Marine Genomics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

Cells to Shells

Mining the transcriptomes of four commercially important shellfish species for single nucleotide polymorphisms within biomineralization genes

David L.J. Vendrami [a,*], Abhijeet Shah [a], Luca Telesca [b], Joseph I. Hoffman [a]

[a] Department of Animal Behaviour, University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany
[b] Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, Cambridgeshire CB2 3EQ, UK

ARTICLE INFO

Article history:
Received 8 October 2015
Received in revised form 17 December 2015
Accepted 23 December 2015
Available online xxxx

Keywords:
Crassostrea gigas
Mya truncata
Mytilus edulis
Pecten maximus
Non-model organism

ABSTRACT

Transcriptional profiling not only provides insights into patterns of gene expression, but also generates sequences that can be mined for molecular markers, which in turn can be used for population genetic studies. As part of a large-scale effort to better understand how commercially important European shellfish species may respond to ocean acidification, we therefore mined the transcriptomes of four species (the Pacific oyster *Crassostrea gigas*, the blue mussel *Mytilus edulis*, the great scallop *Pecten maximus* and the blunt gaper *Mya truncata*) for single nucleotide polymorphisms (SNPs). Illumina data for *C. gigas*, *M. edulis* and *P. maximus* and 454 data for *M. truncata* were interrogated using GATK and SWAP454 respectively to identify between 8267 and 47,159 high quality SNPs per species (total = 121,053 SNPs residing within 34,716 different contigs). We then annotated the transcripts containing SNPs to reveal homology to diverse genes. Finally, as oceanic pH affects the ability of organisms to incorporate calcium carbonate, we honed in on genes implicated in the biomineralization process to identify a total of 1899 SNPs in 157 genes. These provide good candidates for biomarkers with which to study patterns of selection in natural or experimental populations.

1. Introduction

Climate change is one of the major factors threatening biodiversity in the 21st century (Millennium Ecosystem Assessment, 2005) and may even surpass habitat destruction as the greatest stressor in the coming decades (Leadley, 2010). Climate change is affecting the distributions, abundances, behaviors, physiologies and phenologies of many organisms (Bradshaw and Holzapfel, 2006, Forcada and Hoffman, 2014, Franks and Hoffmann, 2012). The sheer pace of ongoing change has led to mounting concerns over whether species will be able to adapt fast enough to survive (Hoffmann and Sgrò, 2011, Parmesan, 2006, Shaw and Etterson, 2012).

One important and pervasive consequence of climate change is ocean acidification (Howes et al., 2015). Atmospheric CO_2 levels have risen from 280 to >390 ppm since the onset of the industrial revolution (http://www.esrl.noaa.gov/gmd/ccgg/trends/ last accessed: October 7 2015) and average surface ocean pH has fallen from 8.16 to 8.05 over the same period (Cao and Caldeira, 2008), leading to a concomitant reduction in the availability of carbonate ions. This affects the ability of organisms such as molluscs to incorporate calcium carbonate, which is essential for building and maintaining robust skeletons and shells (Doney et al., 2009).

Genetic and genomic studies are instrumental to understanding the impacts of ocean acidification and the mechanisms by which various species might adapt to changing oceanic pH (Pespeni et al., 2013, 2012). Approaches that exploit recent advances in high-throughput sequencing, such as transcriptional profiling, are particularly powerful as they can generate vast amounts of sequence data without the need for prior genomic resources (Ekblom and Galindo, 2011). The resulting data can provide insights into patterns of gene expression among other things, and can also be used to develop single nucleotide polymorphisms (SNPs) for use in population genetic studies. Custom SNPs can be genotyped in their tens to millions, depending on the specific assay used, but most technologies rely on allele-specific oligonucleotide probes designed from the SNP flanking sequences (Qian et al., 2015).

Crassostrea gigas (the Pacific oyster), *Mytilus edulis* (the blue mussel), *Pecten maximus* (the great scallop) and *Mya truncata* (the blunt-gaper clam), are four non-model bivalve species whose shells show differences in composition with respect to the amounts of the two calcium carbonate polymorphs (calcite and aragonite), microstructure crystal size and organic matrix content (Checa and Rodríguez-Navarro, 2005; Griesshaber et al., 2013, Taylor et al., 1969). Despite their ecological and economic importance within the European Union and wider afield, relatively little is known about the capacity of these species to respond

* Corresponding author at: Department of Animal Behavior, University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany.
E-mail address: david.vendrami@student.unife.it (D.L.J. Vendrami).

http://dx.doi.org/10.1016/j.margen.2015.12.009
1874-7787/

to environmental change. An important first step has been to sequence their mantle (the tissue that secretes the shell) transcriptomes (Sleight et al., in this issue; Yarra et al., in this issue) in order to help illuminate the mechanisms underlying shell formation and maintenance through the analysis of patterns of differential gene expression. However, the same data also lend themselves to SNP discovery. Large genome-wide distributed panels of SNPs would be useful for population and quantitative genetic studies, while SNPs mined within genes specifically relating to biomineralization could be used in a candidate gene approach to explore the genetic basis of phenotypic variability and to elucidate patterns of selection in relation to prevailing environmental conditions.

Here we mined the transcriptomes of the four species using stringent criteria to identify many tens of thousands of high-quality SNPs residing within functionally annotated transcripts. Through comparisons with other studies, we then identified SNPs located within genes known to play important roles in the biomineralization process. These markers could be used for genetic studies of the four species.

2. Material and methods

2.1. Raw data

Details of the transcriptomic data that we mined for SNPs are provided in Table 1. For *C. gigas*, *M. edulis* and *P. maximus*, we used Illumina HiSeq data generated by Yarra et al. (in this issue) whereas for *M. truncata* we used 454 data generated by Sleight et al. (in this issue). Briefly, transcriptomes were assembled by Yarra et al. (in this issue) for the first three species using between 13 and 14 unrelated, wild-caught individuals per species that were subjected to a shell repair experiment, in which experimental individuals had holes drilled in their shells and transcriptional profiling of the mantle tissue was subsequently carried out. For *M. truncata*, RNA from the mantle tissues of 9 unrelated wild-caught individuals was sequenced. The raw sequence reads and assembled contigs are available both in the original papers and via SRA (accession number: SRP064949) and http://bit.ly/1QcFiVH (*M. truncata*) and via SRA (accession number: SRP067223) and http://molluscdb.afterparty.bio.ed.ac.uk (*C. gigas*, *M. edulis* and *P. maximus*) respectively. The transcriptomes of Yarra et al. (in this issue) have already been stringently quality filtered by the author, with all reads trimmed to remove low-quality base calls. We therefore conducted a quality control check on the *M. truncata* data using the program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc last accessed: October 7 2015). Afterwards, we trimmed away the last bases of the reads with Phred scores below 30 and discarded reads shorter than 40 bp (after trimming) using FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit last accessed: October 7 2015).

2.2. SNP discovery in C. gigas, M. edulis and P. maximus

We used the Bowtie2 package (Langmead and Salzberg, 2012) to index the previous *de novo* assembled transcriptomes, align the raw reads back to the respective transcriptomes and add read groups. The resulting SAM files were then merged together and processed using Samtools (Li et al., 2009) and Picard (http://broadinstitute.github.io/picard last accessed: October 7 2015) to generate a unique indexed BAM file from which duplicate reads were removed. Subsequently, the Genome Analysis Toolkit (GATK, McKenna et al., 2010) was applied to detect variant sites with high confidence as follows: firstly, we used 'RealignerTargetCreator' and 'IndelRealigner' to identify areas containing Indels and perform a realignment step. Then, 'HaplotypeCaller' was used to create a vcf file containing a list of all of the sites recognized either as Indels or as SNPs as well as relevant information concerning the accuracy of each call. Finally, we used 'VariantsToTable' to export the data into a tab delimited file, and filtered the resulting SNPs within R (R Core Team, 2015) to retain only high-quality, informative SNPs. Specifically, we retained only SNPs that met the following criteria: total number of alleles in called genotypes (AN) ≥20, minor allele frequency (MAF) ≥0.1, Phred-scaled *p*-value using Fisher's exact test to detect strand bias (FS) ~0, depth of coverage (DP) ≥8, Z-score from Wilcoxon rank sum test of alternative *versus* reference base qualities (BaseQRankSum) ≤−1.96 or ≥1.96, root mean square of the mapping quality of reads across all samples (MQ) ≥40 and variant confidence normalized by unfiltered depth of variant samples (QD) ≥5. Finally, we evaluated the SNP parameter space for each set of called SNPs by computing a two dimensional Kernel density estimation using the 'kde2d' function in the R package MASS (Venables and Ripley, 2002). This approach uses a smoothing function to estimate a probability distribution in two dimensions (MAF and DP) for data visualization.

2.3. SNP discovery in M. truncata

We used SWAP454 (Brockman et al., 2008), which first maps the raw reads back to the assembled contigs and then determines, while taking in account an error model for the 454 data, which positions are called as SNPs according to two user-specified thresholds. The first of these, 'MIN_RATIO' corresponds to the percentage of reads that differ from the reference sequence at a given position and the second, 'MIN_READS' to the number of copies present of the minor allele. To minimize the possibility of false positives arising from sequencing error, MIN_RATIO was set to 0.1 and MIN_READS was set to 5. The SNP parameter space was evaluated as described for the other three species.

2.4. Flanking sequences

We used the BEDtools software package (Quinlan, 2014) to extract the flanking sequences of each SNP (60 bp on either side) for probe design. We discarded SNPs with <50 bp of flanking sequence due to proximity to the start or end of contigs.

2.5. Sequence annotation

For each species, we filtered the original transcriptome to create a subset FASTA file comprising only those contigs within which at least one SNP was found. We performed blastx sequence similarity searching of these contig sequences against the GenBank nr and Swiss-Prot protein databases with a standard e-value cutoff of $1e^{-10}$. When a search produced more than one result for a given contig, only the annotation with the lowest e-value was retained. Finally, we retrieved Gene Ontology (GO) annotations for each contig. These operations were performed as described by De Wit et al. (2012). Subsequently, we classified GO annotations according to three major functional categories: biological processes, molecular function and cellular component. The Web Gene Ontology Annotation Plotting tool (Ye et al., 2006) was then utilized to obtain GO slim annotations and these were used to plot the distribution (%) of gene ontology terms among the annotated unique

Table 1
Summary of the raw data available for the four shellfish species.

Species	Literature reference	Type of data	Number of individuals sequenced	Collection site (s)	Number of reads
C. gigas	Yarra et al. (in this issue)	Illumina paired end (125 + 125 bp)	13	Barmore Bay (Scotland, UK)	208.1 million
M. edulis	Yarra et al. (in this issue)	Illumina paired end (125 + 125 bp)	14	Tarbert (Scotland, UK)	286 million
P. maximus	Yarra et al. (in this issue)	Illumina paired end (125 + 125 bp)	14	Eilean Buidhe Island (Scotland, UK)	180.4 million
M. truncata	Sleight et al. (in this issue)	454 single end (500 bp)	9	Dunstaffnage Bay (Scotland, UK)	702,006

Please cite this article as: Vendrami, D.L.J., et al., Mining the transcriptomes of four commercially important shellfish species for single nucleotide polymorphisms within biomineraliz..., Mar. Genomics (2016), http://dx.doi.org/10.1016/j.margen.2015.12.009