A NGS approach to the encrusting Mediterranean sponge *Crella elegans* (Porifera, Demospongiae, Poecilosclerida): transcriptome sequencing, characterization and overview of the gene expression along three life cycle stages

A. R. PÉREZ-PORRO,*† D. NAVARRO-GÓMEZ,† M. J. URIZ* and G. GIRIBET†

*Center for Advanced Studies of Blanes (CEAB-CSIC), c/Accés a la Cala St. Francesc 14, Girona, Blanes 17300, Spain, †Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA

Abstract

Sponges can be dominant organisms in many marine and freshwater habitats where they play essential ecological roles. They also represent a key group to address important questions in early metazoan evolution. Recent approaches for improving knowledge on sponge biological and ecological functions as well as on animal evolution have focused on the genetic toolkits involved in ecological responses to environmental changes (biotic and abiotic), development and reproduction. These approaches are possible thanks to newly available, massive sequencing technologies—such as the Illumina platform, which facilitate genome and transcriptome sequencing in a cost-effective manner. Here we present the first NGS (next-generation sequencing) approach to understanding the life cycle of an encrusting marine sponge. For this we sequenced libraries of three different life cycle stages of the Mediterranean sponge *Crella elegans* and generated *de novo* transcriptome assemblies. Three assemblies were based on sponge tissue of a particular life cycle stage, including non-reproductive tissue, tissue with sperm cysts and tissue with larvae. The fourth assembly pooled the data from all three stages. By aggregating data from all the different life cycle stages we obtained a higher total number of contigs, contigs with BLAST hit and annotated contigs than from one stage-based assemblies. In that multi-stage assembly we obtained a larger number of the developmental regulatory genes known for metazoans than in any other assembly. We also advance the differential expression of selected genes in the three life cycle stages to explore the potential of RNA-seq for improving knowledge on functional processes along the sponge life cycle.

Keywords: *Crella elegans*, *de novo* assembly, life cycle stages, Mediterranean sponge, poecilosclerida, transcriptome characterization

Received 8 August 2012; revision received 16 January 2013; accepted 18 January 2013

Introduction

Sponges are an important component of benthic aquatic communities. They are widely distributed in all oceans and freshwater bodies at all latitudes from shallow waters to the deep sea (Hooper & Van Soest 2002). They play a key role in our understanding of early metazoan evolution (Giribet *et al.* 2007; Philippe *et al.* 2009) and have an ever-increasing biotechnological potential (Hunt & Vincent 2006; Fusetani 2010; Çelik *et al.* 2011). However, despite their interest, fundamental aspects of their

basic biology remain poorly investigated (Wörheide *et al.* 2005). Sponge genomics and transcriptomics have the potential to address these lacunae and are rapidly expanding fields of research, although until now, only the genome of a single sponge, *Amphimedon queenslandica*, has been available (Gauthier *et al.* 2010), and gene sequences and expressed sequence tag for few other sponges have been published (Adell & Müller 2004; Adell *et al.* 2007; Harcet *et al.* 2010; Holstien *et al.* 2010).

Progress in sequencing technologies for 'next-generation sequencing' (NGS)—such as Roche 454 (Rothberg & Leamon 2008), Solexa Illumina (Kircher *et al.* 2011) or SoliD™ (Tariq *et al.* 2011), among others—now offers the possibility of generating cost-effective

Correspondence: Alicia R. Pérez-Porro, Fax: +1 617-495-5667; E-mail: arodriguezperzporro@g.harvard.edu

genome-level sequence data. NGS has furthermore been widely used to generate transcriptomic data. The latter has a plethora of uses, including primer development, evaluation of alternative splice variants, targeted sequencing assays, gene regulation and DNA-protein interaction studies and gene expression profiling (Collins *et al.* 2008; Meyer *et al.* 2009; Ekblom & Galindo 2011; Ewen-Campen *et al.* 2011; Feldmeyer *et al.* 2011). The increasing use of NGS to generate transcriptomes for non-model organisms is also tightly correlated to the development of new bioinformatic tools for *de novo* assembly and analysis in the absence of a reference genome (Surget-Groba & Montoya-Burgos 2010; Kircher *et al.* 2011).

These advances in sequencing technologies coupled with a broad interest in the evolution of development (Adamska *et al.* 2011) have piqued scientific and commercial interest in sponges. Despite this interest, the phylogenetic position of sponges remains contentious (Dunn *et al.* 2008; Hejnal *et al.* 2009; Philippe *et al.* 2009; Sperling *et al.* 2009). Defining the complete genetic toolkit of sponges should thus further contribute towards confidently placing sponges phylogenetically (author's work in progress).

This study pursues to obtain and characterize the transcriptome of the encrusting Mediterranean demosponge *Crella elegans*. We further provide a first approach to the developmental genetic toolkit and gene expression of *C. rella elegans* along three stages of its life cycle. This species has emerged as a model encrusting sponge. Encrusting sponges experience strong spatial competition with fast growing neighbours (e.g. algae, briozoans, colonial ascidians) and competition-associated stress. Furthermore, detailed anatomical and reproductive data are available for this species (Pérez-Porro *et al.* 2012). With the aims of characterizing the developmental transcriptome and to explore differential gene expression through development, we collected samples of *C. elegans* at different stages of its reproductive cycle, obtained high-quality mRNA, and sequenced short reads from three cDNA libraries with an Illumina platform for their posterior *de novo* assembly. By comparing the three *de novo* assemblies of the different life cycle stages we were able to obtain a first overview of the gene expression along the life cycle of *C. elegans*. Obviously, by pooling data from the different cDNA libraries we were able to improve the *de novo* assembly (in number and length of contigs), resulting in a more accurate characterization of the *C. elegans* transcriptome. We thus used this combined assembly as a reference transcriptome to test the depth and reliability of our sequencing strategy by examining a list of genes involved in the developmental toolkit of metazoans. The new transcriptomic data will be further used as a molecular resource for future studies in sponge

development, physiology, phylogenomics and molecular biology. Future efforts will focus on a more in depth study of the differential gene expression through the life cycle of *C. elegans*.

Materials and methods

Species studied, sample collection and treatment

Crella elegans (Schmidt, 1862) is an Atlanto-Mediterranean, thick encrusting demosponge particularly abundant in the rocky-bottom assemblages of the western Mediterranean. Multiple individuals from the same population were collected via SCUBA diving during three distinct life cycle stages: non-reproductive (March 2010), beginning of the reproductive season with spermatogenic cysts (April 2010), and end of the reproductive season with larvae (October 2009) (Pérez-Porro *et al.* 2012). We refer to these three stages as samples 1, 2 and 3 respectively. Single individuals were used for RNA extractions.

To prevent contamination by other organisms living inside or on the sponge surface, the sampled tissue was cleaned under a stereomicroscope. Following the sponge-specific protocols of Riesgo *et al.* (2011), fragments of clean tissue from 0.25 cm to 0.5 cm in thickness and ~80 mg in wet weight were immediately excised and fixed in RNAlater® (Invitrogen) to avoid RNA degradation.

RNA extraction, quantity and quality control of mRNA

Total mRNA was extracted using the Dynabeads® mRNA DIRECT™ Kit (Invitrogen) following manufacturer's instructions, with minor modifications (Riesgo *et al.* 2011).

Quantity and quality (purity and integrity) of mRNA were assessed by two methods. First, the absorbance at different wavelengths was measured with a NanoDrop ND/1000 UV spectrophotometer (Thermo Fisher Scientific). The ratios of absorbance at 260/280 nm and 260/230 nm were used to assess RNA purity. A 260/230 ratio was used to estimate the presence of contaminants such as salts, carbohydrates, or peptides, among others, while an A260/280 ratio was used to estimate the purity of mRNA. Both ratios should show values close to 2.0–2.2 for pure mRNA. Second, capillary electrophoresis in an RNA Pico 6000 chip was performed using an Agilent BIOANALYZER 2100 System with the 'mRNA pico Series II' assay (Agilent Technologies). Integrity of mRNA was estimated by the electropherogram profile and lack of rRNA contamination (based on rRNA peaks for 18S and 28S rRNAs given by the BIOANALYZER software) (Riesgo *et al.* 2011).

RNA sequencing

RNA Sequencing was carried out using a GENOME ANALYZER II (GAII) or HiSeq 2000 (Illumina, Inc.) at the FAS Center for Systems Biology at Harvard University. Three cDNA libraries were built from mRNA concentrations of 14.7 ng/ μ L, 24.84 ng/ μ L, and 12.4 ng/ μ L (Samples 1, 2 and 3 respectively). An initial volume of 10.5 μ L of mRNA from samples 1 and 3 was used for random primed first-strand synthesis using Superscript II (Invitrogen), followed by second-strand synthesis with DNA Polymerase I and enzymatic fragmentation using the NEBNext[®] dsDNA Fragmentase (New England BioLabs). End repair of the double stranded cDNA (ds cDNA) was performed with NEBNext[®] End Repair Module (New England BioLabs) and an additional dAMP was incorporated with the NEBNext[®] dA-Tailing Module (New England BioLabs). Homemade adapters were ligated to the ds cDNA fragments with the NEBNext[®] Quick Ligation Module (New England BioLabs). Size-selected cDNA fragments of around 300 bp excised from a 2% agarose gel were amplified using Illumina PCR Primers for Paired-End reads (Illumina, Inc.) and 18 PCR cycles (98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s) with an initial step of 98 °C for 30 s followed by an extension step of 5 min at 72 °C. Fragments resulting from PCR amplification were sequenced on the GAII with a read length of 77 bp (sample 1), or 101 bp (sample 3).

Starting from an initial volume of 5 μ L of mRNA for sample 2 we used the TruSeq RNA Sample Prep Kit (Illumina, Inc.) to prepare the cDNA library following the manufacturer's instructions. Fragmentation of mRNA was done for 1.5 min, with 350 bp fragments targeted. Fragments resulting from PCR amplification were sequenced on the HiSeq 2000 with a read length of 101 bp.

No normalization of the cDNA libraries was conducted to allow for future use in studies of differential gene expression (Bogdanova *et al.* 2010). The concentration of the cDNA libraries was measured with the Qubit[®] dsDNA High Sensitivity (HS) Assay Kit and the Qubit[®] Fluorometer (Invitrogen). The quality of the library was checked by an 'HS DNA assay' in a DNA chip for Agilent Bioanalyzer 2100 (Agilent Technologies). Library concentrations were brought to 10 nM to run on the sequencing platforms.

Sequence assembly

Trimming (discarding of poor-quality terminal bases), thinning (trimming based on limit quality scores using the Phred scale) and adapter removal were performed with the software package CLC Genomics Workbench 4.6.1 (CLC bio). Quality of reads before and after

trimming and thinning with different parameters (with or without trimming and with two different thinning limits: 0.0496 and 0.05) was checked with the free access software FASTQC (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>). These results helped us decide which parameters should be used for trimming and thinning for each sample. Global assemblies were done also using CLC Genomics Workbench with the following parameters for *de novo* assembly: Mismatch cost 2; Limit 8; Insertion cost 3; Deletion cost 3; Length fraction 0.5; Similarity 0.8; Minimum distance 180; Maximum distance 250.

Sequence annotation

After assembly, only contigs above 300 bp were blasted against nr databases (NCBI: Metazoa + Fungi and Porifera) using BLASTX. All BLAST searches were carried out using BLAST 2.2.23 + with an e-value cut-off of 1e-5. Contigs with a positive match were annotated and associated to a Gene Ontology (GO) term using the free access software BLAST2GO (Conesa *et al.* 2005).

An 'ortholog hit ratio' (OHR) was calculated following Ewen-Campen *et al.* (2011). The searches of homologue sequences to general metabolic genes consisted of simple text searches based on standard general names or synonyms using the BLAST2GO search tool. Every probable homologue (i.e. contig) was re-blasted and main domains were checked with SMART to assure homology. Redundancy of the proteins in each assembly was calculated with the same script as in Riesgo *et al.* (2012).

To check that the libraries from the three different life cycle stages were comparable (i.e. the number of BLAST hits and annotations were similar), even though sample 2 produced more than double of raw reads (it was sequenced with HiSeq instead of GAII) than samples 1 and 3, the initial FASTA files containing the original reads were collapsed using the FASTX toolkit in Galaxy (Blankenberg *et al.* 2010). To randomly select 5 000 000 reads from each individual assembly we used a Python script (available by contacting the corresponding author). The *de novo* assembly of each sub-sample was also executed with CLC Genomics Workbench 4.6.1 under the parameters discussed above. BLAST analysis and annotations were performed against the same databases and using the same software packages as for the initial assemblies.

Gene expression analysis

To compare gene expression in the three life cycle stages we performed an RNA-Seq analysis for each assembly of samples 1, 2 and 3 using CLC Genomics Workbench 4.6.1. As reference we used the contigs from the assembly obtained by pooling reads of samples 1, 2 and 3 (only

contigs larger than 300 bp and with an average read coverage above 10, and with a positive BLAST hit against the Porifera nr NCBI database). Default mapping settings parameters were used: Minimum length fraction 0.9; Minimum similarity fraction 0.8; Maximum number of hits for a read 10. Reads Per Kilobase of exon model per million mapped reads (RPKM) (as defined by Mortazavi *et al.* 2008) were chosen as an expression value, with the requirement that because we were using as a reference a contig list without annotations and based on cDNA, all sequences were treated as if they had no introns. Only the unique gene reads were considered in the analysis.

Results and discussion

Illumina sequencing, sequence analysis and assembly

A cDNA library was prepared from each of three individual tissue samples of *C. elegans* (See Materials and methods): sample 1 (non-reproductive tissue), 2 (tissue with spermatid cysts) and 3 (tissue with larvae). The cDNA libraries were built and sequenced accordingly with the availability of chemistries, kits, optimized protocols and sequencing platforms at the Bauer Center for Genomics Research: samples 1 and 3 were sequenced using the Illumina GAII platform, each in one lane, while sample 2, processed 6 months later, was sequenced using the newer Illumina HiSeq platform multiplexed with two other samples. The three independent runs produced 50 590 348 reads with an average length of 77 bp; 69 643 680 reads with an average length of 101 bp; and 26 513 534 reads with an average length of 101 bp (Table 1) respectively. We then proceeded to remove the adapter sequences for all samples and trim the last 8 bp at the 3' end just for sample 2 (to increase the quality of the reads in sample 2, according to the FastQC results), thinning (limit = 0.0496 for sample 3 and 0.05 for samples 1 and 2) and filter on length for all samples (discarding all the reads below 20 bp). This left us with 31 295 458 reads with an average length of 60.7 bp for sample 1 (61.86% of the reads); 66 863 694 reads with an average length of 100.4 bp for sample 2 (96.00%); and 25 951 906 reads with an average length of 93.1 bp for

sample 3 (97.88%) (Table 1). As expected, the library run in the HiSeq platform generated more raw reads, trimmed, thinned and size-selected reads and with a longer average length than the two libraries run in the GAII platform (Table 1).

Four different *de novo* assemblies were conducted using the CLC Genomics Workbench 4.6.1 (CLC bio) by using the sequences from the independent single life cycle stages or by combining the three stages. Assembly A was built only with sample 1 (non-reproductive tissue) with 46.36% of total reads (with an average length of 75.37 bp) assembled into contigs (Table 2). Contigs ranged from 70 to 4323 bp in size (Mean contig length 393 ± 273.77) (Table 3); with a N50 of 419 bp (i.e. 50% of the assembled bases were incorporated into contigs 419 bp or longer) (Table 2). An overview of the size distribution on these contigs is shown in Fig. 1.

Assembly B was built with sample 2 (tissue with spermatid cysts) with 86.36% of total reads (with an average length of 87.4 bp) assembled into contigs (Table 2). Contigs ranged from 65 to 40 831 bp in size (Mean contig length 483 ± 462.44) (Table 3, Fig. 1) and N50 of 579 bp (Table 2).

Assembly C was built with sample 3 (tissue with larvae) with 63.44% of total reads (with an average length of 93.65 bp) assembled into contigs (Table 2). Contigs ranged from 94 to 4637 bp in size (Mean contig length 372 ± 261) (Table 3, Fig. 1) and N50 of 390 bp (Table 2).

Finally, Assembly D was built by combining the reads from samples 1, 2 and 3. 78.14% of the total reads (with an average length of 80.28 bp) assembled into contigs (Table 2). Contigs ranged from 90 to 31 766 bp in size (Mean contig length 453 ± 413.67) (Table 3, Fig. 1) and N50 of 517 bp (Table 2).

We determined the optimality of each *de novo* assembly by verifying various previously established parameters (Riesgo *et al.* 2012). Although the number of raw reads (Table 2) used for building the assemblies is not entirely comparable—due to the different Illumina platforms used to sequence the three samples—it should be highlighted that aggregating the reads of the three reproductive stages clearly yielded the highest number of reads as a starting material for assembly D (all stages)

Table 1 Sources of *Crella elegans* sequence reads

Tissue		Initial mRNA concentration (ng/ μ L)	Illumina platform	Raw reads	Avg. Length (bp)	Trimmed, thinned and size-selected reads	Bases (Mb) after trim	Avg. Length (bp) after trim	% trimmed
Sample 1	Non-reproductive	14.7	GAII	50 590 348	77	31 295 458	1898	60.7	61.86
Sample 2	With spermatid cysts	24.84	HiSeq 2000	69 643 680	101	66 863 694	5445	100.4	96
Sample 3	With larvae	12.4	GAII	26 513 534	101	25 951 906	2416	93.1	97.88
Total						98 159 152		88.88	

Table 2 *Crella elegans* transcriptome assembly statistics

	Assembly A (non-reproductive)				Assembly B (spermatic cysts)				Assembly C (larvae)				Assembly D (all stages)			
	Sequences	%	Average length (bp)	Bases (Mb)	Sequences	%	Average length (bp)	Bases (Mb)	Sequences	%	Average length (bp)	Bases (Mb)	Sequences	%	Average length (bp)	Bases (Mb)
Reads	31 295 458		60.65	1898	67 634 319		86.39	5843	25 951 906		93.65	2416	124 881 683		81.34	10 158
Matched*	14 507 616	46.36	75.37	1093	58 411 086	86.36	87.40	5105	16 464 495	63.44	93.65	1541	97 588 686	78.14	80.28	7834
Contigs	65 396		393.00	25	171 635		483.00	83	71 524		372.00	26	203 078		453.00	91
Not matched*	16 787 842	53.64	47.94	804	9 223 233	13.64	79.99	737	9 487 411	36.50	92.23	874	27 292 997	21.85	85.13	2323
N50	419				579				390				517			

*Referred to the number of reads that assembled or not into contigs.

(Table 2). Assemblies A (non-reproductive) and C (larvae) both had low and roughly equal quantities of contigs while assemblies B (spermatic cysts) and D (all stages) contained more than double this amount—with assembly D having the largest number of contigs (Table 3). In all four assemblies, the number of contigs was inside the expected range (Angeloni *et al.* 2011; Feldmeyer *et al.* 2011; Iorizzo *et al.* 2011). Similar results were observed with the N50 for each assembly except that the value for assembly B was slightly higher than that of assembly D (Table 2).

Contig average coverage (defined as the average number of reads included to create a contig and averaged per nucleotide) for all assemblies ranged from 60.46 ± 1152.27 (lowest value corresponding to assembly A) to 98.27 ± 2891.55 (highest value corresponding to assembly D) (Supporting information Table S1) with all values within the range of comparable *de novo* transcriptomes generated with Illumina platforms (Angeloni *et al.* 2011; Riesgo *et al.* 2012). As expected for a randomly fragmented transcriptome (Meyer *et al.* 2009; O'Neil *et al.* 2010; Parchman *et al.* 2010), there is a positive relationship between the length of a given contig and the number of reads it contains (Supporting information Fig. S2).

Transcriptome blasting and annotation

We blasted our four transcriptomes against two different selections of the NCBI nr database, Metazoa + Fungi and Porifera, to identify not only the number of metazoan known genes present (i.e. positive BLAST hits), but also how many of these were specifically known in sponges. BLAST against the Metazoa + Fungi nr NCBI database resulted in more than 15% of contigs >300 bp with a positive BLAST in our assemblies (Supporting information Table S3)—a value similar to those of other studies (Meyer *et al.* 2009; Parchman *et al.* 2010; Angeloni *et al.* 2011; Riesgo *et al.* 2012). This value was higher for longer contigs (i.e. increasing the length of assembled sequences makes them more likely to be identified by BLAST (Ewen-Campen *et al.* 2011; Parchman *et al.* 2010), resulting in 81.08% of the contigs >6000 bp with BLAST hits for assembly D (Supporting information Table S3). The number of contigs with a positive BLAST hit ranged from 12 441 in assembly C to 20 459 in assembly B (Fig. 2). When using the Porifera nr NCBI database the number of positive BLAST hits for each transcriptome increased slightly, compared to the Metazoa + Fungi BLAST, and resulted in a positive match for more than 20% of contigs >300 bp (Supporting information Table S4). The percentage increased for the longer contigs, i.e. 97.30% of the contigs >6000 bp yielded a positive BLAST hit in Assembly D against the Porifera nr NCBI database (Supporting information Table S4). Overall, the number of contigs

Table 3 *Crella elegans* contig statistics by size

Contig length	Assembly A (non-reproductive)		Assembly B (spermatic cysts)		Assembly C (larvae)		Assembly D (all stages)	
	Number of contigs	%	Number of contigs	%	Number of contigs	%	Number of contigs	%
<300	34 193	52.14	75 418	43.96	41 224	56.34	95 188	46.86
300–500	18 326	27.94	48 967	28.54	19 270	26.33	58 136	28.62
500–1000	10 198	15.55	31 764	18.52	10 090	13.79	34 920	17.19
1000–2000	2708	4.13	12 607	7.35	2416	3.30	12 366	6.09
2000–3000	147	0.22	2168	1.26	163	0.22	1946	0.96
3000–4000	4	0.01	411	0.24	8	0.01	398	0.20
4000–5000	3	0.00	139	0.08	3	0.00	107	0.05
5000–6000	0	0.00	45	0.03	0	0.00	36	0.02
>6000	0	0.00	29	0.02	0	0.00	37	0.02
Total	65 579		171 548		73 174		203 134

Fig. 1 Number of contigs per assembly, shown in three size ranges. Assemblies B (with spermatic cysts) and D (all stages) presented not only the largest number of contigs, but also the largest contigs (i.e. assemblies A and C do not present contigs above 5000 bp).

identified as sponge genes ranged from 13 475 (Assembly C) to 23 304 (Assembly D) (Fig. 3).

We then proceeded with the functional annotation of the contigs with BLAST hits (both against the Metazoa + Fungi and Porifera nr database) by using the free software BLAST2GO (Conesa *et al.* 2005) (see Materials and methods). Unlike the BLAST results, a large difference was observed when comparing the number of annotated genes (sequences matching known genes with assigned GO terms) of Metazoa and Porifera. The percentage of contigs with at least one BLAST hit against the Porifera nr database with annotations ranged from 5.00% (Assembly B) to 6.05% (Assembly C) (Supporting information Table S4). However, more than 25% of contigs with at least one BLAST hit (e.g. 40.8% for Assembly D) against the Metazoa + Fungi nr database were annotated in all four assemblies (Supporting information

Table S3). These numbers are considerably lower than those observed in similar studies (Mizrachi *et al.* 2010; Angeloni *et al.* 2011; Ewen-Campen *et al.* 2011; Iorizzo *et al.* 2011), stressing the lack of knowledge of specific function for most sponge genes. In both cases, using the Metazoa + Fungi and Porifera nr databases the largest number of annotated genes was found for Assembly D (7789 and 1183 respectively), as expected due to comprising reads of the three life cycle stages of *C. elegans* (Supporting information Table S3 and S4, Figs 2 and 3).

The functional categories of the known metazoan genes sequenced in this study were explored. An overview of the percentage of metazoan genes mapping to a selection of GO terms is shown in Fig. 4. For most GO terms the highest percentage corresponded to Assembly D. In particular, for some terms (e.g. developmental

Fig. 2 Percentile box plot showing an OHR (ortholog hit ratio) mean value (displayed with the dashed line) of all the assemblies above 0.3. Dots at the beginning and end of the box indicate the 5th and 95th percentiles respectively. Error bars indicate the 10th and 90th percentiles. The box delimits the 25th and 75th percentiles, the solid line within the box marking the median.

process, cell communication, cytoplasm) the difference regarding the other three assemblies was substantial. A more detailed analysis of 6 selected genes of special interest is shown in Table 4.

Transcriptome coverage and annotation assessment

To assess the coverage and to investigate how closely our contigs approached full-length genes we calculated the OHR using the method proposed by O'Neil *et al.* (2010). The average OHR ranged from 0.30 ± 0.25 in assembly A to 0.32 ± 0.27 in assembly D (Supporting information Table S5, Fig. 5). These values were similar to those found in other studies (O'Neil *et al.* 2010; Ewen-Campen *et al.* 2011; Riesgo *et al.* 2012).

We also tested the completeness of our transcriptome and annotation effectiveness by looking at a collection of genes belonging to three different metabolic pathways shared by all animal phyla, finding annotated contigs related to the three major pathways in our data (Table 5).

Fig. 3 Number of total contigs, contigs with BLAST hits against the Metazoa nr NCBI database, annotated contigs and contigs without hits. A positive relationship between size range and number of contigs with BLAST hits (light grey) is shown. Dark grey indicates the number of contigs without a BLAST hit.

Fig. 4 Number of total contigs, contigs with BLAST hits against the Porifera nr NCBI database, annotated contigs and contigs without hit. Note that the number of contigs with positive BLAST hits is remarkably superior when we BLAST against the Porifera nr database than when we BLAST against the Metazoa nr database, though the number of annotated contigs decreases. A positive relationship between the contig size range and the number of contigs with Porifera BLAST hits is shown.

Table 4 Number of sequences per assembly matching a specific GO term

GO Id	GO term	# sequences				Example gene (match accession)
		Assembly A (non-reproductive)	Assembly B (spermatic cysts)	Assembly C (larvae)	Assembly D (all stages)	
GO:0033554	Cellular response to stress	11	16	13	18	Serine threonine-protein kinase atr (XP_003416220.1)
GO:0006950	Response to stress	15	19	12	19	Catalase-like (XP_003389936.1)
GO:0019953	Sexual reproduction	1	1	0	1	Piwi-like protein 2-like (XP_003383170.1)
GO:0007276	Gamete generation	0	2	1	1	Muts protein homologue 4-like (XP_001627391.1)
GO:0048232	Male gamete generation	3	8	3	8	Calreticulin (XP_003388074)
GO:0050896	Response to stimulus	25	27	29	33	Heat shock protein 70 (CAA70695.1)

Fig. 5 GO term distribution of BLAST hits from the 4 assemblies of *Crella elegans*. GO terms are divided in the three top categories: Biological processes, Molecular function and Cellular components. Columns represent the percentage of contigs of each transcriptome mapping to a given GO term.

All selected homologues were recovered in assemblies A, C and D with three genes missing from assembly B [e.g. isocitrate dehydrogenase, 2-oxoglutarate dehydrogenase E1 component and isocitrate dehydrogenase (NAD⁺)] (Table 5). These results may have different explanations: (i) that despite assembly B having both more contigs than the other individual assemblies and the most contigs with BLAST hits (Supporting information Table S3) it did not yield the largest number of characterized genes; (ii) the genes could have low expression values and thus be difficult to detect or; (iii) the genes are not expressed in this specific developmental stage; and (iv) Overrepresentation of reads at the initial pool of data due to a PCR enrichment artefact during the last step of the library construction protocol can result in redundancy of some proteins and silencing of others.

Effect of overrepresented reads and initial sample size assessment

One of the above-mentioned possible explanations for the absence of some homologues of general metabolic genes in assembly B is the overrepresentation of reads at the initial pool of data due to a PCR enrichment artefact during the last step of the transcriptomic library construction protocol (Dong *et al.* 2011). As mentioned, overrepresentation of reads can lead to redundancy of some proteins and silencing of others (Dong *et al.* 2011). Redundancy of certain proteins was found in the four

assemblies (Fig. 6) with percentages just above 15% for all identified proteins (data not shown), as found in other studies (Meyer *et al.* 2009; Ewen-Campen *et al.* 2011). The identity of the five most redundant proteins in each assembly was checked to discard the possibility of contamination. For all assemblies the most redundant proteins were homologues to predicted proteins of the sponge *A. queenslandica*, suggesting that the missing transcripts in assembly B may actually not be the result of a library construction artefact. Proteins related to cell communication and collagen production were found to be the most redundant (Supporting information Table S6).

To avoid the size effect (i.e. number of raw reads) due to the different platforms (GAII and HiSeq) employed to sequence the samples and to test the effectiveness of each assembly corresponding to a life stage we collapsed the initial samples (collapsing identical reads in a FASTQ/A file into a single sequence while maintaining read counts) and sub-sampled 5 000 000 random reads for each of assemblies A, B and C. We then proceeded with the *de novo* assembly of each separate sub-sample—now controlling for size. This process was repeated three times for each sample while calculating the average total number of contigs, number of contigs with BLAST hits, number of contigs without BLAST hits and number of annotated contigs. The results of this analysis are shown in Fig. 7. When comparing the three assemblies, they are all similar in the number of contigs with a BLAST hit and functional annotation, but the total number of contigs is

Table 5 Selected metabolic pathway genes identified in the *Crella elegans* assemblies

Pathways	# Hits (contig length range)			
	Assembly A (non-reproductive)	Assembly B (spermatic cysts)	Assembly C (larvae)	Assembly D (all stages)
Glycolysis/gluconeogenesis				
Alcohol dehydrogenase (NADP+)	3 (727–1216)	3 (304–1410)	2 (561–610)	6 (342–995)
Hexokinase	1 (866)	1 (1522)	2 (402–847)	2 (487–489)
Pyruvate kinase	1 (1103)	1 (1875)	1 (1360)	3 (393–641)
Phosphoglycerate kinase	2 (664–748)	2 (518–2389)	2 (523–1216)	5 (493–796)
Enolase	3 (646–1232)	3 (488–3049)	3 (571–987)	5 (488–1041)
Aldose 1-epimerase	2 (354–431)	2 (373–2073)	1 (492)	3 (305–1098)
Phosphoglucomutase	1 (1491)	3 (532–1220)	1 (706)	3 (404–693)
ADP-dependent glucokinase	2 (1132–1565)	2 (1250–1845)	2 (505–586)	3 (839–3081)
Phosphoglucomutase/phosphopentomutase	1 (1491)	4 (532–1220)	1 (706)	3 (404–693)
Citrate cycle				
Malate dehydrogenase	4 (572–1123)	8 (453–1611)	2 (931–1100)	4 (504–1586)
Isocitrate dehydrogenase	2 (315–471)	0	1 (396)	4 (407–2527)
Pyruvate dehydrogenase E1 component subunit alpha	2 (307–656)	1 (990)	1 (326)	1 (872)
Pyruvate dehydrogenase E1 component subunit beta	2 (396–728)	1 (1945)	1 (728)	3 (664–1213)
2-oxoglutarate dehydrogenase E1 component	2 (412–1248)	0	1 (812)	3 (504–1276)
Citrate synthase	4 (379–947)	9 (304–1612)	6 (317–884)	7 (345–1646)
Pentose phosphate				
Isocitrate dehydrogenase (NAD+)	1 (445)	0	3 (509–769)	2 (748–2007)
Succinate dehydrogenase (ubiquinone) flavoprotein subunit	2 (458–1579)	3 (508–1536)	2 (419–1477)	3 (552–1593)
Succinate dehydrogenase (ubiquinone) iron-sulphur subunit	1 (499)	1 (352)	1 (866)	1 (430)
Succinate dehydrogenase (ubiquinone) cytochrome b560 subunit	1 (360)	2 (596–1494)	2 (514–1552)	1 (1562)
Aconitate hydratase 1	3 (530–882)	2 (773–2018)	4 (447–693)	2 (877–1971)

usually smaller in assembly A, as it was based on shorter (77 bp) reads. This trend is more visible towards the larger size contigs. The number of functional annotation for the size class >4000 is anecdotal, as there are just three annotated genes for assembly B.

In spite of assembly B presenting more contigs ($71\,528 \pm 133$) than assemblies A and C ($38\,458 \pm 48$ and $42\,946 \pm 300$) with the same number of initial reads, the number of contigs with BLAST hits ($12\,573 \pm 79$; $13\,780 \pm 106$ and $13\,047 \pm 72$; assemblies A, B and C respectively) and the number of annotated contigs (4800 ± 64 ; 5048 ± 43 and 5007 ± 64 ; assemblies A, B and C respectively) is very similar. This is most reasonably explained by the superior quality of reads from the HiSeq platform when compared to those of GAII (www.illumina.com/systems).

Developmental genetic toolkit of *Crella elegans*

To test the suitability of using our sponge for future phylogenetic and developmental genetic studies, we investigated developmental regulatory genes whose presence

has been reported for other metazoans, including sponges (Ono *et al.* 1999; Adell & Müller 2004; Hill *et al.* 2010; Holstien *et al.* 2010; Srivastava *et al.* 2010; Adamska *et al.* 2011). We focused our exploration on assembly D, as it is the most complete.

Our results showed that *C. elegans* possesses many transcription factors including Fox, Pax, Tbox and Homeobox genes (Tables 6, 7), as seen in other sponges (Adell & Müller 2004; Larroux *et al.* 2006; Holstien *et al.* 2010). In some cases, we found homologues to metazoan genes, such as Brachyury, a T-box family transcription factor universally expressed in metazoan gastrulation (Adamska *et al.* 2011), not found in *A. queenslandica* (Larroux *et al.* 2008), though it has been reported for other sponges (Manuel *et al.* 2004; Adell & Muller 2005). In a previous study, we also reported the presence of homologues to developmental signalling pathway components such as Notch, TGF- β and Hedgehog (Riesgo *et al.* 2012). With this study we were able to add genes related to the Wnt pathway—a signalling pathway involved in the regulation of morphogenesis in sponges (Adell *et al.* 2007) (Table 7).

Fig. 6 Redundancy of proteins hits. Two-part graph showing the number of protein hits with and without redundancy (left) and the percentage of protein hits with and without redundancy (right).

Fig. 7 Number of total contigs from randomly selected reads, contigs with BLAST hits against the Metazoa nr NCBI database, annotated contigs and contigs without hits. The number of contigs with BLAST hits and functional annotations is similar across the different life cycle stages.

Table 6 Selected developmental genes previously identified in different marine sponges with a homologue sequence (contig) in *Crella elegans*

Pathways	Assembly D (all stages)		
	# Hits	Contig length range (bp)	NCBI Homologue length (bp)
Transcription factor genes			
Brachyury	1	900	1176
TbxA	1	3508	891
Tbx4/5	2	960–2908	1188
TbxC/D	1	1756	999
Tbx1/15/20	1	2573	1887
Homeobox	1	716	1092
protein HOXA1			
Homeobox	1	1775	1035
protein HOXB1			
Homeobox	1	505	438
protein HOXC1			
BarX/Bsh	1	1494	759
SixC	1	1578	1353
SoxB1-like	2	662–710	1311
Lhx5-like	1	781	816
Lim3	1	742	1014
NF-κB	1	1332	3288
Pax 2/5/8	1	1033	1728
Forkhead box protein J2/3	1	555	1083
FoxL2	1	1414	825
FoxD	1	1406	1332
FoxP	1	1292	2010
FoxF	1	1757	1410
POUI	1	695	972

Table 7 Selected Wnt pathway genes previously identified in different marine sponges with a homologue sequence (contig) in *Crella elegans*

Wnt Pathway	Assembly D (all stages)		
	# Hits	Contig length (bp)	NCBI Homologue length (bp)
sFRPA	1	679	570
LZIC	1	554	570
Activated CDC42 kinase 1	1	842	474
RAC	1	464	576
TCF/LEF	1	1078	1248
GSK3	1	322	1308
FzdA	1	822	1776

Our results indicate not only that our sequencing effort was deep enough to sequence homologues of most developmental regulatory genes but also that the demosponge *C. elegans* shares many developmental gene

Fig. 8 Heat map (distances measured: 1-Pearson correlation; clusters linkage criteria: single linkage) performed with CLC bio showing an overview of the expression level per contig per assembly.

families with other metazoans, as recently reported for other sponges (Larroux *et al.* 2006, 2008; Hargett *et al.* 2010; Adamska *et al.* 2011).

A first look into the gene expression along the life cycle of Crella elegans

An initial comparison of the expression level per contig between the assemblies for the three life cycle stages was performed taking into account all contigs longer than 300 bp and above 10 average read coverage with a positive BLAST hit against the Porifera NCBI database (i.e. 23 605 contigs per assembly) (Fig. 8). To simplify the results, we categorized genes with different levels of expression by plotting the log₂ transformed RPKM values vs. the expressed genes (Wickramasinghe *et al.* 2012). According to the graph obtained (not shown) three categories were established: high (≥ 250 RPKM), medium (≥ 160 to 250 RPKM) and low (<160 RPKM) expressed genes. As shown in Table 8 the number of total expressed genes for the non-reproductive stage (Assembly A) and the stage with larvae (Assembly C) were comparable, while the greatest number of expressed genes (including the number of highly expressed genes, i.e. 206) was found at the stages with spermatic cysts (Assembly B). A list of the 10 highest expressed genes per life cycle stage is provided in Supporting information Table S7. Except for the predicted kinesin-like protein KIFC3-like gene, which is the most expressed gene in the non-reproductive stage and in the stage with larvae (but not a top

Table 8 Number of expressed genes in RNA-Seq analysis per life cycle stage and category

	Assembly A (non-reproductive)	Assembly B (spermatic cysts)	Assembly C (larvae)
Highly expressed genes (≥ 250 RPKM)	54	206	160
Medium expressed genes (≤ 160 RPKM to 250 RPKM)	725	1231	881
Lowly expressed genes (<160 RPKM)	3672	5077	3409
Total expressed genes	4451	6514	4450
Non expressed genes	10 908	8845	10 909

Table 9 GO-terms and sequence description of the highly expressed genes per life cycle stage

Sequence Description	GO id
Assembly A (non-reproductive)	
Ribosomal protein 3 large subunit	P:GO:0010467; C:GO:0030529; F:GO:0005198
Protein tyrosine kinase	F:GO:0004713; P:GO:0009987
Heat shock protein 70	F:GO:0032559; P:GO:0050896
Rab8	F:GO:0032561; P:GO:0006810; P:GO:0035556
Assembly B (spermatic cysts)	
Calcyphosine	F:GO:0046872
Heat shock protein 70	F:GO:0032559; P:GO:0050896
Protein tyrosine kinase L27	F:GO:0016301; C:GO:0030529
Pre-b-cell colony-enhancing factor	F:GO:0016763; P:GO:0019359; C:GO:0044424
Protein tyrosine kinase	F:GO:0016740
Polyprotein	F:GO:0005488
Protein kinase c-related kinase	P:GO:0006464; F:GO:0004672; C:GO:0044464; P:GO:0023052; F:GO:0032559
Assembly C (larvae)	
Elongation factor 1 alpha	P:GO:0006412; F:GO:0017111; P:GO:0009207; F:GO:0008135; F:GO:0032561; C:GO:0044424
Protein tyrosine kinase	F:GO:0004713; P:GO:0009987
Heat shock protein 70	F:GO:0032559; P:GO:0050896
L10a	F:GO:0003676; P:GO:0010467; C:GO:0030529; F:GO:0005198
Elongation factor 1 alpha	P:GO:0006412; F:GO:0017111; P:GO:0009207; F:GO:0008135; F:GO:0032561; C:GO:0044424
Alpha- partial	F:GO:0017111; C:GO:0032991; C:GO:0015630; P:GO:0043623; P:GO:0009207; F:GO:0032561; P:GO:0007017
Myosin ii	F:GO:0017111; C:GO:0015629
Actin	F:GO:0032559

10 most expressed protein in the stage with spermatic cysts), the top 10 highly expressed genes were non-overlapping among life cycle stages (Table 9).

To examine the functional changes associated with the expression of genes along the three life cycle stages we explored the GO terms associated to the expressed genes

using BLAST2GO (Fig. 9a). Despite some GO terms being higher (in terms of numbers of sequences mapping against a specific GO term) in some stages, most were represented by comparable number of sequences among the three stages. However, some GO terms were not present in one or more stages. A good example of that is the GO term associated to apoptosis (i.e. Biological processes category) (Fig. 9a). We did not find any apoptosis gene corresponding to this GO term in the non-reproductive stage and in the stage with spermatic cysts, but we found one gene in the stage with larvae. A possible explanation is that at the end of the reproductive season, individuals of *C. elegans* decrease drastically in size and some individuals even disappear (Pérez-Porro *et al.* 2012). We assumed that the disruption of the tissue (i.e. aquiferous system) due to gametogenesis and brooding embryos and larvae that is reflected in the shrinking of the individuals is also accompanied by cellular apoptosis. A remarkable observation was that the number of GO terms found per life stage and category (i.e. biological process, molecular function and cellular component) were comparable, even if those may correspond to different genes (Fig. 9b).

Focusing on the highly expressed genes with GO annotations we found that two genes were presented in the three stages: protein tyrosine kinase (PTK) and heat shock protein 70 (HSP70) (table 9). PTKs are involved in critical signalling events that control fundamental physiological processes such as cell growth and differentiation, cell cycle and cytoskeleton (Müller *et al.* 1999), which occur all along the sponge life-cycles. HSPs are molecular chaperones encoded by highly conserved genes; HSP70 in one of the families typically related with thermal stress (Parsell & Lindquist 1993; Feder & Hofmann 1999) but also expressed in many other stressful situations (Agell *et al.* 2001, 2004; Rossi *et al.* 2006). The RPKM values for the PTK at the three different life cycle stages were 259.31, 267.97 and 284.68 (non-reproductive sample, sample with spermatic cysts and sample with larvae respectively). The RPKM values for the HSP70 were 257.65, 269.68 and 282.86 (non-reproductive sample, the sample with spermatic cysts and the sample with larvae respectively). In both cases, the RPKM values increased along the life cycle of *C. elegans*. We hypothesized a correlation between the increased expression

Fig. 9 (a) GO terms distribution of expressed genes from the three life cycle stages corresponding to assemblies A, B and C; (b) Number of GO terms per life cycle stage and GO category.

level of HSP70 and the sponge reproductive events due to the numerous cellular changes associated with cell transformation in gametes and embryo development (a stressful process involving cell-protein destruction and reorganization). Furthermore, as described by Rossi *et al.* (2006) variations in HSP70 expression levels over an annual cycle can be indirectly caused by sea water temperature oscillation along the year which is also often associated to food availability. Previous studies (Pérez-Porro *et al.* 2012) showed that the reproductive season for *C. elegans* begins with the increase of the sea water temperature, an event that is also coincident with a period of food low, which can contribute to increasing the expression levels of HSP70. In the case of the PTKs, some studies have postulated a role in egg activation (Sato *et al.* 2000; Runft *et al.* 2002) and thus, even if this role has not been demonstrated in sponges, they may be related with reproduction in *C. elegans*.

Our efforts for future studies will focus on the gene expression along the life cycle of *C. elegans*, to answer general questions on reproduction of basal metazoans, specifically marine sponges.

Conclusions

Next-generation sequencing techniques allowed the characterization of the transcriptome of a sponge, a

non-model animal of key phylogenetic and ecological importance for which transcriptomic and genomic data remain scarce (Uribe & Turon 2012). The new data on three stages of the life cycle of the sponge and the results presented here make *C. elegans* closer to becoming a model encrusting sponge for future research and points in new directions for generating more complete transcriptomes in poorly known organisms. The first examination of the gene expression along the life cycle of this sponge identifies that PTKs play a role during the reproduction of marine invertebrates (Sato *et al.* 2000; Runft *et al.* 2002) and highlights the importance of this approach for future directions.

Acknowledgements

Initial protocols for RNA sequencing were shared by Steve Vollmer (Northeastern University), who kindly hosted A.R.P.-P. in his laboratory. We thank Janina González, Andrea Blanquer and Sònia de Caralt for fieldwork assistance and Nick Petschek and Prashant Sharma for critically reading the manuscript. Editor Dr. Andrew DeWoody and three anonymous reviewers provided comments that helped to improve this article. A.R.P.-P. was supported by a FPI pre-doctoral Fellowship (BES-2008-003009) from the Spanish *Ministerio de Economía y Competitividad* project CTM2007-66635-C02-01 and a grant through the *Benthomies* (CTM2010-22218-C02-01) project to M.J.U. This work is based in part upon research funded by the US National Science

Foundation under grants 0844881, 0732903 and 0531757 from the Assembling the Tree of Life and Phylogenetic Systematics programs to G.G. The Bauer Center for Genomics Research and the Harvard research Computing group made this research possible.

References

- Adamska M, Degnan BM, Green K, Zwafink C (2011) What sponges can tell us about the evolution of developmental processes. *Zoology*, **114**, 1–10.
- Adell T, Muller WEG (2005) Expression pattern of the Brachyury and Tbx2 homologues from the sponge *Suberites domuncula*. *Biology of the Cell*, **97**, 641–650.
- Adell T, Müller WEG (2004) Isolation and characterization of five Fox (Forkhead) genes from the sponge *Suberites domuncula*. *Gene*, **334**, 35–46.
- Adell T, Thakur AN, Muller WE (2007) Isolation and characterization of Wnt pathway-related genes from Porifera. *Cell Biology International*, **31**, 939–949.
- Agell G, Uriz M-J, Cebrian E, Martí R (2001) Does stress protein induction by copper modify natural toxicity in sponges? *Environmental Toxicology and Chemistry*, **20**, 2588–2593.
- Agell G, Turon X, De Caralt S, López-Legentil S, Uriz MJ (2004) Molecular and organism biomarkers of copper pollution in the ascidian *Pseudodistoma crucigaster*. *Marine Pollution Bulletin*, **48**, 759–767.
- Angeloni F, Wagemaker CAM, Jetten MSM *et al.* (2011) *De novo* transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. *Molecular Ecology Resources*, **11**, 662–674.
- Blankenberg D, Kuster GV, Coraor N *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, **89**, 19.10.1–19.10.21.
- Bogdanova EA, Shagina I, Barsova EV, Kelmanson I, Shagin DA, Lukyanov SA (2010) Normalizing cDNA libraries. *Current Protocols in Molecular Biology*, **90**, 5.12.1–5.12.27.
- Çelik İ, Cirik Ş, Altunağaç U *et al.* (2011) Growth performance of bath sponge (*Spongia officinalis* Linnaeus, 1759) farmed on suspended ropes in the Dardanelles (Turkey). *Aquaculture Research*, **42**, 1807–1815.
- Collins LJ, Biggs PJ, Voelckel C, Joly S (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Informatics*, **21**, 3–14.
- Conesa A, Götz S, García-Gómez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
- Dong H, Chen Y, Shen Y *et al.* (2011) Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System. *Acta Biochimica et Biophysica Sinica*, **43**, 496–500.
- Dunn CW, Hejnal A, Matus DQ *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 665–780.
- Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
- Ewen-Campen B, Shaner N, Panfilio KA *et al.* (2011) The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. *BMC Genomics*, **12**, 61.
- Feder ME, Hofmann GE (1999) Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annual Review of Physiology*, **61**, 243–282.
- Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics*, **12**, 317.
- Fusetani N (2010) Biotechnological potential of marine natural products. *Pure and Applied Chemistry*, **82**, 17–26.
- Gauthier MEA, Du Pasquier L, Degnan BM (2010) The genome of the sponge *Amphimedon queenslandica* provides new perspectives into the origin of Toll-like and interleukin 1 receptor pathways. *Evolution & Development*, **12**, 519–533.
- Giribet G, Dunn CW, Edgecombe GD, Rouse GW (2007) A modern look at the animal tree of life. *Zootaxa*, **1668**, 61–79.
- Harcet M, Roller M, Cetkovic H *et al.* (2010) Demosponge EST sequencing reveals a complex genetic toolkit of the simplest metazoans. *Molecular Biology and Evolution*, **27**, 2747–2756.
- Hejnal A, Obst M, Stamatakis A *et al.* (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 4261–4270.
- Hill A, Boll W, Ries C *et al.* (2010) Origin of pax and six gene families in sponges: single *PaxB* and *Six1/2* orthologs in *Chalinula loosanoffi*. *Developmental Biology*, **343**, 106–123.
- Holstien K, Rivera A, Windsor P *et al.* (2010) Expansion, diversification, and expression of T-box family genes in Porifera. *Development Genes and Evolution*, **220**, 251–262.
- Hooper JNA, Van Soest RWM (2002) *Systema Porifera: A Guide to the Classification of Sponges*. Kluwer Academic/Plenum Publishers, New York. pp. 1810.
- Hunt B, Vincent ACJ (2006) Scale and sustainability of marine bioprospecting for pharmaceuticals. *Ambio*, **35**, 57–64.
- Iorizzo M, Senalik DA, Grzebelus D *et al.* (2011) *De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics*, **12**, 389.
- Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, **12**, 382.
- Larroux C, Fahey B, Liubicich D *et al.* (2006) Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evolution & Development*, **8**, 150–173.
- Larroux C, Luke GN, Koopman P *et al.* (2008) Genesis and expansion of metazoan transcription factor gene classes. *Molecular Biology and Evolution*, **25**, 980–996.
- Manuel M, Le Parco Y, Borchellini C (2004) Comparative analysis of Brachyury T-domains, with the characterization of two new sponge sequences, from a hexactinellid and a calcisponge. *Gene*, **340**, 291–301.
- Meyer E, Aglyamova G, Wang S *et al.* (2009) Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFLX. *BMC Genomics*, **10**, 219.
- Mizrachi E, Hefer C, Ranik M, Joubert F, Myburg A (2010) *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics*, **11**, 681.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**, 621–628.
- Müller WEG, Kruse M, Blumbach B, Skorokhod A, Müller IM (1999) Gene structure and function of tyrosine kinases in the marine sponge *Geodia cydonium*: autapomorphic characters of Metazoa. *Gene*, **238**, 179–193.
- O'Neil S, Dzurisin J, Carmichael R *et al.* (2010) Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics*, **11**, 310.
- Ono K, Suga H, Iwabe N, K-i Kuma, Miyata T (1999) Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of animals before the parazoan-eumetazoan split. *Journal of Molecular Evolution*, **48**, 654–662.
- Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.
- Parsell DA, Lindquist S (1993) The function of heat-shock proteins in stress tolerance - degradation and reactivation of damaged proteins. *Annual Review of Genetics*, **27**, 437–496.
- Pérez-Porro AR, González J, Uriz M (2012) Reproductive traits explain contrasting ecological features in sponges: the sympatric poecilosclerids *Hemimyscale columella* and *Crella elegans* as examples. *Hydrobiologia*, **687**, 315–330.

- Philippe H, Derelle R, Lopez P *et al.* (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, **19**, 706–712.
- Riesgo A, Pérez-Porro AR, Carmona S, Leys SP, Giribet G (2011) Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. *Molecular Ecology Resources*, **12**, 312–322.
- Riesgo A, Andrade SC, Sharma PP *et al.* (2012) Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Frontiers in Zoology*, **9**, 33.
- Rossi S, Snyder M, Gili J-M (2006) Protein, carbohydrate, lipid concentrations and HSP 70–HSP 90 (stress protein) expression over an annual cycle: useful tools to detect feeding constraints in a benthic suspension feeder. *Helgolander Marine Research*, **60**, 7–17.
- Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nature Biotechnology*, **26**, 1117–1124.
- Runft LL, Jaffe LA, Mehlmann LM (2002) Egg activation at fertilization: where it all begins. *Developmental Biology*, **245**, 237–254.
- Sato K-i, Tokmakov AA, Fukami Y (2000) Fertilization signalling and protein-tyrosine kinases. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, **126**, 129–148.
- Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of eumetazoa. *Molecular Biology and Evolution*, **26**, 2261–2274.
- Srivastava M, Larroux C, Lu D *et al.* (2010) Early evolution of the LIM homeobox gene family. *BMC Biology*, **8**, 4.
- Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research*, **20**, 1432–1440.
- Tariq MA, Kim HJ, Jejelowo O, Pourmand N (2011) Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research*, **39**, e120.
- Uriz MJ, Turon X (2012) Chapter five - Sponge ecology in the molecular era. In: *Advances in Marine Biology* (eds Mikel A, Becerro MJUMM, Xavier T), pp. 345–410. Academic Press, New York, NY.
- Wickramasinghe S, Rincon G, Islas-Trejo A, Medrano JF (2012) Transcriptional profiling of bovine milk using RNA sequencing. *BMC Genomics*, **13**, 45.
- Wörheide G, Solé-Cava AM, Hooper JNA (2005) Biodiversity, molecular ecology and phylogeography of marine sponges: patterns, implications and outlooks. *Integrative and Comparative Biology*, **45**, 377–385.

A.R.P.P. participated in the conception of the study, produced, assembled, analysed the data and wrote the

manuscript. D.N.G. helped with the development of scripts. M.J.U. and G.G. participated in the conception of the study, supervised the research, and helped to draft the manuscript.

Data Accessibility

Raw sequence data: NCBI SRA: SRX217043, SRX216921, SRX216926. Final DNA sequence assemblies: DRYAD entry doi:10.5061/dryad.50dc6.

Supporting Information

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

Table S1 Average coverage statistics for the different assemblies of *Crella elegans*

Fig. S2 Log/log plots showing the positive correlation between contig length and the number of reads for a given contig for all four assemblies.

Table S3 Number and % of contigs without BLAST hit, with BLAST hit (but without annotation) and with functional annotation, by size range. (BLAST performed against the Metazoa + Fungi nr database of NCBI)

Table S4 Number and % of contigs without BLAST hit, with BLAST hit (but without annotation) and with functional annotation, by size range. (BLAST performed against the Porifera nr database of NCBI)

Table S5 OHR (Ortholog hit ratio) statistics for the different assemblies of *Crella elegans*

Table S6 Identity of the five most redundant proteins in each assembly of *Crella elegans*

Table S7 Definition (accordingly with NCBI) and RPKM values of the 10 most expressed contigs per life cycle stage