Ryu *et al. BMC Genomics* (2016) 17:158
DOI 10.1186/s12864-016-2501-0

BMC Genomics

RESEARCH ARTICLE Open Access

Hologenome analysis of two marine sponges with different microbiomes

Taewoo Ryu[1,2,12*], Loqmane Seridi[1,2], Lucas Moitinho-Silva[3], Matthew Oates[4], Yi Jin Liew[2,5], Charalampos Mavromatis[1,2], Xiaolei Wang[6,7], Annika Haywood[2], Feras F. Lafi[2,8], Marija Kupresanin[1,2], Rachid Sougrat[9], Majed A. Alzahrani[6,7], Emily Giles[2], Yanal Ghosheh[1,2], Celia Schunter[1,2], Sebastian Baumgarten[2,5], Michael L. Berumen[2,5], Xin Gao[6,7], Manuel Aranda[2,5], Sylvain Foret[10], Julian Gough[4], Christian R. Voolstra[2,5], Ute Hentschel[11] and Timothy Ravasi[1,2*]

Abstract

Background: Sponges (Porifera) harbor distinct microbial consortia within their mesohyl interior. We herein analysed the hologenomes of *Stylissa carteri* and *Xestospongia testudinaria*, which notably differ in their microbiome content.

Results: Our analysis revealed that *S. carteri* has an expanded repertoire of immunological domains, specifically Scavenger Receptor Cysteine-Rich (SRCR)-like domains, compared to *X. testudinaria*. On the microbial side, metatranscriptome analyses revealed an overrepresentation of potential symbiosis-related domains in *X. testudinaria*.

Conclusions: Our findings provide genomic insights into the molecular mechanisms underlying host-symbiont coevolution and may serve as a roadmap for future hologenome analyses.

Keywords: Sponge, *Stylissa carteri*, *Xestospongia testudinaria*, Innate immune system, Host, Microbial symbionts, Hologenome

Background

Microbial symbionts are being increasingly recognised as deeply integral components of multicellular organisms that affect core host functions such as development, immunity, nutrition, and reproduction [1]. The holobiont (synonym with "metaorganism", and defined as the host organism and its collective microbial community) [2] is thus considered a biological unit of natural selection (the "hologenome theory") [3]. However, the molecular mechanisms (e.g., immune system evasion and tolerance) that have resulted in these symbiotic partnerships are poorly understood.

Sponges (Porifera) represent one of the oldest, still extant animal phyla. Fossil evidence dating back 580 million years ago shows their existence in the Precambrian long before the radiation of all other animal phyla [4]. Sponges are globally distributed in all aquatic habitats from warm tropical reefs to the cold deep sea and are even present in freshwater lakes and streams. As sessile filter feeders, they pump many thousands liters of water per day through the aquiferous canal system that is embedded within the sponge body and are constantly exposed to a plethora of microorganisms from the environment [5]. Many species are colonised by dense and diverse microbial consortia that are contained extracellularly within the mesohyl matrix ("high microbial abundance" (HMA)), while other species are nearly devoid of microorganisms ("low microbial abundance" (LMA)) [6–8].

To investigate factors involved in sponge-microbe interactions, we herein sequenced and analysed hologenome data including genome, transcriptome, and metatranscriptome of *S. carteri* (an LMA sponge and hereafter referred to as "*SC*") and *X. testudinaria* (an HMA sponge and hereafter referred to as "*XT*") (Fig. 1, Additional file 1 and Additional file 2). These two sponges were collected from the same habitat, which ensured systematic comparison by minimizing environmental effect such as different planktons and temperature. *SC* is the first species in the order Halichondrida to have its genome sequenced while *XT* (order Haplosclerida) is the first HMA sponge to have

* Correspondence: TaewooRyu16@apcc21.org; timothy.ravasi@kaust.edu.sa
[1]KAUST Environmental Epigenetic Program (KEEP), King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
[2]Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
Full list of author information is available at the end of the article

© 2016 Ryu et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

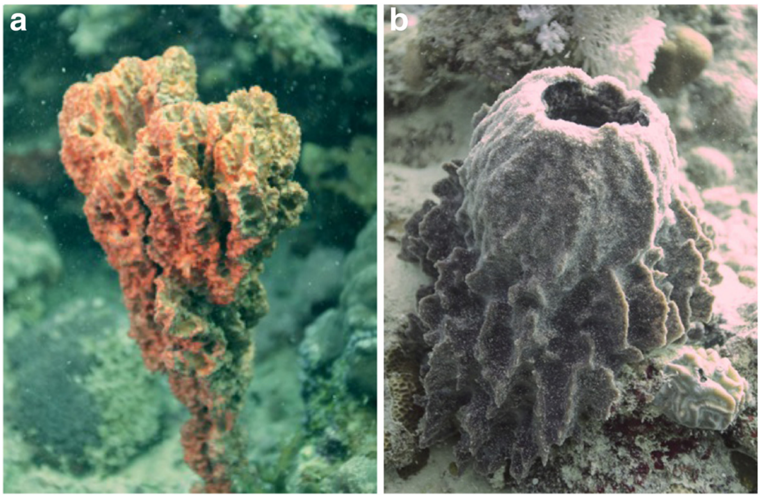


Fig. 1 Sponge species. Underwater images of *Stylissa carteri* (**a**) and *Xestospongia testudinaria* (**b**) taken by Michael L. Berumen

its genome sequenced [8]. Thus, our work provides a unique and valuable resource for future studies.

Results and discussion

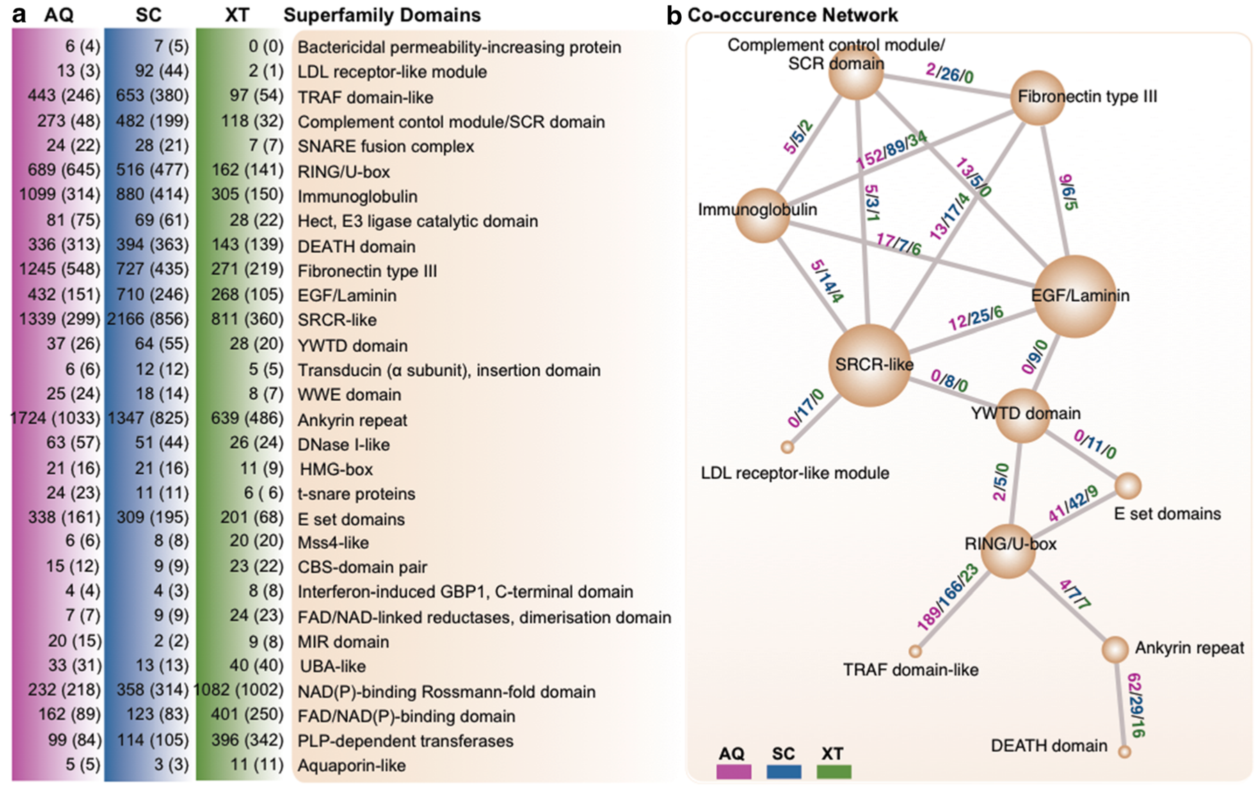
Genome assembly and gene annotation

Our assemblies of the *SC* and *XT* genomes yielded 97,497 and 97,640 scaffolds with base-level coverages of 109 X and 59 X, respectively (Additional file 3). The estimated genome sizes obtained from our assemblies were comparable to experimentally determined genome sizes measured by flow cytometry (386.31 and 161.37 Mbp for *SC* and *XT*, respectively, see Methods). 26,967 and 22,337 gene models of high quality were predicted covering 94.5 and 94.3 % of the core eukaryotic genes for *SC* and *XT*, respectively (see Methods). Comparison to publicly available sponge gene models from draft genomes or transcriptomes [9–11] shows that our gene models have reasonable quality in terms of the number of coding sequences (CDSs), the representation of core eukaryotic gene set [12], the number of genes with protein domains, and the number of protein domains (Additional file 4).

Expansion of innate immunological domains in sponge hosts

We checked for protein domains that were unusually over- or under-represented in the studied sponges compared to other eukaryotes compiled in the SUPERFAMILY database [13]. Most protein domains were neither over- nor under-represented, indicating that our gene models are comparable to those of other eukaryotes (Additional file 5). Both sponges (particularly *SC*) showed substantial expansions of immunological and receptor domains (Additional file 5) [14]. We focused our analysis on protein domains relevant to host-microbe interactions by using four keywords ("symbio," "innate immunity," "antimicrobial peptides," and "antibacterial") in functional annotations of the SUPERFAMILY database [13] (Fig. 2a and Additional file 6). We also included *Amphimedon queenslandica* (hereafter referred to as "*AQ*") in this analysis because its genome has been stably annotated [9] and also because its status with respect to microbial load is well known (LMA, Sandie Degnan, personal communication). Furthermore, overall gene contents of *AQ* are comparable to our studied sponges: among 30,060 gene models for *AQ*, 17,567 genes contained 32,326 SUPERFAMILY domains (28,027 and 29,156 SUPERFAMILY domains from 17,074 and 17,664 genes for *SC* and *XT*, respectively. Additional file 4). Other sponges with public transcript models were excluded due to incompleteness of gene models as shown in Additional file 4 and lack of exact status of LMA and HMA.

Our results revealed that there had been a striking expansion of the Scavenger Receptor Cysteine-Rich (SRCR)-like domain in the three tested sponges compared to all other eukaryotes compiled in the SUPERFAMILY database (see Additional file 6 for selected taxa). One known function of SRCR-like domains is recognition of large and diverse patterns of macromolecules (e.g., modified low-density lipoprotein; LDL) on microbial surfaces and enhancement of the phagocytic clearance of microbes [7]. In mammals, malfunctions in SRCR-like-domain-containing proteins have been linked to diseases and bacterial/viral infections [15]. It has been suggested that a protein containing this domain in the Mediterranean sponge (*Petrosia ficiformis*) may function in the recognition of photosymbionts [16]. We found that the *SC* genome contains 2166 SRCR-like domains, which is the highest number found among the 427 eukaryotes compiled in the SUPERFAMILY database (average, 28 copies). Interestingly, the next-highest known copy number for this domain family is



AQ	SC	XT	Superfamily Domains
6 (4)	7 (5)	0 (0)	Bactericidal permeability-increasing protein
13 (3)	92 (44)	2 (1)	LDL receptor-like module
443 (246)	653 (380)	97 (54)	TRAF domain-like
273 (48)	482 (199)	118 (32)	Complement contol module/SCR domain
24 (22)	28 (21)	7 (7)	SNARE fusion complex
689 (645)	516 (477)	162 (141)	RING/U-box
1099 (314)	880 (414)	305 (150)	Immunoglobulin
81 (75)	69 (61)	28 (22)	Hect, E3 ligase catalytic domain
336 (313)	394 (363)	143 (139)	DEATH domain
1245 (548)	727 (435)	271 (219)	Fibronectin type III
432 (151)	710 (246)	268 (105)	EGF/Laminin
1339 (299)	2166 (856)	811 (360)	SRCR-like
37 (26)	64 (55)	28 (20)	YWTD domain
6 (6)	12 (12)	5 (5)	Transducin (α subunit), insertion domain
25 (24)	18 (14)	8 (7)	WWE domain
1724 (1033)	1347 (825)	639 (486)	Ankyrin repeat
63 (57)	51 (44)	26 (24)	DNase I-like
21 (16)	21 (16)	11 (9)	HMG-box
24 (23)	11 (11)	6 (6)	t-snare proteins
338 (161)	309 (195)	201 (68)	E set domains
6 (6)	8 (8)	20 (20)	Mss4-like
15 (12)	9 (9)	23 (22)	CBS-domain pair
4 (4)	4 (3)	8 (8)	Interferon-induced GBP1, C-terminal domain
7 (7)	9 (9)	24 (23)	FAD/NAD-linked reductases, dimerisation domain
20 (15)	2 (2)	9 (8)	MIR domain
33 (31)	13 (13)	40 (40)	UBA-like
232 (218)	358 (314)	1082 (1002)	NAD(P)-binding Rossmann-fold domain
162 (89)	123 (83)	401 (250)	FAD/NAD(P)-binding domain
99 (84)	114 (105)	396 (342)	PLP-dependent transferases
5 (5)	3 (3)	11 (11)	Aquaporin-like

Fig. 2 Domain expansion and contraction in LMA vs. HMA sponges. Innate immune- and symbiosis-related protein domains were selected from the functional annotation. **a** The number of times a domain occurs in each genome. Shown are domains found ≥ 1.5 times more in *SC* compared to *XT*, or vice versa. *AQ* is included for comparison. The domain occurrence is highly correlated in the LMA sponges (*AQ* and *SC*; Pearson correlation coefficient = 0.92). The numbers of genes containing the domains are shown in the bracket. **b** The co-occurrence of innate immune domains in sponges. Nodes represent domains, while the numbers above the edges indicate the number of proteins in which the two domains co-occur for each sponge (*the color key is shown at the bottom-left*). Each node size is proportional to the number of outgoing edges. Edges are shown only when the domain pairs were observed ≥ 5 times in ≥ 1 species

found in *AQ*, followed closely by the sea urchin, *Strongylocentrotus purpuratus* (1339 and 1328 copies, respectively). In contrast, *XT* was found to have 811 copies.

The SRCR-like domains also show unique combinations with other immune system domains in sponges (Fig. 2b). For example, SRCR-like domains co-occur with LDL receptor-related domains (i.e., the LDL receptor-like module and the YWTD domain [17]) only in *SC*. SRCR-like domains are also associated with a broad collection of immunoglobulin-like beta-sandwich-folds in the three sponges, but most prominently in *SC*; these associated domains include the fibronectin type III and immunoglobulin domains, which are involved in cell surface recognition [18]. Compared to *XT*, *AQ* and more notably *SC* have undergone considerable expansions in combinations of the SRCR-like and abovementioned domains. In the SRCR-like domains, most of the amino acid residues are highly variable except for certain key residues including specific cysteine residues that enable the SRCR-like domains to recognise a vast array of ligands [15]. Clustering of the SRCR-like domain sequences from the three sponges yielded a large number of groups (169 clusters with ≥ 5 domains) whose members showed distinct patterns in their cysteine residues and levels of sequence conservation (Fig. 3a; see Methods). Thus, these domains are characterised by great diversity at the sequence level. The clusters were also distinct from one another in terms of their species compositions and expansion levels (Fig. 3b). The largest clusters contained the SRCR-like domains of *SC* and *AQ*, indicating that these domains are diversified to a greater extent in these species than in *XT*.

Additionally, we observed the expansions of other innate immune domains in *SC* and *AQ* (Fig. 2a). Among the selected examples are bactericidal permeability-increasing proteins. These host-defending antibiotic molecules, which selectively kill gram-negative bacteria [19, 20], were found only in *AQ* and *SC*. High-mobility group (HMG)-box domains were also found to be expanded in *SC* and *AQ* over *XT*. HMG proteins are primarily nucleosome-binding proteins, but some members are released extracellular milieu and propagate danger signal upon infection and tissue damage to active innate and adaptive immune

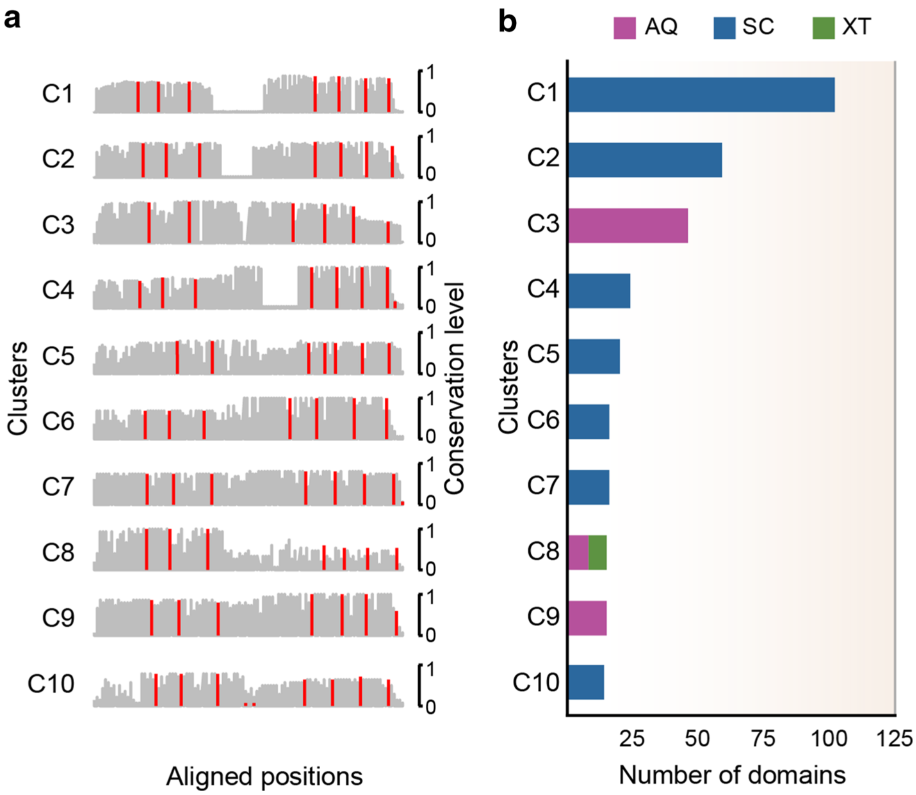


Fig. 3 Diversity of sponge SRCR-like domains. **a** Clustering and sequence alignments revealed diverse primary structures of the sponge SRCR-like domains. The conservation levels of cysteine residues are shown in red, while those of other amino acids are shown in gray. Only the top ten largest clusters are shown. **b** Lineage-specific expansions of each cluster in (**a**). This analysis suggests that the SRCR-like domains had expanded in LMA sponges, possibly enabling the recognition of various microbial ligands

responses in higher eukaryotes [21–23], which is known as "alarmin" functions which sense exogenous microbe- or pathogen-associated molecular patterns (MAMPs/PAMPs) or endogenous danger-associated molecular patterns (DAMPs) and then modulate downstream immune responses. Although not listed in Fig. 3 due to our SUPERFAMILY-based annotation scheme [13], another set of alarmins, the NACHT domains (PF05729), were also found to be enriched in *AQ* (230 copies) and *SC* (64 copies) compared to *XT* (21 copies). This domain is a component of the nucleotide-binding domain and leucine-rich repeat (NLR) proteins, which are major intracellular pattern recognition receptors (PRRs) [14]. DEATH domains, which are often found in MYD88 and NLR proteins, and the TRAF domain-like domains, which functions downstream of the classic Toll/Toll-like receptor pathway, were also found to be enriched in *SC* and *AQ* [11, 14, 24] (Fig. 2). Notably, however, the copy number of the Toll/interleukin receptor domain, which is a component of another set of PRR proteins [25], did not follow the above-described pattern, with 8, 17, and 16 copies found in *AQ*, *SC*, and *XT*, respectively (Additional file 6).

Interestingly, consistent with its symbiont-containing status (Additional file 1), *XT* was enriched over *SC* and *AQ* in protein domains that contribute to controlling symbiosis in some eukaryotes (Fig. 2). These include GBP1, which has been associated with the parasitophorous vacuole (responsible for host defense) [26], and aquaporin, which controls pH and the salt concentration in the symbiosome compartment in legumes, corals, and sponges (a symbiotic interface between host and microbes) [27–30].

We observed further strong correlations in the patterns of protein domain expansion between the LMA sponges, *AQ* and *SC*. In contrast, fewer similarities were found between the protein domains of *AQ* and *XT*, even though these two species belong to the same order (Haplosclerida). Analyses of antimicrobial peptides on the sponge genomes (Additional file 7, Additional file 8, Additional file 9, Additional file 10, and Additional file 11) and evolutionary rates of protein domains (Additional file 7 and Additional file 12) also provided consistent results.

Host-interaction factors in microbial symbionts

Previous studies showed that *SC* and *XT* harbor distinct microbial symbionts encompassing about 27 bacterial and archaeal phyla [31–35], but it is unclear how the unique microbiomes of LMA and HMA sponges are shaped in the context of the holobiont. We therefore analysed the metatranscriptomes of the microbial consortia in *SC* and *XT*. Since the metatranscriptome of *AQ* is not available, it could not be included in the present study. Although most well-known protein domains for symbiosis or pathogenesis [36–38] were not over-represented in any of the sponge symbionts, the fibronectin type III domain was among the most abundant domains in both sponge microbiomes, suggesting that this eukaryotic-like domain [37] may be a major contributor for the maintenance of host-microbe interactions (Additional file 13). Differential expression analysis of the microbiome genes identified several intriguing protein domains that were significantly over-represented in *XT* over *SC* (Fig. 4 and Additional file 14), including: the “Xylose isomerase-like TIM barrel” domain (PF01261), which is thought to be involved in the symbiosis of microbes with leguminous plants and the termite hindgut [39, 40]; the “HicB family” domain (PF05534), which is related to pilus formation and required for niche invasion [41]; the “PIN domain” (PF01850), which is found in the toxin-antitoxin operons of prokaryotes [42]; and the “Mycoplasma protein of unknown function” domain (PF03382), which has been detected in many pathogenic bacteria [43].

The sponge microbiomes also showed distinct community functions (Fig. 4 and Additional file 14). Consistent with our previous findings [33], light-harvesting functions were significantly enriched in the *SC* metatranscriptome, implying that photosynthesis is a major source of nutrients for the symbionts of *SC*. Additionally, virus-related functions were significantly enriched in the *SC* metatranscriptome, corroborating the idea that defence mechanisms against viruses, which are abundant in seawater, may be relevant to this community [37, 44]. On the other hand, the *XT* metatranscriptome was enriched for transposases, which may ensure the exchange of mobile genetic elements and help distribute selectable traits across diverse species [45].

Conclusions

Sponges serve as important organisms for the study of host-microbe interactions in lower marine invertebrates. Our present work identified expansion of potential immune system components especially SRCR-like domains in marine sponges compared to other eukaryotes probably as a result of coevolution with residing microbes. We also identified that sponge genomes expanded protein domains to a different extent by the microbiome contents. Our findings on the putative molecular underpinnings of sponge-microbe interactions provide a foundation for a better understanding of the mechanisms of host-microbe interactions in early branching metazoans [7, 46].

Methods

Ethics statement

This study did not include protected or endangered species and require ethical approval.

Sample collection

Specimens of *SC* and *XT* were collected from 2010 to 2013 via SCUBA at Fsar Reef (22.228408 N, 39.028187E) on the Red Sea coast of Saudi Arabia (Additional file 2). Sponge samples were collected at a depth of 13-14 m. Immediately (i.e., on board the vessel), a scalpel was used to cut the sponges into 2 to 3 cm^3 pieces, and the pieces were washed three times with autoclaved artificial

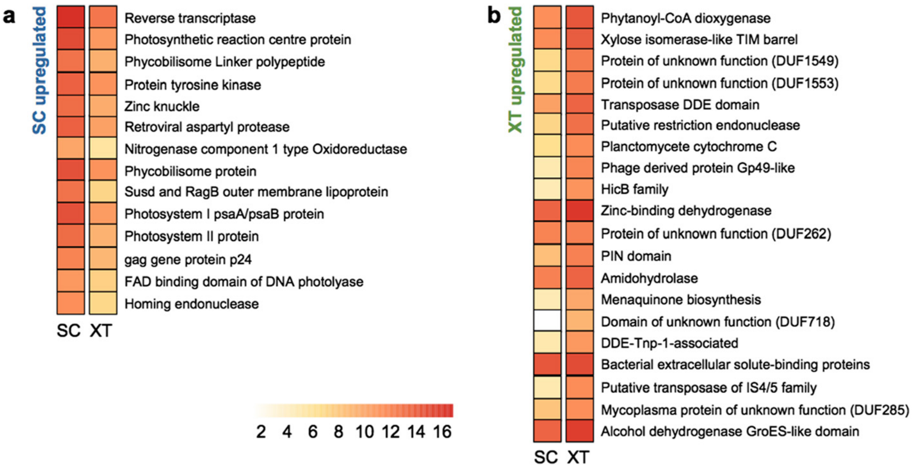


Fig. 4 Comparison of the enriched protein domains in the sponge symbionts. Significantly enriched PFAM domains among the genes found to be differentially expressed in the studied sponges (false discovery rate < 0.05) are shown for (**a**) *SC* and (**b**) *XT*, along with the expression levels of the domain-containing genes. Colors in the heatmap represent log$_2$-normalised expression levels

seawater (ASW). Thereafter, the samples were either frozen in dry ice for DNA extraction, or incubated overnight at 4 °C in RNAlater (Ambion, USA) and stored at – 80 °C for RNA extraction. These cooled samples were transported to the laboratory for experiments.

Transmission electron microscopy

The 3 mm^3 pieces of sponge were fixed with 2.5 % glutaraldehyde in seawater for ≥ 48 h, treated with reduced osmium (1:1 mixture of 2 % aqueous potassium ferrocyanide) for 1 h as described previously [47], gradually dehydrated using an ethanol series (70, 80, 90, 95, and 100 %), and embedded in Epoxy resin. Thereafter, 80 to 120 nm-thick sections were collected on copper grids and contrasted with lead citrate. Imaging was performed using a Tecnai transmission electron microscope operating at 120 kV (FEI, USA). Images were recorded on a 2 K × 4 K CCD camera (Gatan Inc., USA).

DNA extraction

Sponge tissues were ground under liquid nitrogen, and genomic DNA was extracted from 20 to 30 mg ground tissue using an All-PrepDNA kit (Qiagen, Germany). The extracted DNA was eluted with 100 μl of water, and its quality and quantity were measured using a NanoDrop 8000 spectrophotometer (Thermo Scientific, USA). To test the level of bacterial DNA in the extracted DNA, PCR was performed using the Qiagen PCR Master Mix solution (Qiagen, Germany) and two primer pairs (Bac27F, 5'-AGAGTTTGATCMTGGCTCAG-3' and Bac1492R, CGGTTACCTTGTTACGACTT; and COX1-D2, AATACTGCTTTTTTTGATCCT GCCGG and CO X1-R1 TGTTGRGGGAAAAARGTTAAATT). The cycling conditions consisted of 15 min at 95 °C, followed by 30 cycles of 95 °C for 30 s, 56 °C for 90 s and 72 °C for 90 s, and a final extension of 10 min at 72 °C. The samples were resolved by 1 % agarose gel electrophoresis. The DNA integrity was checked, and samples that showed a brighter band for bacterial DNA compared to sponge genomic DNA were excluded from further analysis.

We additionally performed whole-genome amplification of isolated sponge cells from *SC* and *XT* (Additional file 2). The sponges were cut into 0.5 cm^3 pieces, rinsed three times with cold calcium-magnesium-free (CMF)-ASW (0.55 M NaCl, 12 mM KCl, 6.3 mM Na_2S0_4, 5 mM Tris–HCl, 5 mM EDTA) at a 1:10 ratio of sponge:CMF-ASW, and agitated at 100 rpm overnight in fresh CMF-ASW. All liquid and the remaining sponge pieces were passed through a 70 μm Nitex filter (Fisher Scientific, UK), and each sample was centrifuged (700 g for 5 min at 4 °C). The pellet was washed with 10 ml of ASW, centrifuged, and suspended in 2 ml ASW. One ml of sample was gently layered atop a 30:50:70 % Percoll gradient in a 15 ml Falcon tube (VWR International, USA), and the sample-loaded gradient was centrifuged at 400 x g for 15 min at 4 °C. Each gradient layer was individually pipetted to a separate 2 ml tube and subjected to microscopic analysis. The layers representing 30:50 and 50:70 % Percoll were found to contain the most sponge cells and the fewest bacterial cells. These layers were washed with 5 ml ASW (300 x *g* for 5 min) and suspended in 200 μl 1xPBS. Micromanipulators (Narishige, Japan) were used to collect 15–30 sponge cells, which were dispensed to 3 μl sterile 1x PBS and subjected to whole-genome DNA amplification using an REPLI-g Midi kit (Qiagen, Germany). Briefly, 3.5 μl of Buffer D2 (83 mM DTT, 917 mM Reconstituted Buffer DLB) and 3 μl of cells in 1 x PBS were vortexed, briefly centrifuged, and incubated for 10 min on ice. Stop solution (3.5 μl) was added, and the sample was vortexed and then briefly centrifuged to yield denatured DNA. A master mix was made by combining 1 x SYBR Green, nuclease free H_20, REPLI-g Midi Reaction Buffer and REPLI-g Midi DNA polymerase (as per the instructions), and 50 μl of this master mix was added to 10 μl of the denatured DNA. The samples were incubated on a Real-Time PCR 7900 (Applied Biosystems, USA) at 30 °C for 16 h followed by 3 min at 65 °C (to inactivate the polymerase). Each sample was then analysed for bacterial contamination (as described above) and then stored at – 20 °C until use.

Extraction of mRNA

Total RNA was extracted as described by Moitinho-Silva et al. [33], and sponge mRNA was isolated from the total RNA (100 μg) using a Poly(A) Purist MAG kit (Ambion, USA) with two rounds of poly(A) purification. The isolated sponge mRNA was linearly amplified using a MessageAmp II-Bacteria kit (Ambion, USA) as described, except that we omitted the polyadenylation of the template RNA (which is required only for prokaryotic RNA). RNA integrity was analysed using an Experion System (Bio-Rad, USA), and the isolated sponge mRNA was stored at – 80 °C until use.

Microbiome RNA extraction

The metatranscriptome of each sponge was extracted as described by Moitinho-Silva et al. [33].

Estimation of sponge genome size

Fresh sponge tissues were rinsed three times in filtered (0.22-μm, 142-mm Express Plus filters; Millipore, USA) seawater, fixed in 95 % ethanol and stored at – 20 °C. Small pieces of ethanol-preserved sponge (0.5–1 cm^3) were subjected to two different nuclear suspension approaches, both involving the standard protocol of the CyStain® PI absolute T kit (Partec GmbH, Germany). For the first (tissue-grinder-based) approach, a piece of sponge was placed in a cryotube, incubated for 15 min in extraction buffer, and mashed with a tissue grinder for 1 min. The sample was filtered through a 40 μm nylon mesh filter, 250 μl of sample was combined with 1.25 ml (5 volumes) of staining solution (staining buffer + propidium iodide + RNase), and the mixture was incubated in the dark for 60 min. As a control, chicken erythrocytes (*Gallus gallus domesticus*, 2C = 2.45 pg) were included in the same tube and analysed in parallel with the sponge sample. For the second (bead-beating-based) approach, the sample was placed in a cryotube, incubated in extraction buffer (Partec GmbH, Germany) for 15 min, and homogenised with an MP FastPrep 24 machine (MP Biomedicals, USA) for 10–20 s (4.0 M/s, 2 ceramic beads). All samples were analysed using a BD Canto II flow cytometer (BD Biosciences, USA) with a 488 nm laser (to excite the PI) and a 585/42 band-pass emission filter.

Both methods yielded very similar genome sizes. For *SC*, the first and second protocols yielded haploid genome sizes of 0.395 pg (386.31 Mbp) and 0.39 pg (381.42 Mbp), respectively. For *XT*, both protocols yielded haploid genome sizes of 0.165 pg (161.37 Mbp).

High-throughput sequencing

Genomic DNA and RNA libraries were prepared using the TruSeq kit (Illumina, USA). For mate-pair library preparation, a Nextera kit (Illumina, USA) was used for fragmentation, size selection, and circularisation, and then a TruSeq kit was used for end repair and adapter ligation. HiSeq2000 technology (Illumina, USA) was used for paired-end and mate-pair sequencing; 454 and Ion proton sequencing were conducted using standard protocols (Additional file 2). All sequencing was performed in the KAUST Bioscience Core Lab (Saudi Arabia).

De novo assembly of sponge genomes and transcriptomes

The low-quality ends of short Illumina reads (spanning from the first base with Q-score < 20 up to the 3’ end) and sequencing adapters were trimmed. Long reads obtained from 454, Ion PGM, and Ion proton sequencing were split at each low-quality base (Q-score < 20), such that all bases in each split sequence had Q-scores ≥ 20.

The preprocessed genomic reads obtained using the different platforms were assembled with Velvet v1.2.09 [48], using *k*-mers from 55 to 75 with steps of 10. The Velvet assembly with $k = 65$ was selected, as it produced the longest scaffold N50. Transcriptomes were assembled with ABySS v.1.3.4 [49] and Trans-ABySS v.1.4.4 [50], using *k*-mers from 45 to 75 with steps of 10; these programs were chosen because benchmark tests [51, 52] showed that it yielded a higher accuracy than other *de novo* assemblers. Genomic scaffolds were further assembled using the transcriptomes, by the L_RNA_SCAFFOLDER [53]. After discarding short scaffolds (<800 bp) based on the genome annotation guideline [54], our analysis yielded 97,497 and 97,640 scaffolds for *SC* and *XT*, respectively. The statistics for our genomic and transcriptomic assemblies are summarised in Additional file 3. To obtain the base-level and mean coverages for each scaffold, we aligned the reads to the relevant scaffolds, and analysed them using BWA [55], SAMtools [56], BEDTools [57], and custom Java scripts. The mean base-level coverages of the *SC* and *XT* genomes were 109 X and 59 X, respectively. The host genome sizes for *SC* and *XT*, which were roughly estimated using scaffolds with GC % < 50, were 407.44 and 173.78 Mbp, respectively.

Gene annotation

MAKER2 was used to annotate the gene models [58]. Assembled transcriptome contigs were used as the mRNA evidence, while proteins from the *Amphimedon queenslandica* (*AQ*), CEGMA, and UniProtKB/Swiss-Prot databases were used as protein homology evidence [12, 59]. Augustus (trained with the gene model from *AQ*) and SNAP were used as *ab initio* gene predictors inside the MAKER2 pipeline [60, 61]. Gene models with an Annotation Edit Distance (AED) score ≤ 0.75 from MAKER2 were selected.

To increase the authenticity of each predicted gene model, we tagged them as eukaryotic (E), prokaryotic (P), or unknown (X). A gene was tagged as “E” if the protein product had a hit to any eukaryotic gene (*e*-value < 10^{-4}) in the NCBI non-redundant (nr) database, as assessed using Blastp [62]. A gene was tagged as “P” if it had a significant hit (*e*-value < 10^{-4}) to prokaryotic genes without any eukaryotic gene hit. A gene was tagged as “X” if it lacked any significant hit. The statistics and properties of the genes identified with each tag are summarized in Additional file 3 and Additional file 15, respectively. We used only “E” genes for our downstream analysis (26,967 and 22,337 genes for *SC* and *XT*, respectively), because they were considered to represent *bona fide* host genes.

The completeness of each assembly was measured using CEGMA v2.4 [12], which revealed that 73 and 81 % of 458 Core Eukaryotic Genes (CEGs) were completely or partially present in the genomes of *SC* and *XT*, respectively. However, as reported in Smith et al. [63], these numbers can differ depending on the utilised search algorithm. Accordingly, we also used Blastp to search 458 CEGs against the sponge gene models, setting the *e*-value threshold to 10^{-4}. This analysis indicated that 433 (94.5 %) and 432 (94.3 %) CEGs had homologs in *SC* and *XT*, respectively.

The quality of predicted gene models was assessed by comparing to those of publicly available Porifera dataset (Additional file 4). Gene models from the draft genomes were used for *AQ* [9], *SC*, and *XT*. Transcriptome assemblies of eight sponges (*Aphrocallistes vastus*, *Chondrilla nucula*, *Corticium candelabrum*, *Ircinia fasciculata*, *Petrosia ficiformis*, *Pseudospongosorites suberitoides*, *Spongilla lacustris*, and *Sycon coactum*) were retrieved from Riesgo et al. [11]. Transcript models for other sponges (*Ephydatia muelleri*, *Leucosolenia complicata*, *Oscarella carmela*, *Oscarella sp*, *Sycon ciliatum*) were retrieved from Compagen [10]. The CDSs of sponges except for *AQ*, *SC*, and *XT* were obtained by applying TransDecoder [64] and cd-hit-est [65] with default setting. Blastp [62] were performed for sponge CDSs against 458 CEGMA core gene set [12] with the threshold 10^{-4}. SUPERFAMILY domains were annotated using Interproscan v5. RC7 [66].

Functional annotation of genes

We annotated SUPERFAMILY and PFAM domains using InterProScan v5. RC7 [66]. The gene ontology (GO) terms were assigned to proteins harboring SUPERFAMILY and PFAM domains using dcGO [67] and InterProScan, respectively. Blast searches of the predicted sponge proteins were performed against the NCBI nr database, and homologs were identified with an *e*-value threshold of 10^{-4}. The GO terms of the identified homologs were retrieved from the NCBI database and transferred to sponge genes using a custom Python script. Whole-genome over/under-representations of GO terms were ranked using Z-scores calculated from a background distribution generated for each annotated GO term (composed of dcGO results from 382 species found in the SUPERFAMILY library as of June 1, 2014). Due to redundancy among the SUPERFAMILY and PFAM domains and the more comprehensive functional annotation of the former by dcGO, we used the SUPERFAMILY domains for our analysis of the sponge domain repertoire.

Analysis of SRCR-like domains

The peptide sequences of the SRCR-like domains from *AQ*, *SC*, and *XT* were queried against each other using Blastp [62]. A threshold of ≥ 90 % positive-scoring matches between two domains was used to identify homology. The Markov Cluster (MCL) Algorithm [68] was used to cluster the SRCR-like domains, with the Blastp bit score applied as a similarity metric. The SRCR-like domain sequences from each cluster were aligned using MAFFT v7.123b [69], with the extension penalty parameter and maximum iterations set to 0.123 and 3, respectively. We computed the amino acid frequency at each aligned position using a custom Python script.

Microbial community analysis

For Illumina reads, the low-quality ends (from the first base with Q-score < 20, which correspond to an error probability of 0.01, to the 3' end) and sequencing adapters were trimmed using custom Java scripts.

The preprocessed metatranscriptome reads were further processed to remove any rRNA fragments, using riboPicker v0.4.3 [70] with thresholds of 90 % alignment coverage and 90 % alignment identity. Blastx [62] was then used to align the reads against the nr database to obtain the best-hit sequence for each aligned read. To create each reference sequence, we measured the similarities between extracted sequences using Blastp, and clustered them into homology groups using MCL [68] with the inflation parameter set to 3.6 and the other parameters set at their default values.

To quantify the expression level of each homology group per sample per sponge, we summed the numbers of reads whose best hits were assigned to each homology group, then further quantile-normalised the read counts across samples using the preprocessCore package in R [71]. The differentially expressed homology groups between two sponges were obtained using GFOLD v1.1.2 [72], with an expected false discovery rate (FDR) ≤ 0.05.

Representative sequences of each homology group were annotated with respect to PFAM [73] domains using InterProScan v5. RC7 [66]. The GO terms for each domain were also obtained [74]. The statistical significance of each domain and the GO term enrichments observed among the differentially expressed homology groups were assessed based on the cumulative hypergeometric distributions and FDRs (≤0.05), which were calculated with a custom R script. Fourteen and 20 PFAM domains were found to be statistically significant in the *SC* and *XT* metatranscriptome datasets, respectively (Fig. 4). SUPERFAMILY [13] domains were annotated in the same way (Additional file 14).

Availability of supporting data

The generated sequencing datasets for *SC* and *XT* are publicly available under NCBI BioProject IDs PRJNA254402 and PRJNA254412, respectively. The genome assemblies, transcripts, and coding sequences for both sponges are available at http://sc.reefgenomics.org and http://xt.reefgenomics.org.

Additional files

Additional file 1: Transmission electron microscope (TEM) images of studied sponges. (a–d) TEM images of *Stylissa carteri* (*SC*). A number of the *SC* cells are packed with vesicle-like inclusions. Microbes are not observed in the mesohyl. Spongin (spincy lines), which gives structure to the sponge tissues, and choanocytes are observed. (e–j) TEM images of *Xestospongia testudinaria* (*XT*). Archaeocytes are seen to be engulfing bacteria for digestion. Unique sponge symbionts, such as cyanobacteria with thylakoid membranes, are frequently observed. Spirochaetes are also observed (arrow in j). Abbreviations: ECM, extracellular matrix; HSC, host sponge cell; ST, storage cell; n, nucleus; b, bacterium; ub, undigested bacterium; db, digested bacterium; and f, flagella. (PDF 7354 kb)

Additional file 2: Sources and statistics of sequences used for the hologenome analysis. (XLSX 19 kb)

Additional file 3: Detailed statistics of our sponge genomes and transcriptomes. (PDF 263 kb)

Additional file 4: Comparison of gene annotation quality among sponge dataset. Transcript models of 16 sponge species are compared to address the quality of gene annotation. (PDF 60 kb)

Additional file 5: Over- or under-represented Superfamily domains in *SC* and *XT*. Superfamily domains that are unusually enriched in *SC* and *XT*. Domains with |deviation| > 0.5 are shown. (PDF 64 kb)

Additional file 6: Superfamily domains related to innate immunity and symbiosis. Sixteen species were compared; their NCBI taxonomic IDs are given in brackets. (XLSX 72 kb)

Additional file 7: Supplementary information describing relationship between *Xestospongia testudinaria* and *Xestospongia muta*, antimicrobial peptides, and evolutionary rates of innate immune domains in analyzed sponge genomes. (PDF 790 kb)

Additional file 8: Statistics of compiled AMPs for broad taxonomic group. (PDF 34 kb)

Additional file 9: Taxonomic origins of the AMPs that were successfully aligned to the sponge genomes. (PDF 47 kb)

Additional file 10: Antimicrobial peptides encoded in the sponge genomes. (PDF 78 kb)

Additional file 11: The number of AMPs with different biological activities on the sponge genomes. (PDF 56 kb)

Additional file 12: Evolutionary rates of selected domains. (a) Boxplot representing the distribution of the mean Ka/Ks for each protein domain between sponge pairs. (b) The mean Ka/Ks of each protein domain is shown. This analysis was restricted to the protein domains given in Fig. 2 that passed our quality control step (see Methods). (PDF 194 kb)

Additional file 13: Expression levels of protein domains in the metatranscriptome dataset. PFAM and SUPERFAMILY domains are ranked by their expression levels, which were quantile-normalised and then summed. (XLSX 462 kb)

Additional file 14: SUPERFAMILY domains enriched among the genes found to be differentially expressed in each sponge metatranscriptome. (PDF 50 kb)

Additional file 15: Properties of gene models. Gene models were tagged based on the presence of eukaryotic or prokaryotic sequences, as assessed by comparison to the NCBI nr database (see Methods). The properties of the gene models were analysed based on the numbers of exons and the transcriptional expression levels. (PDF 481 kb)

Abbreviations

AED: Annotation edit distance; AQ: *Amphimedon queenslandica*; ASW: Artificial seawater; CEGs: Core eukaryotic genes; CMF: Calcium-magnesium-free; DAMPs: Endogenous danger-associated molecular patterns; GO: Gene ontology; HMA: High microbial abundance; HMG: High-mobility group; LDL: Low-density lipoprotein; LMA: Low microbial abundance; MAMPs: Microbe-associated molecular patterns; NLR: Leucine-rich repeat; PAMPs: Pathogen-associated molecular patterns; SC: *Stylissa carteri*; SRCR: Scavenger receptor cysteine-rich; XT: *Xestospongia testudinaria*.

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

TR1 (Taewoo Ryu) and TR2 (Timothy Ravasi) conceived the overall study. UH and MLB identified the Red Sea sponges. TR1, RS, LMS, and UH performed the TEM analysis. MK, SB, CRV, and MA performed the genome-size analysis. LMS, AH, FFL, EG, and CS purified the DNA and RNA. TR1, LMS, LS, HM, and UH analysed the microbial components. TR1, SF, and LS performed the sequence assembly. TR1 annotated the gene models. TR1 and MO annotated the gene functions. YJL and MA constructed the website. TR1, MO, and JG performed the domain enrichment analysis. TR1, LS, YG, and MAA performed the evolutionary analysis. TR1, XW, and XG performed the antimicrobial peptide analysis. HM contributed to preparing the figures. TR1 integrated the results and wrote the manuscript. TR2, UH, CRV, and LMS contributed to the data interpretation of data and writing of the manuscript. TR2 supervised the project. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the King Abdullah University of Science and Technology. LMS was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Wuerzburg. We thank the KAUST Core Lab for performing the sequencing and assisting with the genome-size analysis. We acknowledge Till Rothig (KAUST) for contributing to the sponge sample collection, Christine Gernert (University of Wurzburg) for assisting with the TEM analysis, and Sebastian Fraune (University of Kiel) for reading the manuscript critically. We also thank several anonymous reviewers for constructive comments.

Author details

[1]KAUST Environmental Epigenetic Program (KEEP), King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [2]Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [3]School of Biotechnology and Biomolecular Sciences & Centre for Marine Bio-Innovation, University of New South Wales Sydney, Sydney, Australia. [4]Department of Computer Science, University of Bristol, 24 Tyndall Ave, Bristol, UK. [5]Red Sea Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [6]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [7]Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [8]Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [9]Imaging and characterization Lab, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [10]Division of Evolution, Ecology and Genetics, Research School of Biology, The Australian National University, Canberra ACT 2601, Australia. [11]GEOMAR Helmholtz Centre for Ocean Research, RD3 Marine Microbiology and Christian-Albrechts University of Kiel, Düsternbrooker Weg 20, D-24105 Kiel, Germany. [12]Present address: APEC Climate Center, Busan 48058, South Korea.

Received: 27 September 2015 Accepted: 18 February 2016
Published online: 29 February 2016

References

1. McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-Loso T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, et al. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci U S A. 2013;110(9):3229–36.
2. Bosch TC, McFall-Ngai MJ. Metaorganisms as the new frontier. Zoology (Jena). 2011;114(4):185–90.
3. Zilber-Rosenberg I, Rosenberg E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiol Rev. 2008;32(5):723–35.
4. Li CW, Chen JY, Hua TE. Precambrian sponges with cellular structures. Science. 1998;279(5352):879–82.
5. Bergquist PR. Sponges. Berkeley: University of California Press; 1978.
6. Taylor MW, Radax R, Steger D, Wagner M. Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. Microbiol Mol Biol Rev. 2007;71(2):295–347.
7. Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. Nat Rev Microbiol. 2012;10(9):641–54.
8. Gloeckner V, Wehrl M, Moitinho-Silva L, Gernert C, Schupp P, Pawlik JR, Lindquist NL, Erpenbeck D, Worheide G, Hentschel U. The HMA-LMA dichotomy revisited: an electron microscopical survey of 56 sponge species. Biol Bull. 2014;227(1):78–88.
9. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature. 2010;466(7307):720–6.
10. Hemmrich G, Bosch TC. Compagen, a comparative genomics platform for early branching metazoan animals, reveals early origins of genes regulating stem-cell differentiation. BioEssays. 2008;30(10):1010–8.
11. Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP. The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. Mol Biol Evol. 2014;31(5):1102–20.

12. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7.
13. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001;313(4):903–19.
14. Yuen B, Bayes JM, Degnan SM. The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. Mol Biol Evol. 2014;31(1):106–20.
15. Martinez VG, Moestrup SK, Holmskov U, Mollenhauer J, Lozano F. The conserved scavenger receptor cysteine-rich superfamily in therapy and diagnosis. Pharmacol Rev. 2011;63(4):967–1000.
16. Steindler L, Schuster S, Ilan M, Avni A, Cerrano C, Beer S. Differential gene expression in a marine sponge in relation to its symbiotic state. Mar Biotechnol. 2007;9(5):543–9.
17. Springer TA. An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components. J Mol Biol. 1998;283(4):837–62.
18. Maness PF, Schachner M. Neural recognition molecules of the immunoglobulin superfamily: signaling transducers of axon guidance and neuronal migration. Nat Neurosci. 2007;10(1):19–26.
19. Levy O, Martin S, Eichenwald E, Ganz T, Valore E, Carroll SF, Lee K, Goldmann D, Thorne GM. Impaired innate immunity in the newborn: newborn neutrophils are deficient in bactericidal/permeability-increasing protein. Pediatrics. 1999;104(6):1327–33.
20. Nupponen I, Turunen R, Nevalainen T, Peuravuori H, Pohjavuori M, Repo H, Andersson S. Extracellular release of bactericidal/permeability-increasing protein in newborn infants. Pediatr Res. 2002;51(6):670–4.
21. Bianchi ME. DAMPs, PAMPs and alarmins: all we need to know about danger. J Leukoc Biol. 2007;81(1):1–5.
22. Pisetsky DS, Erlandsson-Harris H, Andersson U. High-mobility group box protein 1 (HMGB1): an alarmin mediating the pathogenesis of rheumatic disease. Arthritis Res Ther. 2008;10(3):209.
23. Yang D, Tewary P, de la Rosa G, Wei F, Oppenheim JJ. The alarmin functions of high-mobility group proteins. Biochim Biophys Acta. 2010;1799(1–2):157–63.
24. Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, Agata K, Bosch TCG. The innate immune repertoire in Cnidaria - ancestral complexity and stochastic gene loss. Genome Biol. 2007;8(4):R59.
25. Wiens M, Korzhev M, Krasko A, Thakur NL, Perovic-Ottstadt S, Breter HJ, Ushijima H, Diehl-Seifert B, Muller IM, Muller WE. Innate immune defense of the sponge Suberites domuncula against bacteria involves a MyD88-dependent signaling pathway. Induction of a perforin-like molecule. J Biol Chem. 2005;280(30):27949–59.
26. Degrandi D, Konermann C, Beuter-Gunia C, Kresse A, Wurthner J, Kurig S, Beer S, Pfeffer K. Extensive characterization of IFN-induced GTPases mGBP1 to mGBP10 involved in host defense. J Immunol. 2007;179(11):7729–40.
27. Rivers RL, Dean RM, Chandy G, Hall JE, Roberts DM, Zeidel ML. Functional analysis of nodulin 26, an aquaporin in soybean root nodule symbiosomes. J Biol Chem. 1997;272(26):16256–61.
28. Kaldenhoff R, Fischer M. Aquaporins in plants. Acta Physiol (Oxf). 2006; 187(1–2):169–76.
29. Muller WE, Belikov SI, Kaluzhnaya OV, Chernogor L, Krasko A, Schroder HC. Symbiotic interaction between dinoflagellates and the demosponge Lubomirskia baicalensis: aquaporin-mediated glycerol transport. Prog Mol Subcell Biol. 2009;47:145–70.
30. Lehnert EM, Mouchka ME, Burriesci MS, Gallo ND, Schwarz JA, Pringle JR. Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. G3 (Bethesda). 2014;4(2):277–95.
31. Moitinho-Silva L, Bayer K, Cannistraci CV, Giles EC, Ryu T, Seridi L, Ravasi T, Hentschel U. Specificity and transcriptional activity of microbiota associated with low and high microbial abundance sponges from the Red Sea. Mol Ecol. 2014;23(6):1348–63.
32. Lee OO, Wang Y, Yang J, Lafi FF, Al-Suwailem A, Qian PY. Pyrosequencing reveals highly diverse and species-specific microbial communities in sponges from the Red Sea. Isme J. 2011;5(4):650–64.
33. Moitinho-Silva L, Seridi L, Ryu T, Voolstra CR, Ravasi T, Hentschel U. Revealing microbial functional activities in the Red Sea sponge Stylissa carteri by metatranscriptomics. Environ Microbiol. 2014;16(12):3683–98.
34. Schmitt S, Tsai P, Bell J, Fromont J, Ilan M, Lindquist N, Perez T, Rodrigo A, Schupp PJ, Vacelet J, et al. Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. Isme J. 2012;6(3):564–76.
35. Webster NS, Taylor MW, Behnam F, Lucker S, Rattei T, Whalan S, Horn M, Wagner M. Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. Environ Microbiol. 2010;12(8):2070–82.
36. Toft C, Andersson SG. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat Rev Genet. 2010;11(7):465–75.
37. Fan L, Reynolds D, Liu M, Stark M, Kjelleberg S, Webster NS, Thomas T. Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. Proc Natl Acad Sci U S A. 2012;109(27):E1878–1887.
38. Bright M, Bulgheresi S. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 2010;8(3):218–30.
39. Omrane S, Ferrarini A, D'Apuzzo E, Rogato A, Delledonne M, Chiurazzi M. Symbiotic competence in Lotus japonicus is affected by plant nitrogen status: transcriptomic identification of genes affected by a new signalling pathway. New Phytol. 2009;183(2):380–94.
40. Isanapong J, Sealy Hambright W, Willis AG, Boonmee A, Callister SJ, Burnum KE, Pasa-Tolic L, Nicora CD, Wertz JT, Schmidt TM, et al. Development of an ecophysiological model for Diplosphaera colotermitum TAV2, a termite hindgut Verrucomicrobium. Isme J. 2013;7(9):1803–13.
41. Mhlanga-Mutangadura T, Morlin G, Smith AL, Eisenstark A, Golomb M. Evolution of the major pilus gene cluster of Haemophilus influenzae. J Bacteriol. 1998;180(17):4693–703.
42. Arcus VL, Rainey PB, Turner SJ. The PIN-domain toxin-antitoxin array in mycobacteria. Trends Microbiol. 2005;13(8):360–5.
43. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 2012; 40(Database issue):D306–312.
44. Thomas T, Rusch D, DeMaere MZ, Yung PY, Lewis M, Halpern A, Heidelberg KB, Egan S, Steinberg PD, Kjelleberg S. Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. Isme J. 2010;4(12):1557–67.
45. Hooper SD, Mavromatis K, Kyrpides NC. Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol. 2009;10(4):R45.
46. Sachs JL, Essenberg CJ, Turcotte MM. New paradigms for the evolution of beneficial infections. Trends Ecol Evol. 2011;26(4):202–9.
47. Karnovsky M. Use of ferrocyanide-reduced osmium tetroxide in electron microscopy, Proceedings of the 11th Annual Meeting American Society for Cell Biology New Orleans, Louisiana. 1971.
48. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
49. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117–23.
50. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909–12.
51. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC bioinformatics. 2011;12(14):S2.
52. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 2013;14:328.
53. Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, Sun XW. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 2013;14:604.
54. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329–42.
55. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
58. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 2011;12:491.
59. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol. 2007;406:89–112.
60. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33(Web Server issue):W465–467.
61. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004;5:59.

62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
63. Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, et al. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet. 2013;45(4):415–21. 421e411-412.
64. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
65. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
66. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
67. Fang H, Gough J. DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. 2013; 41(Database issue):D536–544.
68. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30(7):1575–84.
69. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
70. Schmieder R, Lim YW, Edwards R. Identification and removal of ribosomal RNA sequences from metatranscriptomes. Bioinformatics. 2012;28(3):433–5.
71. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
72. Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics. 2012;28(21):2782–8.
73. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al. The Pfam protein families database. Nucleic Acids Res. 2010;38(Database issue):D211–222.
74. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central