Mar. Drugs **2012**, *10*, 2265-2279; doi:10.3390/md10102265

OPEN ACCESS

Marine Drugs

ISSN 1660-3397

www.mdpi.com/journal/marinedrugs

Article

Digital Marine Bioprospecting: Mining New Neurotoxin Drug Candidates from the Transcriptomes of Cold-Water Sea Anemones

Ilona Urbarova [1,†], Bård Ove Karlsen [1,†], Siri Okkenhaug [1], Ole Morten Seternes [2], Steinar D. Johansen [1,3,*] and Åse Emblem [1]

[1] RNA and Transcriptomics Group, Department of Medical Biology, Faculty of Health Sciences, University of Tromsø, N9037 Tromsø, Norway; E-Mails: ilona.urbarova@uit.no (I.U.); baard.ove.karlsen@northernbiolabs.no (B.O.K.); siri.okkenhaug@gmail.com (S.O.); ase.emblem@uit.no (A.E.)

[2] Pharmacology Group, Department of Pharmacy, Faculty of Health Sciences, University of Tromsø, N9037 Tromsø, Norway; E-Mail: ole-morten.seternes@uit.no

[3] Marine Genomics Group, Faculty of Biosciences and Aquaculture, University of Nordland, N8049 Bodø, Norway

[†] These authors contributed equally to this work.

[*] Author to whom correspondence should be addressed; E-Mail: Steinar.Johansen@uit.no; Tel.: +47-77-64-53-67; Fax: +47-77-64-53-50.

Received: 31 August 2012; in revised form: 8 October 2012 / Accepted: 10 October 2012 / Published: 18 October 2012

Abstract: Marine bioprospecting is the search for new marine bioactive compounds and large-scale screening in extracts represents the traditional approach. Here, we report an alternative complementary protocol, called digital marine bioprospecting, based on deep sequencing of transcriptomes. We sequenced the transcriptomes from the adult polyp stage of two cold-water sea anemones, *Bolocera tuediae* and *Hormathia digitata*. We generated approximately 1.1 million quality-filtered sequencing reads by 454 pyrosequencing, which were assembled into approximately 120,000 contigs and 220,000 single reads. Based on annotation and gene ontology analysis we profiled the expressed mRNA transcripts according to known biological processes. As a proof-of-concept we identified polypeptide toxins with a potential blocking activity on sodium and potassium voltage-gated channels from digital transcriptome libraries.

Keywords: deep sequencing; drug discovery; marine bioprospecting; neurotoxin; sea anemone; transcriptomics

1. Introduction

Marine bioprospecting has significant potential for the discovery of novel drugs, nutritional supplements and industrial biotechnology. The traditional approach is to extract bioactive compounds from a sample by bioassay-guided fractions and thereafter determine the structure, chemical composition and exact function [1]. *In silico* analysis and genetic discovery of marine biomolecules complement the traditional methods in bioprospecting [2–4]. By sequencing genes, genomes and transcriptomes, the search for gene homologs, motifs or transcripts with a certain expression profile can be identified.

Two years ago we reviewed the concept idea of using massive parallel deep sequencing of transcriptomes in the systematic screening for marine drug candidates [2]. Deep sequencing technologies have revolutionized the field of biology by making it achievable to sequence whole transcriptomes of non-model organisms at a relative low cost [5–7]. Three deep sequencing platforms have dominated the research of whole transcriptome analysis; the 454 pyrosequencing platform from Roche, the Genome Analyzer platform from Ilumina Sequencing technologies, and the SOLiD ligation sequencing from Life technologies [8]. These technology platforms produce raw sequence reads with a length from 50 to more than 500 nucleotides, generating billions of nucleotides in a single run.

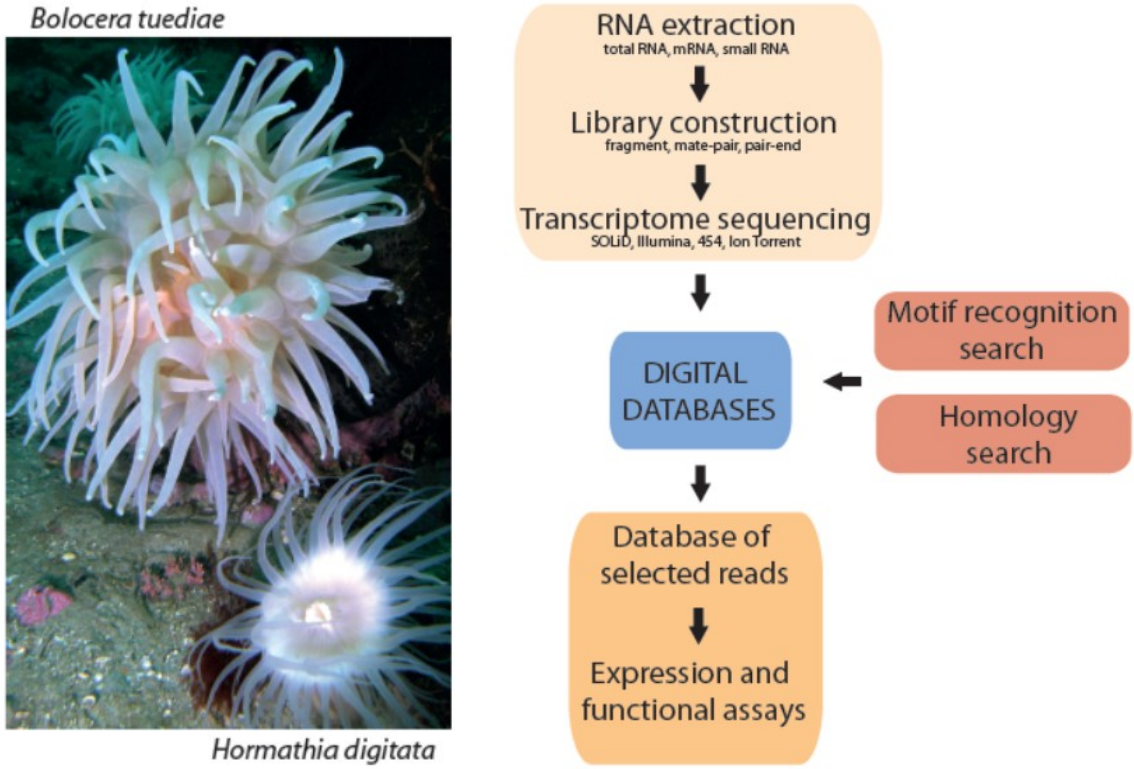
The class Anthozoa, where sea anemones and corals belong, has an interesting evolutionary position as one of the most basal eumetazoans, and recent genome analyses have revealed a gene content and structure more similar to vertebrates than earlier expected [9,10]. Corals and sea anemones are mainly sessile, and must adapt to the changing environment, catch prey, and defend themselves from predators and disease-causing agents. Thus, corals and sea anemones are promising candidates in bioprospecting of novel drug compounds [2,11,12]. Understanding coral and sea anemone biology is also essential in preserving the biodiversity that inhabit coral reefs. Despite this, only a limited number of Anthozoa transcriptomes have been sequenced. *Acropora millepora* and *A. palmate*, the main reef-builders of the Great Barrier Reef and the Caribbean reefs, respectively, are both subjected to genome and transcriptome sequencing. 454 transcriptome data from the coral larvae were already previously annotated with names and Gene Ontology (GO) terms [13,14], and applied in comparative and environmental studies. Expressed sequence tag (EST) and transcriptome projects have been initiated for selected sea anemones and corals [15,16] including the upcoming sea anemone model *Aiptasia* [17–19]. Many EST analyses are however still carried out by the use of Sanger sequencing of cloned libraries [3,20,21].

Neurotoxins are produced by a diverse group of organisms, including sea anemones [22,23]. Typically, they are relatively small peptides with conserved cysteine residues, forming disulfide bridges critical for the peptide structure [24]. Many neurotoxins are translated as inactive precursors with an *N*-terminal leader peptide sequence and a *C*-terminal mature peptide toxin. The active peptide is produced by proteolytic cleavage of a conserved dyad (Lys-Arg) [25]. Neurotoxins block cellular

processes in the nervous system and other tissues by binding to voltage-gated ion channels. In sodium channels, six neurotoxin binding sites have so far been identified [26]. These neurotoxins either block the channel pore, or modify the gating, which causes a massive release of neurotransmitters and inactivation delay. The potassium channels represent a diverse group of proteins and a variety of potassium channel toxins block these channels by different mechanisms and thereby facilitate release of the neurotransmitter acetylcholine. Potassium toxins act in synergism with other peptides such as anti-cholineesterases and sodium channel toxins [27].

We used 454 GS FLX Titanium deep sequencing to profile transcriptomes and identify expressed genes and derived gene products in the adult polyp stage of two distantly related cold-water sea anemone species, *Bolocera tuediae* and *Hormathia digitata* (Figure 1A). Here we present a protocol for digital marine bioprospecting in order to identify new peptide drug candidates derived from transcriptome sequencing libraries.

Figure 1. (**A**) The cold-water sea anemone species *B. tuediae* and *H. digitata* included in this study; (**B**) Flowchart describing the pipeline in digital bioprospecting from RNA extraction to prediction of candidate biomolecules, which can be expressed in functional trials. Photo by SDJ.



2. Results and Discussion

2.1. Transcriptome Sequencing and Assembly

Transcriptome sequencing was performed by 454 GS FLX Titanium (pyrosequencing) and resulted in 546,903 and 546,846 quality-filtered sequencing reads after adapter trimming with an average size

of 333 and 331 nt from *B. tuediae* and *H. digitata*, respectively (Table 1). From these reads, 64,442 (*Bolocera*) and 54,293 (*Hormathia*) contigs were assembled. Transcripts found in one copy number (single reads) counted for about 20% of all transcripts. The raw sequence data from *B. tuediae* and *H. digitata* transcriptomes in this study were archived at NCBI's Sequence Read Archive (SRA) under the accession number SRP011434.

Table 1. Transcriptome sequencing and assembly [a].

Species	Reads/Contigs	Number	Average size (nt)	Total nt
B. tuediae	Raw reads	547,061	547	299,232,484
	Trimmed reads	546,903	333	182,128,133
	All contigs	64,442	591	38,101,858
	Large contigs	5072	1380	6,997,895
	Single reads	118,104	279	33,008,862
H. digitata	Raw reads	546,974	543	296,833,666
	Trimmed reads	546,846	331	181,169,361
	All contigs	54,293	613	33,255,104
	Large contigs	5083	1430	7,272,471
	Single reads	105,695	260	27,786,964

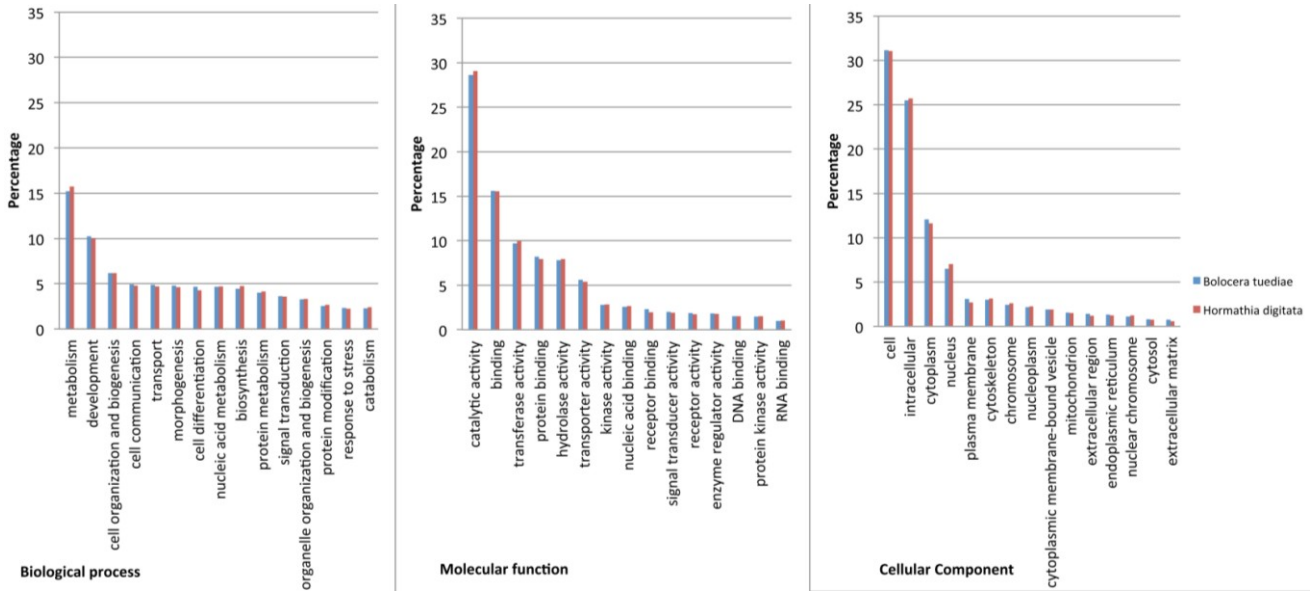
[a] Number of sequencing reads obtained from 454 pyrosequencing of the transcriptomes of the two sea anemones *B. tuediae* and *H. digitata*. Raw Reads, represent all sequence reads obtained from the transcriptome sequencing. Trimmed reads, represent raw reads after trimming of key tag (TCAG) at the 5′ end and removal of low quality and adapter sequences. All contigs, represent all contigs assembled by MWG Eurofins. Large Contigs, represent assembled contigs with size larger than 1000 bases. Single reads, represent reads that are only found in one copy number in the dataset.

2.2. Annotation and Gene Ontology Analysis

All contigs together with single reads (182,546 sequences for *Bolocera* and 159,988 for *Hormathia*) were analyzed in Blast2GO. From the 64,442 *Bolocera* contigs, 25,447 (40%) had BLAST hits to known proteins. Furthermore, 17,153 (27%) were assigned with GO terms and 11,666 (18%) were annotated. As expected, a much smaller fraction of the 118,104 single reads had BLAST hits (21,268 reads corresponding to 18%). For the 54,293 *Hormathia* contigs, 22,210 (48%) had BLAST hits, 13,787 (25%) were assigned with GO terms, and 9514 (18%) were annotated. From 105,695 single reads for *Hormathia*, 22,864 (22%) had BLAST hits. Additionally, 3674 and 4680 single reads were annotated for *Bolocera* and *Hormathia*, respectively, and assigned GO terms added to the final analyses. The GO terms assigned to the contig sequences were then exported to CateGOrizer, and GO slim analyses were run externally and graphs were produced in Microsoft Excel. GO slim terms are higher-level GO ontology categories, which provide a better profile for specie comparison [28]. The sequences were then classified according to three main GO categories, molecular function, biological process and cellular components and visualized in bar charts (Figure 2). After GO slim was performed, there were 8655 GO terms in total for *B. tuediae*, assignments to the biological process category made up the majority (5,562; 64%), followed by molecular function (2180; 25%) and cellular components (913; 11%). From 9066 GO terms for *H. digitata*, biological process category was also represented in majority (5799; 64%), compared to molecular function (2257; 25%) and cellular components (1010;

11%). *B. tuediae* and *H. digitata* belong to the distantly related families Actiniidae and Hormathiidae, respectively, of the order Actiniaria [29]. Interestingly they possess very similar transcriptome profiles in adult polyps (Figure 2). This may be explained by the identical environmental growth conditions since these sessile individuals were sampled side-by-side at 25 meters depth (Figure 1A). Additionally, both species showed a very high similarity to *Nematostella vectensis* sea anemone [9]. Here, 17,295 out of 25,447 and 18,277 out of 26,210 TOP-BLAST hits for *Bolocera* and *Hormathia*, respectively, were assigned to *N. vectensis*.

Figure 2. Gene Ontology (GO) assignment for *B. tuediae* and *H. digitata* from 454 pyrosequencing. All assembled contigs together with single reads were blasted and annotated. For the 182,546 and 159,988 contig sequences together with single reads for *B. tuediae* and *H. digitata*, respectively, 104,622 and 128,814 GO terms in total were assigned. Furthermore, 127 GO slim ancestor term were assigned to both species. Transcripts were annotated in three main categories: cellular components, molecular function and biological processes. Top 15 classes from each GO category were chosen as representatives for transcriptome comparison. A single transcript could be assigned in more than one category.



2.3. A Protocol for Digital Marine Bioprospecting

A main objective of this study was to present a workflow for digital bioprospecting. The reported protocol is based on whole transcriptome sequencing of a desired organism, recognizing sequences or motifs by bioinformatical tools, and thereafter expressing the candidate gene to perform further functional characterizations. A flowchart of the approach is presented in Figure 1B.

At least two different approaches can be pursued. First, regular BLAST homology searches can be performed at different stringencies. Stringencies must be evaluated for each query, depending on expected conservation between query sequence, database, and acceptable degree of false positives. Evaluation can be performed by inspection of reciprocal searches when applying different parameters. Sequencing data are arranged into a local database that represents a digital library, and annotated

homologs of desired molecules are then employed as the query, either at the nucleotide or amino acid sequence level. The database can be utilized as a collection of crude sequences from the original reads, or sequence assemblies can be produced, representing longer and more complete contigs. One drawback of databases with assembled contigs is the risk of producing false assemblies due to the presence of more than one gene copy or closely related homologs, as well as alternative splicing or RNA editing. This can however be largely avoided by stringent settings for contig assemblies. False assemblies can be a challenge for toxin peptides, which often are expressed together with closely related isoforms [30].

Searches in this study were performed on assembled contigs and on single reads. First, 284 sodium and 268 potassium channel toxins from different species were downloaded from SwissProt/UniProtKB protein database and used as query sequences. After closer examination of possible hits, it was concluded that candidate toxins are more similar to published sea anemone toxins and 78 annotated sea anemone neurotoxins were therefore used as query sequences for follow-up analyses (Table S1).

All possible hits to the query sequences with e-value lower than 1×10^{-6} were evaluated (after reciprocal searches to the reviewed SwissProt/UniProtKB protein database) and resulted in 15 hits, 4 for *Bolocera* and 11 for *Hormathia* (Table S2). Another level of stringency was added by allowing only sequences with e-value lower than 1×10^{-6} when blasting against the reviewed SwissProt/UniProtKB database. This search finally resulted in four hits, one for *Bolocera* and three for *Hormathia*, the alignments are shown in Figure 3A. An additional four hits to potential neurotoxin candidates were also included in this study (Table S2).

The second approach identifies potential sequences of interest in the digital library based on recognition of conserved domains. The domain architecture is sometimes the only indication to derived protein function, and domain analysis will thus increase the probability of discovering novel compounds. Recognition of conserved domains is also based on homology searches, but here multiple sequence alignment models based on experimentally verified structures make up the basis of the search tools. Most prediction pipelines are designed for single sequence analyses, and thus not suited for NGS data. Only a few studies have applied motif recognition on large-scale data set. Kozlov and co-workers [31,32] have developed a motif recognition program called Single Residue Distribution Analysis (SRDA) where predicted motifs, based on conserved amino acid sites in a certain group of proteins or peptides, are used in scanning of translated EST databases. This method was successfully applied on spider and sea anemone EST databases in the identification of potentially novel neurotoxins. Here, conserved domains were recognized using the NCBI's Conserved Domain Database (CDD) with the Batch CD-Search interface, which can process up to 100,000 sequence predictions at one time [33]. The CDD input data are amino acid sequences, and nucleotide data have to be translated into the correct reading frames prior to analysis. The complete search results are then compiled into a temporary database, which is downloaded or viewed graphically. The output domain footprint is either shown as specific hits, as domain super families, or as multi-domain models.

2.4. Identification of New Potential Neurotoxin Drug Candidates from Sea Anemones

As a proof-of-concept we applied the protocol in pursuing neurotoxin transcripts in the 454 pyrosequenced transcriptome data from the two cold-water sea anemone species *B. tuediae* and *H. digitata*. Unlike many tropical species these sea anemones are non-symbiotic, meaning they are not associated with a zooxanthellae. Although we cannot exclude co-extraction of RNA from protozoa or microbial species, we find it highly likely that transcripts investigated here originate from the sea anemone tissue. Sodium and potassium channel toxins blast searches (blastx) were performed on our local *Bolocera* and *Hormathia* transcriptome databases, applying published peptide sequences from other Actiniaria as query sequences (Table S1). Most of these toxins are small peptides, less than a hundred amino acids long. Based on the sequence similarity with published sea anemone neurotoxins, we predicted one sodium channel toxin and three potassium channel toxins, which passed our quite stringent criteria for homology prediction (Figure 3A,C). Reciprocal searches against SwissProt/UniProtKB protein database were performed to assure that the blast hits for neurotoxin genes truly represent the best matches for the sequences. One of the classification parameters of sodium channel toxins is the size of the mature toxins, although the number and positions of cysteine residues seem also to be of greater importance [34,35]. It is well known that the 3D structure of neurotoxins is essential for appropriate binding to the specific ion channel and therefore deletions in loop regions might not be vital. The predicted *Hormathia* type III sodium channel toxin (HdNa3) was aligned to the type III homolog from *Calliactis parasitica* (CLX-1). The type III sodium channel toxins are not well defined. The predicted HdNa3 has, however, sequence similarity to the CLX-1 peptide. Surprisingly, no sodium channel toxins were predicted from the *Bolocera* sequence data. This was unexpected since sodium channel toxins are abundant in other sea anemones, and because a sodium channel toxin has previously been reported from *B. tuediae* [36].

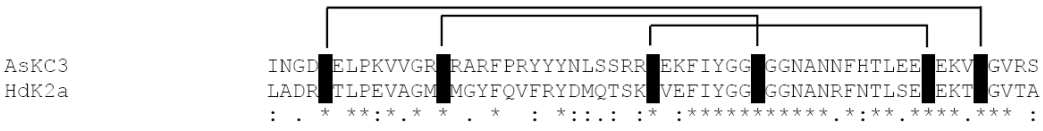
Transcripts representing type II class of potassium channel toxins were predicted from both *B. tuediae* and *H. digitata* (Figure 3A,C). Type II toxins appear to be well conserved both regarding sequence and structure prediction. HdK2a aligns well to the type II toxin from *Anemonia sulcata* (AsKC3) (Figure 3A), two additional toxins BtK2 and HdK2b align both well to the type II toxin from *Anthopleura elegantissima* (APEKTx1).

Some toxins are represented as precursors that include an *N*-terminal signal peptide (Figure 3A). Peptide cleavage is usually initiated at a cleavage tandem site (Lys-Arg) leaving a mature peptide at the *C*-terminal part [37]. With the exception of a few conserved domains, neurotoxins generally have limited sequence conservation. The predicted neurotoxin sequences were therefore structure determined by SWISS-MODEL in order to establish conserved structural motifs, and thereby support the sequence predictions. SWISS-MODEL 3D predictions resulted in four β-strands configurations of HdNa3 (Figure 3B). This is in agreement with the information stating that most sodium channel toxins seem not to possess α-helix motifs [22]. The potassium channel toxin, HdK2 (Figure 3B), was predicted by SWISS-MODEL to possess a *C*-terminal, and also small *N*-terminal α-helix, presence of helices is comparable to other type II potassium channel toxins [27].

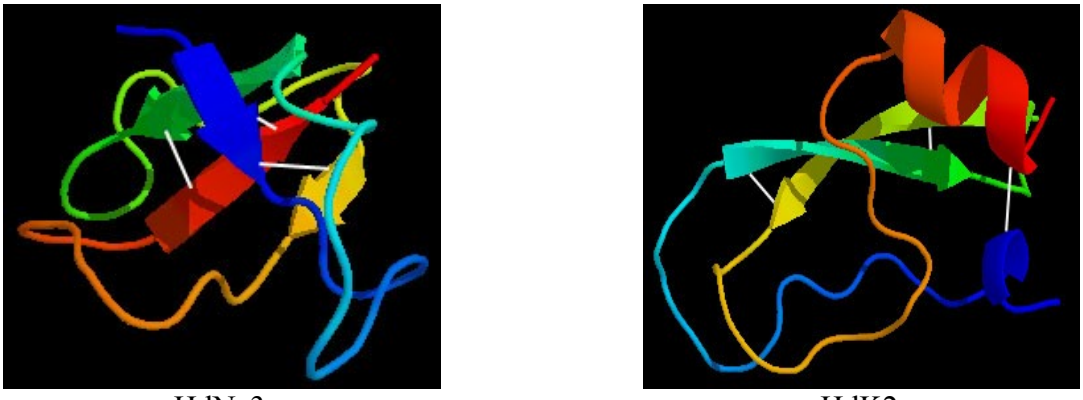
Figure 3. Representative examples of predicted neurotoxin candidates in *H. digitata* transcriptome libraries. (**A**) Recognition of one sodium channel (HdNa3) and one potassium channel (HdK2a) neurotoxin candidates from *H. digitata* based on amino acid sequence alignments. Observed cysteine residues involved in disulfide bridges are indicated. The *N*-terminal leader peptide sequences (italics) are proposed to be cleaved off at the cleavage tandem sequence (KR). (**B**) Structure predictions of the HdNa3 and HdK2a mature peptide regions. Predictions were made in SWISS-MODEL. The sodium channel neurotoxin predictions contain only β-sheets and loops, in contrast with the potassium channel neurotoxin that also contains an α-helix motif. Disulfide bridges are indicated by white lines between β-sheet motifs. (**C**) Additional two potassium channel neurotoxin candidates from group II, one predicted for *Bolocera* (BtK2) and one for *Hormathia* (HdK2b). 3D structure predictions of both of these type II potassium channel toxins are similar to HdK2a potassium channel neurotoxin from *H. digitata*. Note that star (*) below alignments in (**A**,**C**) indicates identical amino acids. Conserved amino acid changes are indicated by (: or ·).

H. digitata sodium channel type III

H. digitata potassium channel type II



A



HdNa3 HdK2

B

C

An additional 15 toxin candidates were predicted using lower stringency settings, 4 for *B. tuediae* and 11 for *H. digitata* (Table S2). These would be also included in possible experimental follow-up studies. However the digital marine bioprospecting approach introduced here serves only for bioinformatic purposes to identify possible homology variants of known proteins by mining next generation sequencing data.

Four of these additional toxin candidates showed high similarity to proteins with Kunitz-type domain (AXPI and Kunitz/BPTI-like toxin). AXPI protein was already previously shown to have sequence similarity to type II potassium channel toxins (AsKC1-3) and was proven to belong to the Kunitz-type family [38]. Proteins from this family have high sequence similarity, concerning especially the position of Cysteine residues, and they share both protease, as well as ion channel inhibitory activity. AXPI protein is also predicted to be structurally similar to these proteins, a sign also very important for neurotoxins, although its possible ion channel inhibitory activity has not been proven yet.

Potential neurotoxin sequences were also predicted by recognizing conserved domains (Table 2). We obtained 144 and 206 neurotoxin candidate transcripts from *B. tuediae* and *H. digitata*, respectively. The output result from the Batch CD-Search tool assigned 131 (*B. tuediae*) and 229 (*H. digitata*) superfamily queries with 151 and 267 positive domain hits, respectively, and a total of nine unique domains (Tables 2, S3 and S4). The most widespread conserved domain was the BPT1/Kunitz superfamily, originally a serine protease inhibitor that has gained a new function as an ion channel blocking toxin [39]. Other serine protease inhibitor domains were also recognized, together with domains characterized by disulfide bridges. A significant fraction of the transcripts was found to contain more than one conserved recognized domain.

Table 2. Conserved domain recognition.

CDD, input and output [a]	*B. tuediae*	*H. digitata*
Query amino acid sequences	864	1236
Queries with domain hits	131	229
Total number of domain hits	151	267
Superfamilies		
KU (Kunitz-type)	135	211
Toxin4	-	23
KAZAL_FS	6	23
Antistatin	6	-
WAP	1	3
TY	2	1
ShK	-	1
VMA21-like	-	1
NTR	-	1

[a] Conserved domain recognition in transcriptome data from *B. tuediae* and *H. digitata*. A neurotoxin-enriched portion of the 454 transcriptome raw reads was translated into six reading frames and ran through the NCBI's Conserved Domain Databases (CDD). Recognized superfamily domains included: KU—Kunitz type toxins (serine proteinase inhibitor); Toxin4—sea anemone neurotoxin; KAZAL_FS—serine protease inhibitor, Antistatin—serine protease inhibitor; WAP—whey acidic protein-type four-sulfide core domains; TY—thyroglobin type I; ShK—three disulfide bridges, potassium channel inhibitor; VMA21-like—two potential transmembrane helicos; NTR-like—beta barrel.

3. Experimental Section

3.1. Sampling and RNA Extraction

The cold-water sea anemones *B. tuediae* (Order Actiniaria; Family Actiniidae) and *H. digitata* (Order Actiniaria; Family Hormathiidae) were collected 2009-10-01 in Tromsø, Norway (69°41′ N; 18°56′ E) at 25 m depth using scuba diving (Figure 1A). RNA was extracted by crushing fresh tissue from body wall and tentacles in TRIzol using a Precellys lysis homogenizer (Stretton Scientific, Stretton, UK) to ensure identical sample handling before extraction [40]. 0.2× volume of chloroform was added, incubated on ice for 20 min, centrifuged and the water phase was transferred to a new tube. The RNA was precipitated in isopropanol at 4 °C and the pellet washed with 75% ethanol before the RNA was rehydrated in water. For some of the samples an additional phenol/chloroform extraction was performed, and subsequently RNA was precipitated in ethanol.

3.2. Large Scale Sequencing

Transcriptome sequencing of fragment libraries was performed by the 454 pyrosequencing platform at Eurofins MWG Operon (Germany). Approximately 10 µg total RNA from each species was shipped to Germany. Poly(A)$^+$ RNA was prepared by Eurofins MWG Operon, first strand cDNA was synthesized applying random hexamers, with successive ligation of 5′ and 3′ adaptors. PCR amplification was performed with a proof-reading enzyme. Normalization was carried out by denaturation and renaturation of the cDNA, with subsequent removal of ds-cDNA before ss-cDNA PCR amplification. The cDNA was size fractionated (500–700 bp) by elution of preparative agarose gels, subjected to shotgun library preparation and subsequent GS FLX Titanium sequencing. All the handling of the samples after RNA isolation, including cDNA library preparation, was done by Eurofins MWG Operon.

3.3. Assembly, Mapping and Annotation

The contig assemblies were performed as a service by the Eurofins MWG Operon. Quality-filtering of the reads was done by Roche/454 sequencer software when performing the base calling. The sequences were additionally trimmed, the key tag (TCAG) at the 5′ end and low quality and adapter sequences were removed from the sequences before assembly by MIRA Assembler software. Only reads ≥40 bp were considered for the assembly by MIRA Assembler. Assembled contigs and single reads were run through Blast2GO [41], they were BLASTed, mapped and annotated. The transcripts were grouped, based on their potential function and visualized in bar charts applying CateGOrizer [42] and Microsoft Excel. Additional statistical data were extracted from Blast2GO. The transcriptome data were collected in two local databases and blastx searches were performed on the contigs and single reads using relevant published anthozoan sequences as queries with a threshold value of $e = 1 \times 10^{-6}$. Furthermore, CLC Genomic Workbench [43] was applied in the mapping of toxins. Different stringencies were used during the neurotoxin searches and also additional reciprocal searches against second, more comprehensive SwissProt/UniProtKB protein database were performed to verify the candidate toxin hits.

3.4. Structure and Domain Predictions

Multiple alignments were made in ClustalW2 [44], and the mature protein length for the aligned known toxins was determined using Protein Knowledgebase (SwissProt/UniProtKB). Cysteine residues in the alignments were highlighted by black colour and the disulphide bridges were marked by lines. 3D structure predictions were performed by SWISS-MODEL [45,46] using translated amino acid sequences. The 3D structures were exported in .pdb format and visualized in PyMol Viewer [47] as a cartoon with α-helices and β-sheets with a colour transition from red to blue, *C*- to *N*-terminal. Additionally, disulphide bridges were marked according to the sequence alignments. For better resolution, figures were exported and run in POV-Ray [48].

A collection of sequences were also translated in all six reading frames applying the Six Frame Translation tool from Max-Planck Institute for Developmental Biology [49], and run through the Batch Conserved Domain Database [33] with default settings. For the purpose of this analysis, the transcriptome sequences were first enriched with potential neurotoxin transcripts by performing a low stringency ($e = 1 \times 10^{-2}$) blastx homology search applying 78 annotated neurotoxins (Table S1) as queries and these candidate sequences were translated into amino acid sequences in all six reading frames before being introduced to the Batch CD-Search tool.

4. Conclusion

Whole transcriptome profiling based on deep sequencing technologies has revolutionized the field of gene expression. In this study we report high-throughput 454 pyrosequencing to generate draft assemblies of adult polyp transcriptomes in two distantly related cold-water sea anemone species. Interestingly, the transcriptome profiles were highly similar between species. The datasets were stored as digital libraries from which desired genes and gene motifs were recognized. As a proof-of-concept we performed searches for neurotoxins by following two different approaches; homology searches and conserved domain recognition. Homology searches obtained precise hits determined by the stringency of the search, while functional domain annotation increased the chances of finding novel molecules with a certain function despite limited recognition at the nucleotide level. The fact that we identified four highly similar and 15 additional new neurotoxin peptide candidates from the *Bolocera* and *Hormathia* transcriptomes confirms the potential of digital bioprospecting. The next step in fulfilling the protocol is to include high-throughput functional analysis of candidate peptide/proteins in an appropriate experimental setting. The recent developments in array-based protein function analyses are very promising and have resulted in cell-free protein synthesis and high-density protein array platforms [50,51]. Combining these fields of biological science (bioinformatics, transcriptomics and proteomics) will create a powerful complementary approach to marine bioprospecting, which require only minute amounts of sample materials in the discovery and investigation of new protein-based drug candidates.

Acknowledgments

We thank members of our research groups at University of Tromsø and University of Nordland for support and interesting discussions. We also thank two anonymous reviewers for constructive

comments and corrections, and colleagues at Eurofins MWG Operon, Germany, for technical discussions and 454 pyrosequencing services. This work was supported by grants to SDJ from the Tromsø Research Foundation, the Research Council of Norway, and the University of Tromsø.

References

1. Koehn, F.E.; Carter, G.T. Rediscovering natural products as a source of new drugs. *Discov. Med.* **2005**, *26*, 159–164.
2. Johansen, S.D.; Emblem, A.; Karlsen, B.O.; Okkenhaug, S.; Hansen, H.; Moum, T.; Coucheron, D.H.; Seternes, O.M. Approaching marine bioprospecting in hexacorals by RNA deep sequencing. *N. Biotechnol.* **2010**, *27*, 267–275.
3. Yamaguchi, Y.; Hasegawa, Y.; Honma, T.; Nagashima, Y.; Shiomi, K. Screening and cDNA cloning of Kv1 potassium channel toxins in sea anemones. *Mar. Drugs* **2010**, *8*, 2893–2905.
4. Sperstad, S.V.; Haug, T.; Blencke, H.M.; Styrvold, O.B.; Li, C.; Stensvåg, K. Antimicrobial peptides from marine invertebrates: Challenges and perspectives in marine antimicrobial peptide discovery. *Biotechnol. Adv.* **2011**, *5*, 519–530.
5. Vera, J.C.; Wheat, C.W.; Fescemyer, H.W.; Frilander, M.J.; Crawford, D.L.; Hanski, I.; Marden, J.H. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* **2008**, *7*, 1636–1647.
6. Parchman, T.L.; Geist, K.S.; Grahnen, J.A.; Benkman, C.W.; Buerkle, C.A. Transcriptome sequencing in an ecologically important tree species: Assembly, annotation, and marker discovery. *BMC Genomics* **2010**, *11*, 180.
7. Wang, Y.; Zeng, X.; Iyer, N.J.; Bryant, D.W.; Mockler, T.C.; Mahalingam, R. Exploring the switchgrass transcriptome using second-generation sequencing technology. *PLoS One* **2012**, *7*, 3.
8. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46.
9. Putnam, N.H.; Srivastava, M.; Hellsten, U.; Dirks, B.; Chapman, J.; Salamov, A.; Terry, A.; Shapiro, H.; Lindquist, E.; Kapitonov, V.V.; *et al*. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* **2007**, *317*, 86–94.
10. Shinzato, C.; Shoguchi, E.; Kawashima, T.; Hamada, M.; Hisata, K.; Tanaka, M.; Fujie, M.; Fujiwara, M.; Koyanagi, R.; Ikuta, T.; *et al*. Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **2011**, *476*, 320–323.
11. Molinski, T.F.; Dalisay, D.S.; Lievens, S.L.; Saludes, J.P. Drug development from marine natural products. *Nat. Rev. Drug. Discov*. **2009**, *8*, 69–85.
12. Rocha, J.; Peixe, L.; Gomes, N.C.M.; Calado, R. Cnidarians as a source of new marine bioactive compounds—an overview of the last decade and future steps for bioprospecting. *Mar. Drugs* **2011**, *9*, 1860–1886.
13. Meyer, E.; Aglyamova, G.V.; Wang, S.; Buchanan-Carter, J.; Abrego, D.; Colbourne, J.K.; Willis, B.L.; Matz, M.V. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **2009**, *10*, 219.
14. Polato, N.R.; Vera, J.C.; Baums, I.B. Gene discovery in the threatened elkhorn coral: 454 sequencing of the *Acropora palmata* transcriptome. *PLoS One* **2011**, *6*, 12.

15. Traylor-Knowles, N.; Granger, B.R.; Lubinski, T.J.; Parikh, J.R.; Garamszegi, S.; Xia, Y.; Marto, J.A.; Kaufman, L.; Finnerty, J.R. Production of a reference transcriptome and transcriptomic database (PocilloporaBase) for the cauliflower coral, *Pocillopora damicornis*. *BMC Genomics* **2011**, *12*, 585.
16. Yuyama, I.; Watanabe, T.; Takei, Y. Profiling differential gene expression of symbiotic and aposymbiotic corals using a high coverage gene expression profiling (HiCEP) analysis. *Mar. Biotechnol.* **2011**, *1*, 32–40.
17. Weis, V.M.; Davy, S.K.; Hoegh-Guldberg, O.; Rodriguez-Lanetty, M.; Pringle, J.R. Cell biology in model systems as the key to understanding corals. *Trends Ecol. Evol.* **2008**, *7*, 369–376.
18. Sunagawa, S.; Wilson, E.C.; Thaler, M.; Smith, M.L.; Caruso, C.; Pringle, J.R.; Weis, V.M.; Medina, M.; Schwarz, J.A. Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone *Aiptasia pallida* and its dinoflagellate endosymbiont. *BMC Genomics* **2009**, *10*, 258.
19. Morgan, M.B.; Parker, C.C.; Robinson, J.W.; Pierce, E.M. Using representational difference analysis to detect changes in transcript expression of *Aiptasia* genes after laboratory exposure to lindane. *Aquat. Toxicol.* **2012**, *110–111*, 66–73.
20. Schwarz, J.A.; Brokstein, P.B.; Voolstra, C.; Terry, A.Y.; Manohar, C.F.; Miller, D.J.; Szmant, A.M.; Coffroth, M.A.; Medina, M. Coral life history and symbiosis: Functional genomic resources for two reef building Caribbean corals, *Acropora palmata* and *Montastraea faveolata*. *BMC Genomics* **2008**, *9*, 97.
21. Sabourault, C.; Ganot, P.; Deleury, E.; Allemand, D.; Furla, P. Comprehensive EST analysis of the symbiotic sea anemone, *Anemonia viridis*. *BMC Genomics* **2009**, *10*, 333.
22. Norton, R.S. Structure and structure-function relationships of sea anemone proteins that interact with the sodium channel. *Toxicon* **1991**, *29*, 1051–1084.
23. *Handbook of Neurotoxicology*; Massaro, E.J., Ed.; Humana Press: Totowa, NJ, USA, 2002; Volume I, p. 685.
24. Wunderer, G.; Fritz, H.; Wachter, E.; Machleidt, W. Amino-acid sequence of a coelenterate toxin: Toxin II from *Anemonia sulcata*. *Eur. J. Biochem.* **1976**, *1*, 193–198.
25. Anderluh, G.; Podlesek, Z.; Macek, P. A common motif in proparts of Cnidarian toxins and nematocyst collagens and its putative role. *Biochim. Biophys. Acta* **2000**, *1476*, 372–376.
26. Stevens, M.; Peigneur, S.; Tytgat, J. Neurotoxins and their binding areas on voltage-gated sodium channels. *Front. Pharmacol.* **2011**, *2*, 71.
27. Diochot, S.; Lazdunski, M. Sea Anemone Toxins Affecting Potassium Channels. *Prog. Mol. Subcell. Biol.* **2009**, *46*, 99–122.
28. Lomax, J. Get ready to GO! A biologist's guide to the gene ontology. *Brief. Bioinform.* **2005**, *6*, 298–304.
29. Emblem, Å. Genomic Analyses of the Cold-Water Coral *Lophelia* and Sea Anemones. PhD Thesis, University of Tromsø, Norway, 2011.
30. St Pierre, L.; Fischer, H.; Adams, D.J.; Schenning, M.; Lavidis, N.; de Jersey, J.; Masci, P.P.; Lavin, M.F. Distinct activities of novel neurotoxins from Australian venomous snakes for nicotinic acetylcholine receptors. *Cell. Mol. Life Sci.* **2007**, *21*, 2829–2840.

31. Kozlov, S.; Malyavka, A.; McCutchen, B.; Lu, A.; Schepers, E.; Herrmann, R.; Grishin, E. A novel strategy for the identification of toxinlike structures in spider venom. *Proteins* **2005**, *59*, 131–140.
32. Kozlov, S.; Grishin, E. The mining of toxin-like polypeptides from EST database by single residue distribution analysis. *BMC Genomics* **2011**, *12*, 88.
33. Marchler-Bauer, A.; Anderson, J.B.; Chitsaz, F.; Derbyshire, M.K.; DeWeese-Scott, C.; Fong, J.H.; Geer, L.Y.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; *et al*. CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.* **2009**, *39*, 225–229.
34. Wanke, E.; Zaharenko, A.J.; Redaelli, E.; Schiavon, E. Actions of sea anemone type 1 neurotoxins on voltage-gated sodium channel isoforms. *Toxicon* **2009**, *54*, 1102–1111.
35. Moran, Y.; Gordon, D.; Gurevitz, M. Sea anemone toxins affecting voltage-gated sodium channels—molecular and evolutionary features. *Toxicon* **2009**, *54*, 1089–1101.
36. Beress, L.; Zwick, J. Purification of two crab-paralyzing polypeptides from the sea anemone *Bolocera tuediae*. *Mar. Chem.* **1980**, *8*, 333–338.
37. Dauplais, M.; Lecoq, A.; Song, J.; Cotton, J.; Jamin, N.; Gilquin, B.; Roumestand, C.; Vita, C.; de Medeiros, C.L.; Rowan, E.G. On the convergent evolution of animal toxins. Conservation of a diad of functional residues in potassium channel-blocking toxins with unrelated structures. *J. Biol. Chem.* **1997**, *272*, 4302–4309.
38. Minagawa, S.; Ishida, M.; Shimakura, K.; Nagashima, Y.; Shiomi, K. Isolation and amino acid sequences of two Kunitz-typeprotease inhibitors from the sea anemone *Anthopleura* aff. *xanthogrammica*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **1997**, *118*, 381–386.
39. Strydom, D.J. Protease inhibitors as snake venom toxins. *Nat. New. Biol.* **1973**, *243*, 88–89.
40. Verollet, R. A major step towards efficient sample preparation with bead-beating. *Biotechniques* **2008**, *44*, 832–833.
41. Conesa, A.; Götz, S.; Garcia-Gomez, J.M.; Terol, J.; Talon, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676.
42. Hu, Z.; Bao, J.; Reecy, J.M. CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories. *Online J. Bioinform*. **2008**, *9*, 108–112.
43. CLCbio. Available online: http://www.clcbio.com/ (accessed on 15 August 2012).
44. Clustal: Multiple Sequence Alignment. Available online: http://www.clustal.org/ (accessed on 15 August 2012).
45. Arnold, K.; Bordoli, L.; Kopp, J.; Schwede, T. The SWISS-MODEL Workspace: A web-based environment for protein structure modelling. *Bioinformatics* **2006**, *22*, 195–201.
46. SWISS-MODEL: Swiss Institute of Bioinformatics. Available online: http://swissmodel.expasy.org/ (accessed on 15 August 2012).
47. PyMOL: A User-Sponsored Molecular Visualization System on an Open-Source Foundation. Available online: http://www.pymol.org/ (accessed on 15 August 2012).
48. POVRAY—Persistence of Vision. Available online: http://www.povray.org/ (accessed on 15 August 2012).
49. Bioinformatics Toolkit. Available online: http://toolkit.tuebingen.mpg.de/sixframe (accessed on 15 August 2012).

50. Hartley, J.L.; Salehi-Ashtiani, K.; Hill, D.E. Proteome expression moves *in vitro*: Resources and tools for harnessing the human proteome. *Nat. Methods* **2008**, *5*, 1001–1002.
51. Goshima, N.; Kawamura, Y.; Fukumoto, A.; Miura, A.; Honma, R.; Satoh, R.; Wakamatsu, A.; Yamamoto, J.; Kimura, K.; Nishikawa, T.; *et al*. Human protein factory for converting the transcriptome into *in vitro*-expressed proteome. *Nat. Methods* **2008**, *5*, 1011–1017.