Marine Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

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

Method paper

A phylogenetic classification of gastropod aquaporins

D.J. Colgan[a,*], R.P. Santos[a,b]

[a] Malacology, The Australian Museum, 1 William St., Sydney, NSW 2010, Australia
[b] Laboratório de Recursos Genéticos, Universidade Federal de São João del Rei, Campus CTAN, São João Del Rei-MG, CEP 36307-352, Brazil

ARTICLE INFO

Keywords:
Mollusca
Evolution
Habitat transitions
Multi-gene family

ABSTRACT

Successful responses to the multifarious challenges of controlling water balance are critical for snails' survival in the great diversity of habitats they occupy. Advances are being made in understanding how such challenges are approached at the molecular level, including through the study of aquaporins, which are proteins functioning to facilitate the passage of water and other small molecules across cellular membranes. Deduced aquaporin amino acid sequences from partial genomic assemblies of three neritimorph species were added to available gastropod data and sequences from other taxa to make a phylogenetic classification of these proteins using maximum likelihood and Bayesian analyses. We identified ten groups, designated as G1 to G10, containing sequences from multiple major gastropod lineages. At least six of the groups appear to be encoded by multiple genes within at least some species. Five weakly-associated sequences from Neritimorpha were not allocated to a group. The designated groups G1, G2, G3, G4, G5 and G7 (previously defined as Malacoglyceroporins) formed clades containing only gastropod sequences and were strongly supported by Bayesian inference. G1, G2, G3 and G5 were also strongly supported by maximum likelihood analyses. Group G6 (previously defined as Malacoaquaporins)was included with sequences from the oyster, *Crassostrea gigas* in a strongly supported clade. Groups G8 and G9 included only gastropod sequences but were not strongly supported. Groups G8 and G10 were designated to include all the gastropod sequences belonging respectively to strongly-supported clades including human aquaglyceroporins and aquaammoniaporins. Most groups have been found in a wide range of gastropod lineages but all identified representatives of group G7 belong to Apogastropoda whereas G2 is known only from Patellogastropoda and Neritimorpha.

1. Introduction

Gastropods are remarkable for the diversity of environments they successfully occupy by individual responses to environmental variation or habitat transitions during lineage evolution. They exhibit a broad range of morphological, physiological and behavioural adaptations to assist individuals' responsive capability (reviewed by Little, 1981; Prior, 1985, 1989; Andrews, 1988). Habitat transitions between marine, brackish and freshwaters and terrestrial environments have occurred at a wide diversity of phylogenetic levels particularly in Panpulmonata (Klussmann-Kolb et al., 2008; Holznagel et al., 2010), Neritimorpha (Scott and Kenny, 1998; Kano et al., 2002; Uribe et al., 2016) and within Caenogastropoda, particularly in Rissooidea (Strong et al., 2008) and Cerithiodea (Strong et al., 2011).

Successful responses to the multifarious challenges of controlling water balance are critical for snails' survival in the great diversity of their habitats. Advances are being made in understanding how such problems are approached at the molecular level, including through the study of aquaporins. These are proteins whose function is generally to provide channels through the membrane into which they are integrated that facilitate the flow of water and other small molecules (including glycerol, urea and volatiles such as carbon dioxide and ammonia) down concentration gradients (Preston et al., 1992; Zardoya and Villalba, 2001; Takata et al., 2004; Kruse et al., 2006). Nearly all aquaporins have some capacity to transport water molecules (Perez Di Giorgio et al., 2014; Finn and Cerdà, 2015). Many have additional capabilities, often including glycerol transport, and frequently including ammonia and urea (Perez Di Giorgio et al., 2014).

The phylogenetic approach to the classification of aquaporins that began with the investigations of Zardoya and Villalba (2001) and Zardoya (2005) has currently converged on the recognition of four main clades of aquaporins in metazoans (Perez Di Giorgio et al., 2014; Abascal et al., 2014; Finn and Cerdà, 2015). Functional classifications of aquaporins are not, however, congruent. For example, although one of the clades in Metazoa includes only proteins with glycerol-transporting capacity, two of the other three clades also include members

* Corresponding author.
E-mail address: don.colgan@austmus.gov.au (D.J. Colgan).

https://doi.org/10.1016/j.margen.2017.12.002
Received 8 August 2017; Received in revised form 31 October 2017; Accepted 2 December 2017

with this capability (Perez Di Giorgio et al., 2014). Most of the four broad classes have been identified in studied invertebrate species (Finn and Cerdà, 2015). However, the two available studies including significant numbers of Mollusca differ as to whether the major group related to vertebrate aquaporins 11 and 12 ("S-aquaporins" or "unorthodox aquaporins") is absent (Abascal et al., 2014) or present (Kosicka et al., 2016) in the phylum.

Only recently have aquaporins been explicitly characterised in snails (Pieńkowska et al., 2014; Kosicka et al., 2016) although some genes are identified as aquaporins in the genomes of *Lottia gigantea* Gray in G. B. Sowerby I 1834 (Simakov et al., 2013) and *Biomphalaria glabrata* (Say 1818) (*Biomphalaria glabrata* genome initiative, http://biology.unm.edu/Biomphalaria-Genome/index.html, 2015) and the transcriptome of *Aplysia californica* J. G. Cooper 1863 (http://neurobase.rc.ufl.edu/aplysia/browse).

The evolutionary homologies of molluscan aquaporin sequences were not tested using a phylogenetic approach until the analyses of Abascal et al. (2014) and Kosicka et al., 2016). Kosicka et al. (2016) identify two new groups of aquaporins, which they name as malacoaquaporins and malacoglyceroporins but do not otherwise classify the large group of molluscan proteins that the careful biochemical investigations in this and a previous study by the group (Pieńkowska et al., 2014) have shown to be "classical" aquaporins.

The present investigation was performed to make a phylogenetically-based classification of gastropod aquaporin proteins. The analysis was based on the compilation of new information from partial genomic assemblies of three species of Neritimorpha generated in this laboratory and genomic and transcriptomic information available from GenBank and other public databases. The three neritimorph species were the marine *Nerita melanotragus* E.A. Smith 1884 (family Neritidae) and the terrestrial *Pleuropoma jana* (Cox 1872) (Helicidae) and *Georissa laseroni* (Iredale1 1937) (Hydrocenidae).

2. Materials and methods

2.1. Data compilation

Gastropod aquaporin cds sequences were downloaded from the GenBank nr/nt database. Representatives from across the topology recovered in initial phylogenetic analyses of these sequences were used in blastn searches of EST and TSA databases retaining up to 200 sequences and restricting taxon to "Gastropoda". Detected sequences with an "expect value" less than 1 were downloaded. Because of the low stringency of selection, the data were expected to include non-aquaporins. Sequences among the downloaded accessions identified as aquaporins were added to our data compilation. Sequences not so identified were checked individually by using them as test sequences in blastn searches of GenBank databases and by using their deduced translations in blastp and tblastn searches. If the highest similarity recorded for the sequences from these searches was found to be with proteins identified as aquaporins, they were added to the dataset.

Additional coding sequences were recovered from available partial or complete gastropod transcriptomes or genomes of L. *gigantea* (Simakov et al., 2013, *B. glabrata* (*Biomphalaria glabrata* genome initiative, 2015: https://www.vectorbase.org/organisms/biomphalaria-glabrata) and *A. californica* (http://neurobase.rc.ufl.edu/aplysia) by keyword searching for "aquaporin" and/or by blastn searches using known aquaporin sequences. Sequences for *Pomacea canaliculata* (Lamarck, 1822) were identified in the transcriptome data of Sun et al. (2012) by the fasty35 option of FASTA3 (Pearson, 1999) using searches with representatives of the various clades of gastropod aquaporins. This procedure was also used to identify potential aquaporin sequences in partial genome assemblies of three neritimorph species, *Nerita melanotragus, Georissa laseroni* and *Pleuropoma jana*, that were generated as described below. The deduced amino acid sequences of the sequences identified as possible aquaporins were tested against the GenBank nr database using tblastn to confirm that they belong to this class of protein. The GenBank Accession numbers for these sequences are MF380517 - MF380527 and MF510930 - MF510933. Intronic sequences in the *Nerita* contigs 12965, 73503 and 897454, and the *Georissa* contigs 33993 and 129968 were not included in the conceptual translation. The intron sequences were identified by confirming the presence of consensus 5′ and/or 3′ consensus splice sites at positions where there was an abrupt change in the similarity of deduced amino acid sequences with the overall alignment of known aquaporins.

Genes described as aquaporins in the annotation for the genome of the bivalve *Crassostrea gigas* (Thunberg 1793) (GCA_000297895.1) were obtained from the Ensembl genome browser at http://metazoa.ensembl.org/Crassostrea_gigas/Info/Index. Sequences from a number of non-molluscan species were downloaded from GenBank and added to the compilation to provide a framework for classifying the gastropod sequences. These included sequences of identified human aquaporin types and aquaporins from five species of Ecdysozoa, which is the sister group of Lophotrochozoa, the major taxon within Protostomia that includes Gastropoda (Aguinaldo et al., 1997). The ecdysozoan sequences were from the nematode *Caenorhabditis elegans* (Maupas 1900), the tardigrade *Milnesium tardigradum* Doyère 1840, the acarine *Ixodes ricinus* (Linnaeus 1758), and the dipteran insects *Drosophila melanogaster* Meigen 1830 and *Bactrocera dorsalis* (Hendel 1912). Duplicate sequences were removed after identification of identical accessions in preliminary alignments in BioEdit (Hall, 1999) and by filtering in ALTER, the Alignment Transformation Environment (Glez-Peña et al., 2010).

Inferred amino acid sequences were obtained from DNA sequences by the EMBOSS Transeq program (Rice et al., 2000; Goujon et al., 2010). The final overall protein dataset was then re-aligned with MAFFT v7 (Katoh and Toh, 2008) on the webserver at http://www.ebi.ac.uk/Tools/msa/mafft/) using a BLOSUM 62 matrix, gap penalties of 1.55 (opening) and 0.123 (extension), tree rebuilding of 2 and 2 iterations.

2.2. Partial genomic assemblies of Neritimorpha

Next generation Illumina de novo shotgun sequencing used three separate lanes of a HiSeq run, one for each species. The sequencing was performed on DNA extractions from individual snails using the CTAB procedure of Saghai-Maroof et al. (1984), modified by precipitating DNA from the liquid phase resulting from chloroform/isoamyl alcohol extraction with an equal volume of isopropanol. DNA from *N. melanotragus* and *P. jana* was used directly in the preparations of libraries for NGS sequencing at the Australian Genomics Research Facility. DNA from *G. laseroni* was amplified by the GenomiPhi V2DNA Amplification Kit (GE Health Care, Life Sciences). The amplified DNA was purified and sequenced by Macrogen (Seoul, Korea).

After quality-checking with FastQC v0.10.1 (Andrews, 2012), the raw reads were assembled into contigs with CLC Genomics Workbench (http://www.clcbio.com/products/clc-genomics-workbench/) on a local installation using a variety of parameter values. The assemblies used here were based on a word length of 20, bubble size of 30 and the "fast" algorithm and had the largest value for N50 (the length shorter than 50% of the contigs) except for *G. laseroni* for which these parameters gave the largest partial assembly. In the absence of genome size data for Neritimorpha, the range of estimates for coverage in the following is based on assumed 1C (i.e. haploid) contents of between 0.5 pg and 2.5 pg. For *N. melanotragus*, there were 6123 Mb of data provided by 61,233,027 unmatched reads giving a coverage range between 2.50 × and 12.52 ×. The partial assembly from these reads produced 920,740 contigs, with an N50 of 429 and an assembled length of 342,840,026 bases. For *P. jana*, there were 6268 Mb of data provided by 62,678,072 unmatched reads giving a coverage range between 2.56 × and 12.82 ×. The partial assembly from these reads produced 1,322,892 contigs, with an N50 of 274 and an assembled length of