RESEARCH NOTES

Identification of genes expressed in the gill tissue of the Pacific oyster (*Crassostrea gigas*) using expressed-sequence tags

Gerard P. Rafferty and Richard Powell

Department of Microbiology, National University of Ireland, Galway, Ireland

Crassostrea gigas (Thunberg, 1793), the Pacific or Japanese oyster, is an important shellfish species in terms of world aquaculture. Genetic research on *C. gigas* has focused on growth rate,[1] karyotyping[2] and polyploidy.[3,4] To complement these studies, research is also likely to include the description of individual oyster genes. A search of the GenBank DNA sequence database reveals that such information is currently limited to the description of 31 full-length *C. gigas* gene sequences, 12 of which are from by the mitochondrial genome, and approximately 250 partial DNA sequences of which the majority are non-coding and <100 bp in length. Expressed-sequence tag (EST) studies are now commonly used for the purpose of large-scale gene identification in species of interest.[5] Currently, 909 *C. virginica* ESTs, derived from haemocyte and embryo cDNA libraries are available in GenBank (Accession Numbers BG624106-BG624961). Here, we describe an EST-based gene identification study targeting the genes expressed in the gill tissue of *C. gigas*.

To construct the cDNA library, gill tissue was dissected from five adult *C. gigas* obtained from a commercial farm (Redbank Oyster Farm, New Quay, Co. Clare, Ireland). Total RNA was extracted using the TRIzol reagent (Gibco BRL Life Technologies Ltd., Renfrewshire, UK) resulting in a yield of 1.9 mg of total RNA per gram wet weight of gill tissue. Messenger RNA was enriched by oligo-(dT) cellulose chromatography using the Poly (A) Quick mRNA isolation kit (Stratagene Cloning Systems, California, USA) resulting in a yield of 5.2 μg (1%) of mRNA per 500 μg of total RNA. The integrity of both the total RNA and mRNA were checked by electrophoresis through 1% agarose gels.[6] A directionally cloned gill cDNA library was constructed using the λ Zap Express cDNA synthesis/Gigapack cloning kit (Stratagene Cloning Systems) and cDNA synthesis was carried out using an oligo-$(dT)_{18}$ primer for the reverse transcription of approximately 5 μg of mRNA. A quality analysis of the resulting cDNA library showed that it contained 5.1×10^4 primary cDNA clones with a background of <2% parental λ Zap Express plaques, after plating on LB using soft-top containing 3.2 mg/ml XGal and 1 mM IPTG. After mass *in vivo* excision of 10^6 cDNA clones from an amplified portion of the cDNA library into pBK-CMV phagemids, 20 random clones were selected to determine the lengths of the cloned cDNA fragments. The cloned cDNA fragments were released by digestion of recombinant phagemid DNA with *Pst*I and *Xba*I and analysed by electrophoresis through 1% agarose gels.[6] The results showed that the cloned DNA fragments ranged from 0.3 to 2 kb in length, and >90% of the clones had cDNA fragments >0.5 kb in length. Therefore, it was possible to proceed directly to EST determination without having to identify and remove clones containing insufficient cDNA sequence.

To determine ESTs, single-pass sequencing of the 5′ termini of 132 randomly chosen mRNA-encoding cDNA clones from the cDNA library was performed using the ABI 310 automatic DNA sequencer (PE Applied Biosystems Inc., California, USA) and the ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction kit (PE Applied Biosystems). After elimination of vector sequences, the 122 (92%) oyster ESTs >250 bp in length were analysed for multi-EST redundancy using the Clustal-X programme[7] with the subjective parameter that ESTs sharing >99% identity over 300 nucleotides were identical. Seven EST clusters were identified and longest EST was chosen as the representative member of each cluster. The sequence data of these seven ESTs and also the 97 singleton ESTs have been deposited in the GenBank dbEST[8] database under accession numbers BM986671-BM986729 and BG467397-BG467443.

Gene identification analysis was performed by submitting the 104 unique *C. gigas* ESTs to the NCBI database[9] using the Blast-X and -N programs located on the NCBI Blast homepage (http://www.ncbi.nim.nih.gov/BLAST/). The Blast-X protein homology search (based on translation of the ESTs in all six reading frames) was used due to the low number of oyster DNA sequences currently described and the greater ability of protein sequences to detect homology over long periods of evolutionary time.[10] The Blast-N nucleotide search was included to search *C. gigas* sequences in the database that may not have been identified at the protein level. For each search, the first 50 most similar sequences were listed and the best ten sequence alignments were examined. The criteria chosen for unambiguous *C. gigas* gene identification was based on a minimum sequence identity of >40% over either (a) a contiguous series of >30 amino acids for the protein search or (b) a contiguous series of >90 nucleotides for the nucleotide search. The results revealed that 47 (45%) *C. gigas* ESTs showed sufficient identity to previously described sequences from other species including eight nucleotide sequences from *C. gigas*.

Analysis of the 47 identified *C. gigas* ESTs showed that they provided unique EST markers for 47 individual oyster genes (Table 1). Only eight (17%) of these genes had previously been identified in *C. gigas*, of which seven were mitochondrial genes (GenBank Accession Number AF177226), while the other was the ubiquitously expressed actin gene. Comparative analysis with genes described from other species showed that 39 (83%) of the 47 identified *C. gigas* genes could be assigned to a functional category (Table 1). Genes encoding proteins involved in translation represented the largest functional category, i.e. 14 of the 47 identified genes, including nine ribosomal protein genes. By comparison, between two and seven genes were identified whose products were associated with the range of different cellular functions including metabolism (two ESTs), defence (two ESTs), cell structure (three ESTs), cell membrane (three ESTs), digestion (three ESTs), binding proteins (seven ESTs) and the mitochondrion (seven ESTs). The remaining eight identified *C. gigas* genes could not be classified as the functions of their best match homologues are as yet unclear. Comparative analysis with the *C. virginica* EST dataset showed that nine genes common to both species were identified (α-tubulin, actin, precerebellim-like protein, cytochrome b, elongation factor 1-α, cytochrome oxidase, cathepsin-b, actin modulator and translation initiation factor). However, the respective ESTs from both species are dissimilar as they are derived from different regions of the relevant mRNA transcripts.

With respect to the 57 unidentified *C. gigas* ESTs, 3 ESTs showed no database match, while 53 ESTs showed both amino acid and nucleotide homology values to database sequence entries, but at values below the selected criteria. Subsequently,

Correspondence: R. Powell; e-mail: richard.powell@nuigalway.ie

Table 1. Classified list of identified ESTs expressed in the gill tissue of *Crassostrea gigas*.

Pacific oyster gill ESTs		Matching sequences			
Access no.	Name	Species	Access no.	E-value	Identity
I. Cell structure					
BG467397	α-tubulin	*Eleusine indica*	022349	1e-09	100% (28/28)
BG467398	actin	*Crassostrea gigas*	017320	4e-47	100% (97/97 nt)
BG467399	β-tubulin	*Paracentrodus lividus*	P11833	7e-82	95% (138/145)
II. Gene/protein expression					
BG467400	Elongation factor 1-α	*Chromolepida pruinosa*	AF124488.1	7e-73	85% (137/161)
BG467401	Ribosomal protein S11	*Homo sapiens*	NP001006	7e-63	74% (112/150)
BG467402	Ribosomal protein S3a	*Xenopus larvis*	P02350	8e-55	88% (104/117)
BG467403	Ribosomal protein L21	*Homo sapiens*	L38826.1	1e-34	57% (68/119)
BG467404	Elongation factor 2	*Caenorhabditis elegans*	P29691	2e-66	76% (132/173)
BG467405	Ribosomal protein L23a	*Homo sapiens*	L13799.1	5e-26	74% (59/79)
BG467406	Elongation factor 1-γ	*Artemia salina*	P12261	7e-65	65% (118/180)
BG467407	Translation initiation factor	*Homo sapiens*	P56537	4e-11	72% (31/43)
BG467408	Ribosomal protein	*Homo sapiens*	L13802.1	7e-59	61% (113/183)
BG467409	Ribosomal protein S10	*Lumbricus rubellus*	AJ011705.1	8e-50	70% (98/140)
BG467410	Ribosomal protein L7	*Caenorhabditis elegans*	001802	1e-44	52% (84/161)
BG467411	Ribosomal protein L7a	*Fugu rubripes*	O57592	1e-14	41% (50/121)
BG467412	Ribosomal protein S30	*Mus musculus*	XP006522	8e-28	58% (65/112)
BG467440	Transcription factor	*Homo sapiens*	JC4020	4e-31	53% (70/132)
III. Mitochondrion					
BG467413	NADH dehydrogenase 2	*Crassostrea gigas*	AF177226	0.0	99% (385/388 nt)
BG467414	Cytochrome b	*Crassostrea gigas*	AF177226	0.0	98% (581/591 nt)
BG467415	NADH dehydrogenase 6	*Crassostrea gigas*	AF177226	1e-103	92% (203/206 nt)
BG467416	16S rRNA	*Crasssostrea gigas*	AF177226	0.0	97% (674/689 nt)
BG467417	ATPase subunit 6	*Crassostrea gigas*	AF177226	0.0	99% (585/589 nt)
BG467418	Unidentified protein	*Crasssostrea gigas*	AF177226	0.0	97% (572/584 nt)
BG467419	Cytochrome oxidase subunit I	*Crasssostrea gigas*	AF177226	0.0	99% (386/387 nt)
IV. Digestion					
BG467420	Proteosome	*Homo sapiens*	P34062	8e-63	91% (122/134)
BG467421	Cathepsin-b	*Bos taurus*	KHBOB	3e-58	54% (104/192)
BG467422	Mitochomdrial processing peptidase	*Homo sapiens*	075439	3e-47	58% (83/141)
IV. Binding proteins					
BG467423	Actin modulator	*Lumbricus terrestris*	S51363	1e-82	67% (146/215)
BG467424	X-box binding protein	*Homo sapiens*	P17861	5e-19	49% (54/109)
BG467425	Cdc 5 like protein	*Rattus norvegicus*	AF000578.2	4e-42	84% (79/94)
BG467426	Arsenite transporter	*Homo sapiens*	U60276.1	4e-39	68% (75/110)
BG467427	Type A/B HnRNP P40	*Rattus norvegicus*	AJ238854.1	2e-31	65% (67/103)
BG467428	Galectin	*Rattus norvegicus*	P38552	1e-17	40% (49/120)
BG467439	Topoisomerase chain A	*Homo sapiens*	3659921	4e-15	39% (50/127)
V. Metabolism					
BG467429	Glutamine synthetase	*Panuhrim argus*	M96798.1	3e-88	74% (140/189)
BG467430	Glycine transporter	*Rattus norvegicus*	A48716	8e-40	39% (69/175)
VI. Cell/organism defence					
BG467431	Heat shock protein 90	*Brugia pahangi*	AJ005785.1	2e-68	75% (146/195)
BG467432	Peptidoglycan recognition protein	*Mus musculus*	NM009402.1	7e-31	41% (75/179)
VII. Membrane					
BG467433	Intermediate filament A	*Aplysia californica*	S24545	3e-38	48% (85/175)
BG467434	Transmembrane receptor	*Homo sapiens*	NM003468.1	1e-12	68% (30/44)
BG467435	Leptin receptor	*Mus musculus*	O89013	2e-21	52% (49/93)
VIII. Unclassified					
BG467436	Hypothetical protein	*Homo sapiens*	NM014886.1	3e-86	76% (146/190)
BG467437	Unknown protein	*Homo sapiens*	AK001477.1	7e-20	35% (47/134)
BG467438	C1q related factor	*Mus musculus*	NM011798.1	2e-05	34% (40/115)
BG467441	Precerebellim	*Oncorhyhcus mykiss*	AF192969.1	1e-05	31% (35/111)
BG467442	Unknown protein	*Arabidopsis thaliana*	ACC6262.3	3e-34	44% (75/170)
BG467433	Nel protein	*Gallus gallus*	JP0076	2e-18	37% (46/122)

*The identity values are based on the overlap of amino acid residues except where indicated as nucleotides (nt).

examination of these E values and the relevant protein and nucleotide alignments did not provide confidence for unequivocal gene identification. The identification of the *C. gigas* genes represented by these ESTs likely requires the description of greater numbers of gene sequences from more closely related species, rather than from comparatively well studied mammals. Also, it cannot be excluded that some of the ESTs described in this study may be derived from mRNA of parasites present in the oyster gill tissue during the RNA extraction procedure.

In summary, the construction and assessment of a *C. gigas* gill tissue cDNA library revealed it to be a valuable resource for the identification of individual oyster genes. Using the EST strategy, it proved possible to assign a gene identification to 45% of the randomly selected *C. gigas* ESTs, resulting in the identification of 39 genes not previously in this species. Furthermore, the identified genes spanned the range of cellular functional categories and should be useful for monitoring gene expression under different physiological conditions, e.g. respiration, feeding, defence.

This work was supported by the EU FAIR Programme (Contract FAIR CT95 0421). G.P.R. was supported by a NUI, Galway postgraduate fellowship.

REFERENCES

1. THIRIOT-QUIÉVREUX, C., POGSON, G.H. & ZOUROS, E. 1992. *Genome*, **35**: 39–45.
2. LEITAO, A., THIRIOT-QUIÉVREUX, C., BOUDRY, P. & MALHEIRO, I. 1999. *Genet. Select. Evol.*, **31**: 519–527.
3. GUO, X. & ALLEN, S.K. 1996. *Genome*, **40**: 397–405.
4. LEITAO, A., BOUDRY, P. & THIRIOT-QUIEVREUX, C. 2001. *Aquaculture*, **193**: 39–48.
5. MARRA, M.A., HILLIER, L. & WATERSON, R.H. 1998. *Trends Genet.*, **14**: 4–7.
6. SAMBROOK, J., FRITSCH, E.F. & MANIATIS, T. 1989. In: *Molecular cloning* (C. Nolan, ed.), 6.1–6.62. Cold Spring Harbor Press.
7. THOMPSON, J.D., GIBSON, T.J., PLEWNIAK, F., JEANMOUGIN, F. & HIGGINS, D. 1997. *Nucl. Acids Res.*, **25**: 4876–4882.
8. BOGUSKI, M.S., LOWE, T.M. & TOLSTOSHEV, C.M. 1993. *Nature Genet.*, **4**: 332–333.
9. BENSON, D.A., BOGUSKI, M.S., LIPMAN, D.J., OSTELL, J. & FRANCIS OULLETE, B.F. 1998. *Nucl. Acids Res.*, **26**: 1–7.
10. PEARSON, W.R., 1997. *Comput. Appl. Biosci.*, **13**: 325–332.