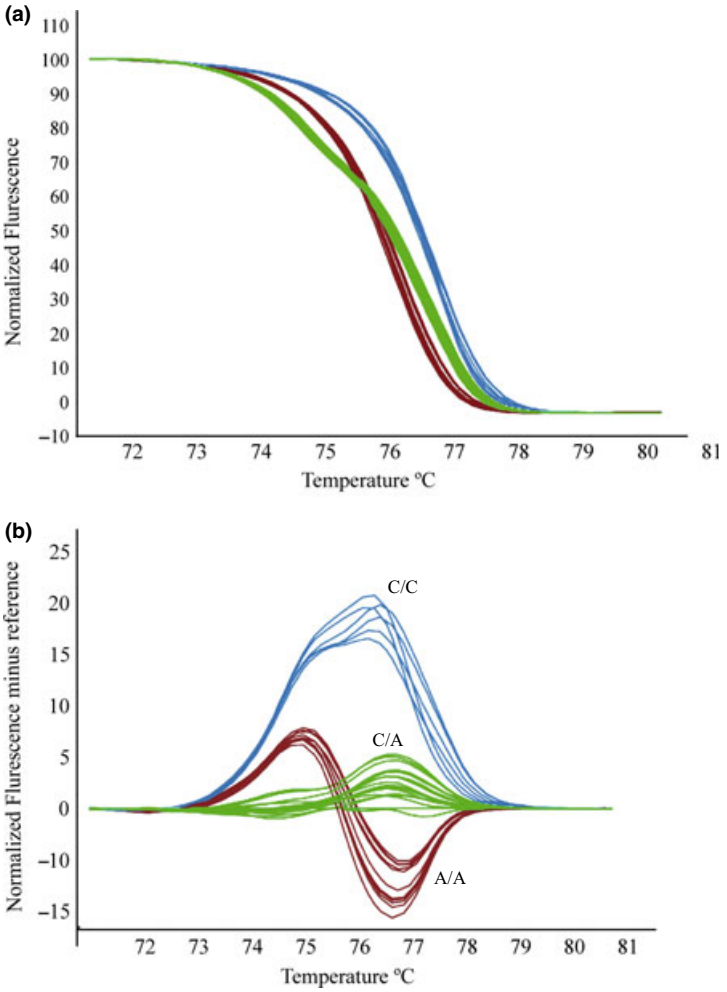

Figure 5 High-resolution melting analysis (HRMA) for heat shock 70 kDa gene (*HSP70*). Normalized HRM curves (a) and difference plots (b).

sex-specific expression and 166 sex-specific SNPs. Moreover, a recent study by Huan, Wang and Liu (2012) reported large-scale RNA 454 pyrosequencing of different larval stages of the clam *Meretrix meretix*, obtaining pivotal genomic information about genes that could be involved in the ontogenetic development of marine invertebrates.

While we did not aim to carry out a deep transcriptome characterization in the soft clam *M. donacium*, generating a high coverage of information with respect to a particular biological process, the objective of this research was to identify a set of SNPs that could be used as markers for future population genetic investigations. Herein, neutral molecular markers, such as mtDNA and microsatellites, have provided important insights into the population structure and demographic history of wild populations, but they cannot detect adaptive genetic variation (Ouborg, Pertoldi, Loeschcke, Bijlsma & Hedrick 2010). However, designing SNP markers from NGS transcriptome data requires substantial bioinformatic effort to avoid paralogous genes, alternative splicing and transcription noise during the assembling process (Cahais, Gayral, Tsagkogeorga, Melo-Ferreira, Ballenghien & Galtier 2012).

De novo assembly was used in the absence of a reference genome to the SNP discovery in the surf clam. From 10 178 assembled contigs, we found 2594 putative SNPs in 613 sequences. The observed transition/transversion ratio (ti/tv) among the fifty individuals using *de novo* assembly approach was approximately 1.10, which is lower than the 2.3 ratio reported for humans (Deutsch, Iseli, Bucher, Antonarakis & Scott 2001), but similar to the ratios in the entire genome of *Salmo salar* of 1.37 (Hayes, Laerdahl, Lien, Moen, Berg, Hindar, Davidson, Koop, Adzhubei & Hoyheim 2007), 1.49 for *Oncorhynchus tshawytscha* (Smith, Elfstrom, Seeb & Seeb 2005) and 1.76 for rat

(Guryev, Berezikov, Malik, Plasterk & Cuppen 2004).

The SNP frequency achieved by *de novo* assembly was found with 1/260 SNPs per base pair (SNP/bp) and 1/2279 SNP/bp. The first frequency was estimated using data set that was exclusively annotated to SNP variants, while the second one considers all data set. These results were lower than those reported by Brumfield, Beerli, Nickerson and Edwards (2003), in which the frequency in non-model species was estimated at one in 200–500 bases of noncoding DNA and one in 500–1000 bases of coding DNA. With respect to bivalves, Clark, Thorne, Vieira, Cardoso and Power (2010) using 454 pyrosequencing of mantle transcriptome of *Laternula elliptica* reported an SNP frequency of one for every 100 bases. Further, reported data from the eastern oyster showed that in 6.8 Kb of genome, 336 putative SNPs were archived. The average density of SNPs was estimated to be one for every 20 bp. The 6.8 Kb sequences included 2462 bp of introns, including 156 putative SNPs. For expressed sequences, the frequency of SNPs was one per 24 bp (Zhang & Guo 2010). The reason for both estimations could be based on the lower transcriptome coverage generated from *M. donacium*, and also, the differences reported in our study with respect to the SNP frequency may result from the detection method applied and the species or population used. However, another reason for the differences among these estimates might be that our depth of coverage criterion excluded thousands of potential candidates with observed frequencies in contigs with a depth <8. The observed frequencies are of course dependent on the parameters selected for assembly, SNP identification and validation, as well as the regions sequenced. In this context, several studies have compared different available assemblers, such as CAP3, MIRA, Newbler, SeqMan, Oases and CLC, to establish best practices for transcriptome assembly (Kumar & Blaxter 2010; Mundry, Bornberg-Bauer, Sammeth & Feulner 2012).

In the present study, consensus sequences containing putative SNPs were selected for further Gene Ontology analysis. A significant percentage of transcripts were assigned protein and nucleotide binding, as well hydrolase activity, ion binding and oxidoreductase activity regarding molecular function level. With respect to biological processes like cellular and metabolic process, biogenesis, localization and biological regulation were highly annotated. Additional process involved in developmental process, response to stimulus, immune system, growth and biological adhesion was less achieved. These results are congruent with previous reports in no-model species (Gao, Luo, Liu, Zeng, Liu, Yi & Wang 2012; Ma, Qiu, Feng & Li 2012). Furthermore, gene expression analysis of these contigs revealed that the most expressed proteins are related to the mitochondrial electron transport chain such as NADH dehydrogenase subunit 4, cytochrome c oxidase subunit I, ATP synthase F0 subunit 6 and cytochrome b. Further genes were also highly expressed such as senescence-associated protein, ubiquitin and actin. Interestingly, some relevant genes related to immune response and biomineralization showed a high abundance, such as TNF-alpha-receptor-like protein, serine protease inhibitor, heat shock protein 70, 90, aragonite-binding protein and ferritin. Although some relevant genes associated with immune response, reproduction, stress and growth have been shown to be highly conserved among taxa because of their critical functions in immune response to stressors (Feder & Hofmann 1999),, high levels of base pair polymorphism in these genes may be necessary. For instance, gene expression related to temperature stress (e.g. *HSPs*) may vary seasonally, ontogenetically and with temperature. Indeed, high levels of sequence variation have been observed in regulatory regions of HSP among a wide variety of organisms (Narum & Campbell 2010). Further, in *Litopenaeus vannamei*, genetic polymorphisms of the *HSP70* gene were shown to be associated with Taura syndrome virus resistance (Zeng, Chen, Li, Peng, Ma, Jiang, Yang & Li 2008). Recently, Yu, He, Wang, Zhang, Bao and Guo (2011) identified SNPs in the serine protease inhibitor gene and studied its association with improved survival after disease-caused mortalities and in disease-resistant eastern oyster strains against *Perkinsus marinus*.

Overall, our study demonstrated successful SNP identification in nonreference genome and identification of relevant genes containing putative SNPs. Herein, validation of putative SNPs was carried out using HRMA. High-resolution melting analysis allows identification of genotypic differences by changes on the melting temperature of each of the homozygotes and allows distinguishing heterozygotes at lower costs than alternative methods (Liew, Pryor, Palais, Meadows, Erali, Lyon & Wittwer 2004), allowing a high-throughput alternative for SNP screening for genetic studies of nonmodel

population and showing reliable results due to its high sensitivity (Smith, Lu & Alvarado Bremer 2013). Our study provides a first panel of 24 SNPs identified in *M. donacium* transcriptome. Future studies will seek to validate a major panel of polymorphic SNPs and to investigate whether there is gene transcription variation that supports genetic adaptive variation among wild populations of *M. donacium*. In addition, some candidate genes will be useful to achieve novel molecular knowledge of the immune response, reproduction and growth of this native bivalve species.

Acknowledgments

This work was supported by a FONDECYT (1120397) and FONDAP (15110027) from CONICYT-Chile.

References

Aguilar-Espinoza A., Guzmán-Riffo B., Haye P.A. & Gallardo-Escárate C. (2012) Mining of EST-SSR from 454 pyrosequencing in the surf clam *Mesodesma donacium* (Lamark, 1818). *Conservation Genetics Resources* **4**, 829–832.

Amar G., Rojas C.P., von Brand E. & Jara-Seguel P. (2008) Karyotype study in the surf clam *Mesodesma donacium* Lamarck, 1818 (Bivalvia : Veneroida : Mesodesmatidae). *Gayana* **72**, 18–22.

Bellin D., Ferrarini A., Chimento A., Kaiser O., Levenkova N., Bouffard P. & Delledonne M. (2009) Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *Bmc Genomics* **10**, 1–9.

Bhasker C.R. & Hardiman G. (2010) Advances in pharmacogenomics technologies. *Pharmacogenomics* **11**, 481–485.

Bouck A. & Vision T. (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16**, 907–924.

Brumfield R., Beerli P., Nickerson D. & Edwards S. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18**, 249–256.

Cahais V., Gayral P., Tsagkogeorga G., Melo-Ferreira J., Ballenghien M. & Galtier N. (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources* **12**, 834–845.

Cardenas L., Sanchez R., Gomez D., Fuenzalida G., Gallardo-Escarate C. & Tanguy A. (2011) Transcriptome analysis in *Concholepas concholepas* (Gastropoda, Muricidae): mining and characterization of new genomic and molecular markers. *Marine Genomics* **4**, 197–205.

Cerda J., Douglas S. & Reith M. (2010) Genomic resources for flatfish research and their applications. *Journal of Fish Biology* **77**, 1045–1070.

Clark M.S., Thorne M., Vieira F.A., Cardoso J.C.R. & Power D.M. (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., Gotz S., Garcia-Gomez J.M., Terol J., Talon M. & Robles M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676.

Deutsch S., Iseli C., Bucher P., Antonarakis S.E. & Scott H.S. (2001) A cSNP map and database for human chromosome 21. *Genome Research* **11**, 300–307.

Everett M.V., Grau E.D. & Seeb J.E. (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources* **11**, 93–108.

Feder M.E. & Hofmann G.E. (1999) Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annual Review of Physiology* **61**, 243–282.

Gao Z.X., Luo W., Liu H., Zeng C., Liu X.L., Yi S.J. & Wang W.M. (2012) Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). *PLoS ONE* **7**, e42637.

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.

Guryev V., Berezikov E., Malik R., Plasterk R.H.A. & Cuppen E. (2004) Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Research* **14**, 1438–1443.

Hayes B., Laerdahl J.K., Lien S., Moen T., Berg P., Hindar K., Davidson W.S., Koop B.F., Adzhubei A. & Hoyheim B. (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture* **265**, 82–90.

Huan P., Wang H.X. & Liu B.Z. (2012) Transcriptomic analysis of the clam *Meretrix meretrix* on different larval stages. *Marine Biotechnology* **14**, 69–78.

Hudson M.E. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**, 3–17.

Johansen S.D., Karlsen B.O., Furmanek T., Andreassen M., Jorgensen T.E., Bizuayehu T.T., Breines R., Emblem A., Kettunen P., Luukko K., Edvardsen R.B., Nordeide J.T., Coucheron D.H. & Moum T. (2011) RNA deep sequencing of the Atlantic cod transcriptome. *Comparative*

Biochemistry and Physiology D-Genomics and Proteomics **6**, 18–22.

Kumar S. & Blaxter M.L. (2010) Comparing de novo assemblers for 454 transcriptome data. *Bmc Genomics* **11**, 571.

Liew M., Pryor R., Palais R., Meadows C., Erali M., Lyon E. & Wittwer C. (2004) Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. *Clinical Chemistry* **50**, 1156–1164.

Ma K.Y., Qiu G.F., Feng J.B. & Li J.L. (2012) Transcriptome analysis of the oriental river prawn, *Macrobrachium nipponense* using 454 pyrosequencing for discovery of genes and markers. *PLoS ONE* **7**, e39727.

Marguerat S. & Bahler J. (2010) RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences* **67**, 569–579.

Marins L.F. & Levy J.A. (1999) High genetic distance between marine bivalves of the genus *Mesodesma* inhabiting the Atlantic and Pacific coasts of South America. *Comparative Biochemistry and Physiology a-Molecular and Integrative Physiology* **124**, 313–319.

Mundry M., Bornberg-Bauer E., Sammeth M. & Feulner P.G.D. (2012) Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PLoS ONE* **7**, e31410.

Narum S.R. & Campbell N.R. (2010) Sequence divergence of heat shock genes within and among 3 oncorhynchids. *Journal of Heredity* **101**, 107–112.

Neira-Oviedo M., Tsyganov-Bodounov A., Lycett G.J., Kokoza V., Raikhel A.S. & Krzywinski J. (2011) The RNA-Seq approach to studying the expression of mosquito mitochondrial genes. *Insect Molecular Biology* **20**, 141–152.

Ouborg N.J., Pertoldi C., Loeschcke V., Bijlsma R. & Hedrick P.W. (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics* **26**, 177–187.

Peralta G. (2008) *Patrones filogeográficos en el bivalvo Mesodesma donacium Lamarck (1818) "macha" en Chile*. Diss. MSc thesis, Universidad de Chile, Santiago.

Renaut S., Nolte A.W. & Bernatchez L. (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology* **19** (Suppl. 1), 115–131.

Rice A.M., Rudh A., Ellegren H. & Qvarnstrom A. (2011) A guide to the genomics of ecological speciation in natural animal populations. *Ecology Letters* **14**, 9–18.

Saavedra C. & Bachère E. (2006) Bivalve genomics. *Aquaculture* **256**, 1–14.

Seeb J.E., Carvalho G., Hauser L., Naish K., Roberts S. & Seeb L.W. (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* **11**, 1–8.

Smith C.T., Elfstrom C.M., Seeb L.W. & Seeb J.E. (2005) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology* **14**, 4193–4203.

Smith B.L., Lu C.-P. & Alvarado Bremer J.R. (2013) Methodological streamlining of SNP discovery and genotyping via high-resolution melting analysis (HRMA) in non-model species. *Marine Genomics* **9**, 39–49.

Tanguy A., Bierne N., Saavedra C., Pina B., Bachere E., Kube M., Bazin E., Bonhomme F., Boudry P., Boulo V., Boutet I., Cancela L., Dossat C., Favrel P., Huvet A., Jarque S., Jollivet D., Klages S., Lapegue S., Leite R., Moal J., Moraga D., Reinhardt R., Samain J.F., Zouros E. & Canario A. (2008) Increasing genomic information in bivalves through new EST collections in four species: development of new genetic markers for environmental studies and genome evolution. *Gene* **408**, 27–36.

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.

Vera J.C., Wheat C.W., Fescemyer H.W., Frilander M.J. & Crawford D.L. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**, 1636–1674.

Wheat C.W. (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica* **138**, 433–451.

Wilhelm B.T. & Landry J.R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257.

Yu H.Y., He Y., Wang X.X., Zhang Q.Q., Bao Z.M. & Guo X.M. (2011) Polymorphism in a serine protease inhibitor gene and its association with disease resistance in the eastern oyster (*Crassostrea virginica* Gmelin). *Fish and Shellfish Immunology* **30**, 757–762.

Zeng D.G., Chen X.H., Li Y.M., Peng M., Ma N., Jiang W.M., Yang C.L. & Li M. (2008) Analysis of Hsp70 in *Litopenaeus vannamei* and Detection of SNPs. *Journal of Crustacean Biology* **28**, 727–730.

Zhang L.S. & Guo X.M. (2010) Development and validation of single nucleotide polymorphism markers in the eastern oyster *Crassostrea virginica* Gmelin by mining ESTs and resequencing. *Aquaculture* **302**, 124–129.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Blast of contigs that contained SNP markers.

Data S2. Gene ontology analysis of sequences that contained SNPs.

Data S3. Sequences of PCR primers used for HRM analysis.