DR. MANUEL ARANDA LASTRA (Orcid ID : 0000-0001-6673-016X)

Article type   : Resource Article

Draft genomes of the corallimorpharians *Amplexidiscus fenestrafer* and *Discosoma* sp.

Running title: **Draft genomes of two Corallimorpharia**

Xin Wang[1], Yi Jin Liew[1], Yong Li[1], Didier Zoccola[2], Sylvie Tambutte[2], Manuel Aranda[1,*]

[1]*King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Biological and Environmental Sciences & Engineering Division (BESE), Thuwal, 23955-6900, Saudi Arabia*

[2]*Centre Scientifique de Monaco, 8 quai Antoine Ier, Monaco, 98000, Monaco*

*Correspondence: Manuel Aranda Lastra, Red Sea Research Center, King Abdullah University of Science and Technology, 4799 KAUST, Thuwal 23955, Saudi Arabia

Phone: +966 544 700 661

E-mail: manuel.aranda@kaust.edu.sa

Keywords: Corallimorpharia, naked corals, *Amplexidiscus fenestrafer*, *Discosoma* sp.

Abstract

Corallimorpharia are the closest non-calcifying relatives of reef-building corals. Aside from their popularity among aquarium hobbyists, their evolutionary position between the Actiniaria (sea anemones) and the Scleractinia (hard corals) makes them ideal candidates for comparative studies aiming at understanding the evolution of hexacorallian orders in general and reef-building corals in particular. Here we have sequenced and assembled two draft genomes for the Corallimorpharia species *Amplexidiscus fenestrafer* and *Discosoma* sp.. The draft genomes encompass 370 Mbp and 445 Mbp respectively and encode for 21,372 and 23,199 genes. To facilitate future studies using these resources, we provide annotations for the predicted gene models—not only at gene level, by annotating gene models with the function of the best-matching homolog, and GO terms when available; but also at protein domain level, where gene function can be better verified through the conservation of the sequence and order of protein domains. Further, we provide an online platform (http://corallimorpharia.reefgenomics.org), which includes a BLAST interface as well as a genome browser to facilitate the use of these resources. We believe that these two genomes are important resources for future studies on hexacorallian systematics and the evolutionary basis of their specific traits such as the symbiotic relationship with dinoflagellates of the genus *Symbiodinium* or the evolution of calcification in reef-building corals.

Introduction

Corallimorpharians, also known as “false” corals, are closely related to reef building corals of the order Scleractinia but lack the ability to form calcium carbonate skeletons. They are widely distributed, and preferentially live in shallow oceans with weak currents (Vandepitte *et al.* 2013). Their vibrant appearance, in addition to their relative ease of maintenance in salt-water aquarium systems, is the primary reason for their popularity among reef aquarium enthusiasts (Murray & Watson 2014).

For the scientific community, corallimorpharians are especially interesting due to their close phylogenetic relationship to corals and their presumed origin from a scleractinian ancestor as proposed by the naked coral hypothesis (Medina *et al.* 2006; Stanley & Fautin 2001). They belong to the subclass Hexacorallia, a monophyletic group that also contains all reef-building corals (Scleractinia) and sea anemones (Actiniaria) among others (Daly *et al.* 2003). While the availability of the genomes of the sea anemones *Aiptasia pallida* (Baumgarten *et al.* 2015), *Nematostella vectensis* (Putnam *et al.* 2007) and the coral *Acropora digitifera* (Shinzato *et al.* 2011) provided valuable insights into the genomic basis of Hexacorallian traits, our understanding of the evolution of more specific traits, such as calcification in reef-building corals, is still hampered by the large evolutionary gap between Actiniaria and Scleractinia which split more than > 500 Mya (Medina *et al.* 2006; Park *et al.* 2012; Schwentner & Bosch 2015; Simpson *et al.* 2011; Stolarski *et al.* 2011). Here, we sequenced the genomes of *Amplexidiscus fenestrafer* and *Discosoma* sp., two members of the order Corallimorpharia that represents the closest non-calcifying relatives of reef-building corals. These genomes constitute important genomic resources for future comparative studies that contrast the individual orders within Hexacorallia, thus providing a better insight into the distinct evolutionary histories and innovations of these organisms.

Materials and Methods

DNA, RNA extraction and sequencing

Samples of *Amplexidiscus fenestrafer* and *Discosoma* sp. were maintained at the Centre Scientifique de Monaco in aquaria supplied with flowing seawater from the Mediterranean Sea (exchange rate 2% h^{-1}) at a salinity of 38.2 ppt, pH 8.1 ± 0.1 under an irradiance of 300 µmol photons $m^{-2}s^{-1}$ at 25 ± 0.5 °C. Individuals were fed 3 times a week with both frozen krill and live *Artemia salina* nauplii. DNA for sequencing libraries was extracted from *Amplexidiscus fenestrafer* and *Discosoma sp*. tissue samples using a nuclei isolation approach to minimize contamination with symbiont DNA. Briefly, cells were harvested using a Water Pick in 50 ml of 0.2 M EDTA solution refrigerated at 4 °C. Extracts were passed sequentially through a 100 µm and a 40 µm cell strainer (Falcon®) to eliminate most of the Symbiodinium. Then extracts were centrifuged at 2,000g for 10 min at 4 °C. The supernatant was discarded and the resulting pellets were homogenized in lysis buffer (G2) of the Qiagen Genomic DNA isolation kit (Qiagen, Hilden, Germany). The DNA was extracted following manufacturer instructions using genomic-tip 100/G. DNA concentration was determined by O.D. with an Epoch Microplate Spectrophotometer (BioTek, Winooski, VT, USA). Contamination with Symbiodinium DNA was assessed via PCR targeting the multicopy gene RuBisCO (Genbank accession number AY996050). Total RNA extraction was performed as described previously for corals (Moya *et al.* 2008) Briefly, total tissue from individuals maintained at the above described culture conditions was snap frozen in liquid nitrogen and ground into powder in a cryogrinder (Freezer/Mill 6770, Spex Sample Prep®). Total RNA was subsequently extracted with Trizol® Reagent (Invitrogen) and quantified on a Bioanalyzer 2100 (Agilent, Santa Clara, CA). Strand specific sequencing libraries were generated using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA).

Sequencing libraries were prepared using the Illumina TruSeq DNA kits for paired-end or mate-pair libraries respectively according to the manufacturer's instructions. A total of 2 paired-end and 4 mate-pair libraries were generated for each species and sequenced on the Illumina HiSeq 2000 platform at the KAUST Bioscience Core Facility. All data were uploaded to NCBI and are available under Bioproject IDs: PRJNA354436 (*A. fenestrafer*) and PRJNA354492 (*Discosoma* sp.).

Sequence filtering and genome assembly

A total of 1,199 and 1,411 million raw paired-end reads were sequenced for *A. fenestrafer* and *Discosoma sp.* respectively. These included 621 and 815 million paired-end as well as 578 and 596 million mate-pair reads respectively. All reads were processed to remove low quality reads (< Q30 and < 30 bp), adapter sequences and duplicated reads using Trimmomatic v0.32 (Bolger *et al.* 2014) (Table S1, Supporting Information). Mate paired reads were additionally evaluated by mapping against the assembled genomes using Bowtie2 v2.1.0 (Langmead & Salzberg 2012) (Fig. S1, Supporting Information). Potential contamination from *Symbiodinium* sources was removed by mapping all paired-end reads to the *S. minutum* (Shoguchi *et al.* 2013) and *S. microadriaticum* (Aranda et al 2016) genomes via Bowtie2 v2.1.0 (Langmead & Salzberg 2012) using default parameters. Mate-pair reads were not filtered for contamination due to low mapping rates to both *Symbiodinium* genomes (< 1%, Table S2, Supporting Information).

Genome sizes for *A. fenestrafer* and *Discosoma* sp. were estimated using Jellyfish (Marçais & Kingsford 2011) and KmerFreq_AR (Luo *et al.* 2012). For both genomes, the frequency distributions of the *k*-mers revealed the presence of many *k*-mers with low coverage (Table S3, Supporting Information). To reduce this considerable sequence heterogeneity a three-pass digital normalization approach was applied, as described in Khmer v1.4 (Crusoe *et al.* 2015). Briefly, the paired-end libraries were initially normalized to a *k*-mer coverage of 20 (-C 20, -x 4e9) followed by removal of low abundant *k*-mers and a final normalization to a coverage of 10 (-C 10, -x 4e9). Construction and scaffolding of contigs were carried using ALLPATH-LG (Butler *et al.* 2008) with HAPLOIDIFY and estimated genome sizes of 400Mbp. Ambiguous and erroneous bases were identified through mapping of filtered pre-normalized paired-end against the assembled genome using Bowtie2 v2.1.0 (Langmead & Salzberg 2012) and subsequently corrected using SAMtools v1.2 (Li *et al.* 2009) and a custom Perl script. Finally, gaps were filled within the genome scaffolds with GapCloser v1.12 (Luo *et al.* 2012) using stringent parameters (-l 125, -p 31) and all trimmed, filtered pre-normalized paired-end and mate pair reads. Assembly statistics were assessed as previously described (Bradnam *et al.* 2013). In order to identify and remove scaffolds originating from bacterial or virus contaminants, BLASTN searches (e-value 10^{-5}) were carried out against two bacterial databases (NCBI complete and draft bacterial genomes) and a viral database (NCBI complete viral genomes) (Pruitt *et al.* 2007). Scaffolds with high degree of contaminating sequences (coverage > 50%, bitscore > 1000 and e-value < 10^{-20}) were subsequently removed from the assemblies.

Identification and classification of repeats and transposable elements

A combination of homology and *ab initio*-based methods was used to identify interspersed repeats and low complexity DNA sequences. To identify repeat boundaries and construct species-specific consensus models, *de novo* identification was carried out with RMBlast as implemented in RepeatModeler v1.0.8 (http://www.repeatmasker.org/RepeatModeler.html). Subsequently, RepeatScout, TRF and Recon were used to identify *de novo* repeats, which produced a consensus library file. RepeatMasker was then run using this high-quality custom repeat library. Subsequently, RepBase v20.02 (Bao *et al.* 2015) and Dfam v1.4 (Wheeler *et al.* 2013) were used to identify and classify different categories of repetitive elements. A hierarchical system was utilized to classify *de novo* repeat elements into five different categories: DNA transposon, LTR, SINEs, LINEs, and simple repeats. Based on the different categories of repeat elements, coverage distributions and repeat divergences for each species were calculated. A Perl script that extracted results from RepeatMasker was written to calculate the percentage of nucleotide divergence of each transposable element to the consensus sequence from the respective genomes (Fig. S2a, b, Supporting Information).

Gene prediction

After repeat masking, a combination of *ab initio*, homology- and expression- based prediction strategies was used to produce the final set of gene models. To generate draft transcriptomes for our gene prediction pipeline, high-quality RNA from both organisms was extracted to produce cDNA sequencing libraries. These libraries (insert size of 180 bp) were sequenced on the Illumina HiSeq2000 platform with a paired-end read length of 101 bp. A total of 52,772,546 and 37,876,678 paired-end reads were sequenced for *A. fenestrafer* and *Discosoma* sp. respectively, of which 52,771,850 and 37,875,788 passed filtering and trimming at Q20. Filtered RNA-seq reads were used for *de novo* assemblies of the *A. fenestrafer* and *Discosoma* sp. reference transcriptomes (Table S4, Supporting Information) using Trinity v2.02 (Evans *et al.* 2012), and PASA (Haas *et al.* 2003) was used to align the newly assembled transcripts to the respective genomes (Table S5, Supporting Information). Full-length transcripts, defined to be transcripts containing start and stop codons, 5' and 3' untranslated regions (UTRs) and a minimum of two exons (4,539 *A. fenestrafer* and 3,739 *Discosoma* sp. passed all criteria) were identified and used as the training set for three *ab initio* gene prediction pipelines: Augustus v3.0.3 (Stanke *et al.* 2006), GlimmerHMM v3.02 (Kelley *et al.* 2012) and SNAP v2013-02-16 (Korf 2004). To refine exon-intron boundaries via Spaln v2.1.4 (Gotoh 2008), gene models were also aligned against protein sequences in UniProt and other cnidarians, i.e. *Nematostella vectensis* (Putnam *et al.* 2007), *Aiptasia pallida* (Baumgarten *et al.* 2015), *Acropora digitifera* (Shinzato *et al.* 2011) and *Hydra magnipapillata* (Chapman *et al.* 2010). The final consensus gene models were generated using EVM (Haas *et al.* 2008), which incorporated the various methods of gene predictions. Gene models with more than two stop codons or < 20 bp in length were subsequently removed. tRNAs were predicted via tRNAscan-SE v1.3.1 (Lowe & Eddy 1997), while rRNAs were identified by aligning rRNA sequences from Rfam v12.0 database (Nawrocki *et al.* 2014) against the genomes using BLASTN with cutoffs of e-value < 10^{-5}, identity > 85% and match length > 50 bp (Table S6, Supporting Information).

Gene functional annotation

To infer putative functions of protein coding genes, gene models were subjected to a successive BLASTP search against SwissProt, TrEMBL and nr databases as previously described (Liew *et al.* 2014). Domain annotations were derived from different databases including Pfam, PRINTS, PROSITE, ProDom, and SMART using InterProScan (Zdobnov & Apweiler 2001) (Table S7). Gene Ontology (GO) (Fig. S3, Supporting Information) annotations were obtained from BLASTP results against SwissProt and TrEMBL. Additional functional information was derived via pathway analysis based on homology to characterized pathways in Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Moriya *et al.* 2007). Detailed annotation files can be retrieved from http://corallimorpharia.reefgenomics.org. To ensure the consistency of the annotations, the same annotation pipeline was carried out on the protein coding gene sets from the other anthozoans used in the subsequent comparative analysis. To assess potential protein domain expansions or contractions in the corallimorpharian lineage, Fisher's exact tests ($p < 0.001$) were conducted on the relative Pfam domain counts, comparing in-group (*A. fenestrafer* and *Discosoma* sp.) and outgroup (all other cnidarians in the analysis). The resulting p-values were corrected using false discovery rate (FDR) (Benjamini & Hochberg 1995). To visualize these significantly enriched Pfam domains, z-scores were calculated and plotted in a heatmap using gplots (in R) (Warnes 2016).

Validation of assemblies and gene models

The completeness and accuracy of our assembled genomes was assessed by mapping all filtered, normalized paired-end libraries to the final genome assemblies using Bowtie v2.1.0 (Li & Durbin 2009) with default parameters. Coverage and mapping rates were calculated using SAMtools and downstream Perl scripts. As assemblies of highly heterogeneous organisms have been shown to be prone to duplicate contig artefacts from unresolved bubbles in the de Bruijn graphs (Kajitani *et al.* 2014), within-species BLASTN analyses of the assembled contigs were carried out to assess the contribution of such erroneously duplicated contigs to our assemblies. Briefly, best matches between two unique contigs were identified using BLASTN. The resulting congruencies were tallied for each contig to produce a plot depicting the bitscore ratios of the second best hits to the first best hits of each contig in order to evaluate the ratio of putative artificially duplicated contigs. The general completeness of our assemblies was assessed using CEGMA (Parra *et al.* 2007). In order to estimate the overall accuracy of the genome annotations, genic properties e.g. mean GC%, gene length, exon length, and number of exons were calculated via Perl and R scripts, and contrasted against the *N. vectensis, A. pallida, A*. *digitifera*, and *H. magnipapillata* genome annotations.

Gene family comparisons and phylogenetic analysis

Based on the gene models of *N. vectensis, A. pallida, A*. *digitifera* and *H. magnipapillata*, an "all-against-all" BLASTP analysis (e-value < 10^{-5}) was performed including OrthoMCL DB (http://orthomcl.cbil.upenn.edu) (Chen *et al.* 2006). OrthoMCL v2.0.9 (Li *et al.* 2003) was used with default parameters to assign proteins into orthologous gene groups. Gene families were modelled from orthologous genes using CAFE (De Bie *et al.* 2006), and functionally annotated with Pfam and BLAST annotation results. Only gene families with more than 30% of proteins sharing the same Pfam domain were retained.

For the phylogenetic analyses, ortholog groups with a single ortholog per species and at least one common Pfam domain were selected, which retained 696 high confidence single copy orthologs (Table S8, Supporting Information). The amino acid sequences for these orthologs were aligned with MUSCLE v3.8.31 (Edgar 2004) at default settings, and poorly-aligned regions were trimmed using Gblock (Talavera & Castresana 2007). The trimmed alignments were concatenated, providing a supermatrix of 115,440 amino-acid positions that was analysed using Prottest v2.4 (Darriba *et al.* 2011) to determine the best evolutionary model (LG+I+G). RAxML v8.1.22 (Stamatakis 2014) was used to reconstruct the maximum-likelihood tree (PROTGAMMAILGF, 1,000 bootstrap replicates, rapid hill-climbing model) while MrBayes v3.2.6 (Ronquist *et al.* 2012) was used for the Bayesian inference of phylogeny (LG+I+G), 4 runs with 4 chains for 5,000,000 generations).

Estimation of duplication rates

To identify duplicated genomic loci, both genomes were aligned against themselves using MUMmer v3.23 (Kurtz *et al.* 2004) (length > 3,000 bp, identity ≥ 85%). To determine putative functions associated with these duplicated regions, we analysed the functional annotation of genes located in these regions.

Results

Validation of assembled genomes

We generated a total of 124 Gbp (334x) and 142 Gbp (319x) of sequence data for *A. fenestrafer* and *Discosoma* sp. respectively using the Illumina HiSeq2000 platform (Table S1, Supporting Information). After quality trimming, filtering and digital normalization, we retained approximately 81.16x and 149.96x coverage for the final assemblies respectively, where a larger proportion of mate pair reads were filtered for *A. fenestrafer* than *Discosoma* sp.. The assembled genome sizes for *A. fenestrafer* (370 Mbp) and *Discosoma* sp. (445 Mbp) were in close agreement to the *k*-mer based estimates of 350 Mbp and 428 Mbp respectively (Table 1, Table S3, Supporting Information). After gap-filling with Gapcloser, the *A. fenestrafer* genome had contig and scaffold N50s of 20 kbp and 510 kbp respectively; for *Discosoma* sp., contig and scaffold N50s were 19 kbp and 770 kbp respectively (Table 1).

To validate our assembled genomes, one library of paired-end data for each corallimorpharian was back-mapped to the assembly. 76.49% (*A. fenestrafer*) and 86.28% (*Discosoma* sp.) of the Illumina paired-end reads could be mapped concordantly. For both genomes, ≥ 99% of the sequence had coverage of at least 2x (Table S9, Supporting Information). To assess the extent of contig-level duplication we performed a BLASTN search for all contigs against themselves for both species. Comparison to other available anthozoan genomes showed less duplicated contigs for both corallimorpharian genomes (Fig. S4, Supporting Information). Further, we found less duplication in *A. fenestrafer* than *Discosoma* sp*.,* which possibly contributes to the observed differences in estimated and assembled genome sizes. Assessment of genome completeness using CEGMA (Table S10, Supporting Information) confirmed that 85.48% and 82.66% of core genes were either completely or partially present in *A. fenestrafer* and *Discosoma* sp*.* respectively, which is close to the completeness of the available genomes of the coral *A. digitifera* and the sea anemone *A. pallida*.

Functional annotation of predicted gene models

Data from *ab initio* gene prediction (AUGUSTUS, SNAP, GlimmerHMM), protein-based homology (Spaln), and transcript evidence (Trinity, PASA) were integrated with EvidenceModeler (EVM) to produce consensus gene sets for both organisms (Table S11, Supporting Information). For *A. fenestrafer*, 21,372 high confidence gene models were predicted with a gene density of 5.78 genes per 100 kbp and an average of 6.5 exons per gene; for *Discosoma* sp., 23,199 high confidence gene models were predicted with a gene density of 5.22 and an average of 6 exons per gene (Table 1). Annotation of these gene models against Swiss-Prot, TrEMBL and nr resulted in predicted gene functions assigned to 84.6% and 83.5% genes in *A. fenestrafer* and *Discosoma* sp*.* respectively, which is similar to the annotation rates in the other three hexacorallian genomes (Table 2).

To assess the accuracy of our predicted gene models, we performed a comparative assessment of the gene models from all five hexacorallian species (*A. fenestrafer*, *Discosoma* sp., *A. digitifera*, *N. vectensis* and *A. pallida*) using several metrics: gene length (Fig. S5a, Supporting Information), exon length (Fig. S5b, Supporting Information), exon count (Fig. S5c, Supporting Information) and GC content of coding regions (Fig. S5d, Supporting Information). As expected from the smaller evolutionary distance (24 Mya) (Simpson *et al*. 2011), both corallimorpharians produced distributions that are more similar to each other than to any of the other three non-corallimorpharians, although potential methodological biases contributing to this finding cannot be entirely excluded.

To further evaluate the quality of our gene models we identified their closest homolog in the Swiss-Prot database via BLASTP and assessed the proportion of the gene model that is covered by its putative homolog. Comparison of the coverage results against the previously published Hexacorallia gene models showed very similar gene coverage distributions, with the exception of *N. vectensis*, as its gene models are present in the Swiss-Prot database (Fig. S6, Supporting Information).

Analysis of known protein domains encoded in the genomes using the Pfam annotation identified 16,045 and 16,966 protein domains in *A. fenestrafer* and *Discosoma* sp. (Table 2), and the proportions of annotated gene models (75.2% and 83.5% respectively) are consistent with the other three cnidarians (Table 2). The analysis of domain abundances highlighted WD domain, G-beta repeat and Ankyrin domains to be the most abundant domains (Fig. S7a, Supporting Information), similar to the other hexacorallian genomes. However, enrichment analyses comparing relative domain abundances between corallimorpharians and the other Hexacorallia species showed that WD domain, G-beta repeats were significantly enriched when compared to the coral *A. digitifera* and the anemone *A. pallida* ($p < 0.001$). Further domains showing significant enrichment in Corallimorpharia were hemopexin, zona pellucida-like domain, EAD/DEAH box helicase domain and putative glycoside hydrolase xylanase, among others (Fig. S7b, Supporting Information).

Confirming the phylogeny within Hexacorallia

While corallimorpharians are commonly accepted to be the sister group of scleractinians, it remains unclear whether, from the corallimorpharians' perspective, scleractinians are genetically closer to them than actiniarians. Using the gene models encoded in the five hexacorallian genomes available, we investigated the overlap of genes between both corallimorpharians and the other non-corallimorpharians. We measured the amount of conservation at an ortholog group level (see Fig. 1, Table S12, Supporting Information) as well as at sequence conservation level using BLAST, and for both measures, the overlap between both corallimorpharians and *A. digitifera* was more extensive than to either of the two actiniarians (see Fig. 1, 2). These findings likely reflect the closer evolutionary distance of corals and corallimorpharians since gene gains or losses as well as sequence divergence are expected to increase proportionally with evolutionary distance. Interestingly though, we also identified a set of 4,952 orthologs that were conserved across all five hexacorallian genomes. A phylogenetic tree based on 696 one-to-one orthologs further supports the observation that Corallimorpharia is a close sister group of Scleractinia (Kitahara *et al.* 2014; Lin *et al.* 2016), and more distantly related to the actiniarians (see Fig. 3).

***Rapid expansion of repeat sequences in Discosoma* sp.**

Homology and structure based searches via Dfam and Repbase identified 666,894 and 1,000,975 repeat elements for *A. fenestrafer* and *Discosoma* sp., which constituted 113.5 Mbp and 167.9 Mbp (30.7% and 37.8%, respectively) of sequences in the assembled *A. fenestrafer* and *Discosoma* sp. genomes (Table 1). Despite the recent evolutionary split, both genomes harboured different proportions of repeat element categories (Table S13, Supporting Information). The predominant types of transposable elements in the two corallimorpharian genomes were DNA/PIF-Harbinger, LINE/L2, and LINE/Penelope (Table S14, Supporting Information). Interestingly we observed that the *Discosoma* sp. genome appears to contain significantly more repetitive elements, including DNA transposons (6.97%) and SINEs (2.5%) than *A. fenestrafer* (5.13% and 2.04% respectively) among others (t-test p < 0.001, Table S13, Supporting Information). The prevalence of repeat sequences in *Discosoma* sp. potentially contributes to the genome size differences observed in comparison to *A. fenestrafer*: the additional 54.4 Mbp of additional repeat sequences in the former is close to the 74.9 Mbp size disparity that separates these two corallimorpharian genomes.

This is further corroborated by self-alignments of the *A. fenestrafer* and *Discosoma* sp. genomes with MUMmer that indicate far more partial duplications (with contiguous length of > 3 kbp) in *Discosoma* sp. than *A. fenestrafer*—these regions, when tallied, totalled 5,281 kbp in the former versus 928 kbp in the latter (Fig. S8a, b, Supporting Information). *A. fenestrafer* and *Discosoma* sp. contained one and seven matches respectively that were > 10 kbp, and these regions tend to be annotated as repeat regions. For example, the largest duplication in *Discosoma* sp. (12,967 bp) showed homology to known retrotransposons, and contained the LTR/Gypsy repeat element.

Discussion

Here, we sequenced and assembled draft genomes for the two Corallimorpharia species, *A. fenestrafer* and *Discosoma* sp.. Validation of the assemblies showed similar quality to currently published genomes of other Anthozoans (Baumgarten *et al.* 2015; Shinzato *et al.* 2011), verifying the suitability of our genomes for comparative studies. We find that the overall genome sizes of the two Corallimorpharia are slightly larger than the published genome of the anemone *A. pallida* but lower than the coral *A. digitifera*. Interestingly, we find that differences in genome size between the two closely related Corallimorpharia species are likely caused by the presence of more repeat elements and genome duplications in *Discosoma* sp. than *Amplexidiscus fenestrafer* as indicated by genome content analyses.

Phylogenetic analysis of single-copy orthologs confirmed the closer relationship of Corallimorpharia with the reef-building coral *A. digitifera* and our genome-level analysis corroborates this. However, while analysis of gene family conservation and protein sequence similarity confirmed the closer phylogenetic relationship of Corallimorpharia and the scleractinian coral *A. digitifera*, we observed a less significant sequence overlap than expected. This might in part be due to the use of *Acropora digitifera* as the sole representative of scleractinians—which might, perhaps, be more derived than other scleractinian species (Lin *et al.* 2016) While our results provide additional proof for the expected closer phylogenetic relationship between corals and Corallimorpharia, it does not resolve the actual phylogenetic relationships and, hence, does not allow conclusions regarding ancestry.

Our findings confirm that Corallimorpharia are the closest living, non-calcifying relatives of reef building corals, therefore closing the current evolutionary gap in the genomic resources available for this important subclass. Consequently we expect these genomic resources to be of great value for future comparative genomic studies analysing the evolution of Hexacorallia species and their specific traits such as the symbiotic relationship with dinoflagellates of the genus *Symbiodinium* or the evolution of calcification in reef-building corals.

Acknowledgements

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST). We would like to also thank Craig Michell for library construction, and the KAUST Sequencing Core Facility for sequencing the libraries.

Contributions

M.A. conceived and designed the study. W.X. assembled and analysed the genomes with help from Y.J.L., Y.L. and M.A.. S.T. and D.Z. provided materials. W.X., Y.J.L. and M.A. wrote the manuscript with input from all authors. All authors agree to be held accountable for the content herein and gave approval for publication.

Data availability

Raw sequence libraries produced for the assembly of the genomes, including whole genome and RNA-seq data are accessible under Bioproject IDs: PRJNA354436 (*A. fenestrafer*) and PRJNA354492 (*Discosoma* sp.). The sequenced genomes, transcriptomes and annotations as well as a detailed bash command protocol and relevant Perl and R scripts can be downloaded at http://corallimorpharia.reefgenomics.org (Liew *et al.* 2016).

Supporting Information

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

Fig. S1 Insert size distributions of mate-pair libraries.

Fig. S2 Divergence rate distribution of repeat elements in the (a) *A. fenestrafer* and (b) *Discosoma* sp. genome.

Fig. S3 Summary of GO annotations across five cnidarian genomes.

Fig. S4 Estimate of artificial contig duplications across different genomes using BLASTN. Bitscore ratios of the second best hit vs. the first best hit are plotted indicating similarity between contigs in each genome.

Fig. S5 Comparison of gene structure across the five cnidarian genomes. (a) Gene length density distribution; (b) Exon length distribution; (c) Exon per gene distribution; (d) GC content of coding sequences.

Fig. S6 Coverage of SwissProt gene homologs.

Fig. S7 (a) Ratios of the top 30 most abundant Pfam domains (depicted as domain ratios * 100). (b) Enrichment of Pfam domains in Corallimorpharia in comparison to *A. digitifera*, *A. pallida* and *N. vectensis* (z-scores of significantly enriched domains, $p < 0.001$)

Fig. S8 Large duplication (> 3 kbp) density in (a) *A. fenestrafer* and (b) *Discosoma* sp..

Table S1 Corallimorpharian genome sequencing statistics.

Table S2 Mapping rate of different libraries against *Symbiodinium* genomes.

Table S3 *k*-mer genome size estimation.

Table S4 Filtering statistics of RNA-seq sequences.

Table S5 Mapping ratio and assembly stats of RNA-seq data.

Table S6 tRNAs and rRNA annotations.

Table S7 InterPro annotations for *A. fenestrafer* and *Discosoma* sp..

Table S8 Single copy orthologs with Pfam and BLAST annotations.

Table S9 Mapping statistics of paired-end sequencing libraries.

Table S10 CEGMA evaluation of genome completeness.

Table S11 Gene prediction statistics.

Table S12 Statistics of shared ortholog groups across five genomes. The colour represents gene presence in the corresponding species.

Table S13 Repeat content and classification in Corallimorpharia.

Table S14 Refinement of repeat element categories.

References

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11.

Baumgarten S, Simakov O, Esherick LY*, et al.* (2015) The genome of Aiptasia, a sea anemone model for coral symbiosis. *Proceedings of the National Academy of Sciences* **112**, 11893-11898.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Bradnam KR, Fass JN, Alexandrov A*, et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**, 1.

Butler J, MacCallum I, Kleber M*, et al.* (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**, 810-820.

Chapman JA, Kirkness EF, Simakov O*, et al.* (2010) The dynamic genome of Hydra. *Nature* **464**, 592-596.

Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**, D363-368.

Crusoe MR, Alameldin HF, Awad S*, et al.* (2015) The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res* **4**, 900.

Daly M, Fautin DG, Cappola VA (2003) Systematics of the hexacorallia (Cnidaria: Anthozoa). *Zoological Journal of the Linnean Society* **139**, 419-437.

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165.

De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

Evans VC, Barker G, Heesom KJ*, et al.* (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature methods* **9**, 1207-1211.

Gotoh O (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic acids research* **36**, 2630-2638.

Haas BJ, Delcher AL, Mount SM*, et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666.

Haas BJ, Salzberg SL, Zhu W*, et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7.

Kajitani R, Toshimoto K, Noguchi H*, et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research* **24**, 1384-1395.

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res* **40**, e9.

Kitahara MV, Lin MF, Foret S*, et al.* (2014) The "naked coral" hypothesis revisited--evidence for and against scleractinian monophyly. *PloS one* **9**, e94774.

Korf I (2004) Gene finding in novel genomes. *BMC bioinformatics* **5**, 1.

Kurtz S, Phillippy A, Delcher AL*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome biology* **5**, 1.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li H, Handsaker B, Wysoker A*, et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178-2189.

Liew YJ, Aranda M, Carr A*, et al.* (2014) Identification of MicroRNAs in the Coral Stylophora pistillata. *PloS one* **9**, e91101.

Liew YJ, Aranda M, Voolstra CR (2016) Reefgenomics.Org **-** a repository for marine genomics data. *Database (Oxford)* **2016**.

Lin MF, Chou WH, Kitahara MV*, et al.* (2016) Corallimorpharians are not "naked corals": insights into relationships between Scleractinia and Corallimorpharia from phylogenomic analyses. *PeerJ* **4**, e2463.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955-964.

Luo R, Liu B, Xie Y*, et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 1-6.

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770.

Medina M, Collins AG, Takaoka TL, Kuehl JV, Boore JL (2006) Naked corals: skeleton loss in Scleractinia. *Proc Natl Acad Sci U S A* **103**, 9096-9100.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**, W182-W185.

Moya A, Tambutté S, Béranger G*, et al.* (2008) Cloning and use of a coral 36B4 gene to study the differential expression of coral genes between light and dark conditions. *Marine biotechnology* **10**, 653-663.

Murray JM, Watson GJ (2014) A critical assessment of marine aquarist biodiversity data and commercial aquaculture: identifying gaps in culture initiatives to inform local fisheries managers. *PloS one* **9**, e105982.

Nawrocki EP, Burge SW, Bateman A*, et al.* (2014) Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, gku1063.

Park E, Hwang DS, Lee JS*, et al.* (2012) Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record. *Mol Phylogenet Evol* **62**, 329-345.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067.

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61-D65.

Putnam NH, Srivastava M, Hellsten U*, et al.* (2007) Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* **317**, 86-94.

Ronquist F, Teslenko M, van der Mark P*, et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* **61**, 539-542.

Schwentner M, Bosch TC (2015) Revisiting the age, evolutionary history and species level diversity of the genus Hydra (Cnidaria: Hydrozoa). *Mol Phylogenet Evol* **91**, 41-55.

Shinzato C, Shoguchi E, Kawashima T*, et al.* (2011) Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* **476**, 320-323.

Shoguchi E, Shinzato C, Kawashima T*, et al.* (2013) Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Current Biology* **23**, 1399-1408.

Simpson C, Kiessling W, Mewis H, Baron-Szabo RC, Muller J (2011) Evolutionary diversification of reef corals: a comparison of the molecular and fossil records. *Evolution* **65**, 3274-3284.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313.

Stanke M, Keller O, Gunduz I*, et al.* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-439.

Stanley GD, Jr., Fautin DG (2001) Paleontology and evolution. The origins of modern corals. *Science* **291**, 1913-1914.

Stolarski J, Kitahara MV, Miller DJ*, et al.* (2011) The ancient evolutionary origins of Scleractinia revealed by azooxanthellate corals. *BMC Evol Biol* **11**, 316.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* **56**, 564-577.

Vandepitte L, Vanhoorne B, Decock W*, et al.* (2013) World Register of Marine Species.

Warnes MGR (2016) Package 'gplots'.

Wheeler TJ, Clements J, Eddy SR*, et al.* (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res* **41**, D70-82.

Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.

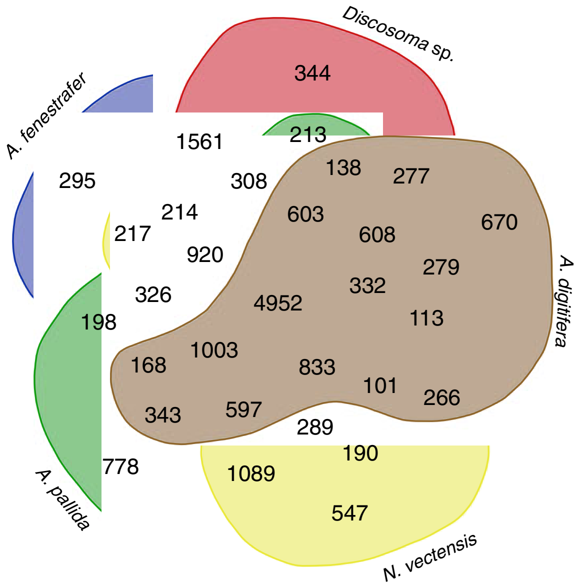

Fig. 1 Venn diagram representing the shared ortholog groups across five hexacorallian genomes. There is extensive conservation of ortholog groups (4,952) across all five genomes.

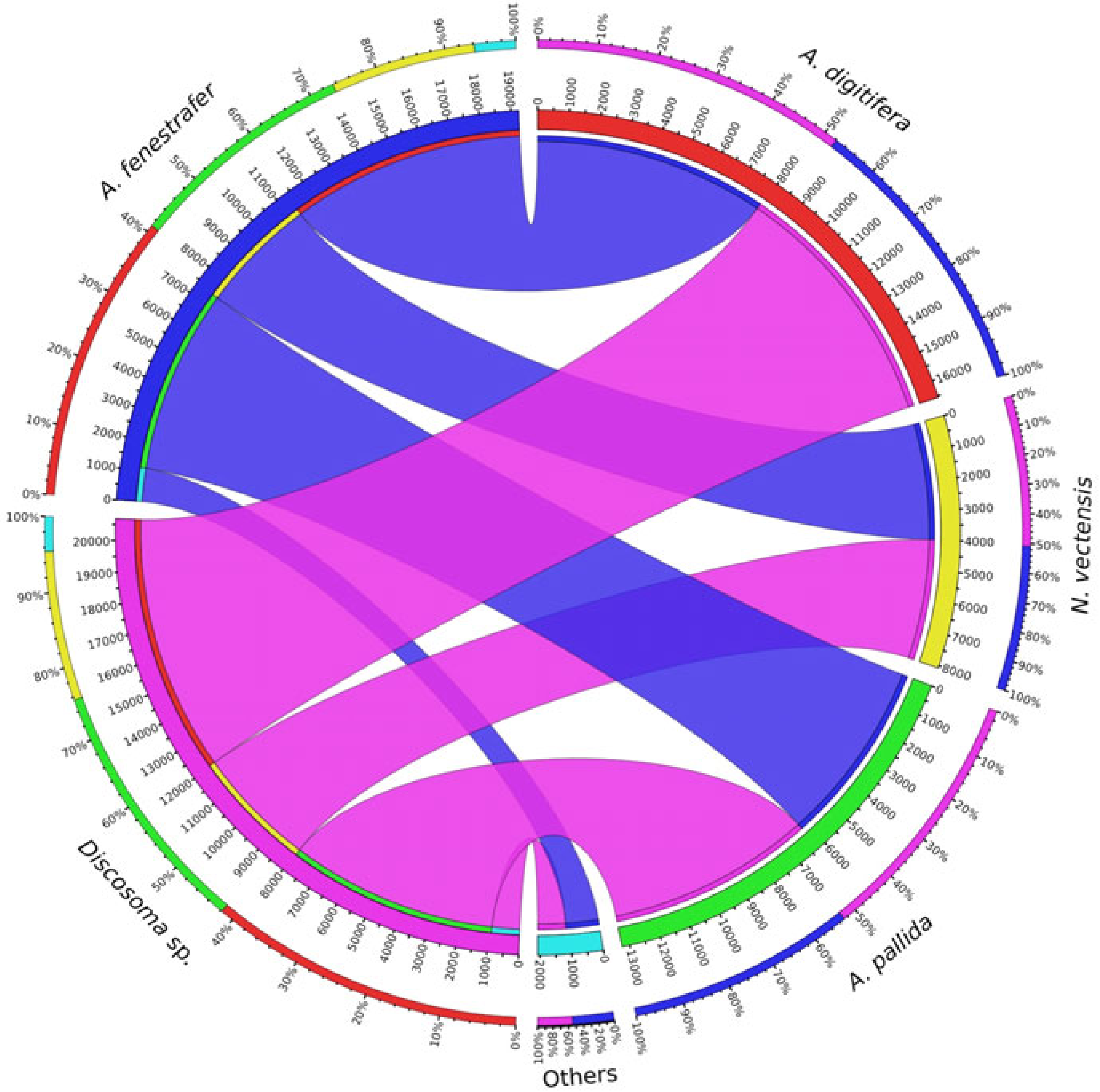

Fig. 2 Chord diagram tallying which non-corallimorpharian species has the best-matching homolog to every corallimorpharian gene based on BLAST. The different colours correspond to the different species as per the inner ring. Numbers on the inner ring depict the number of genes, and the numbers on the outer ring indicate their percentage in the respective species.

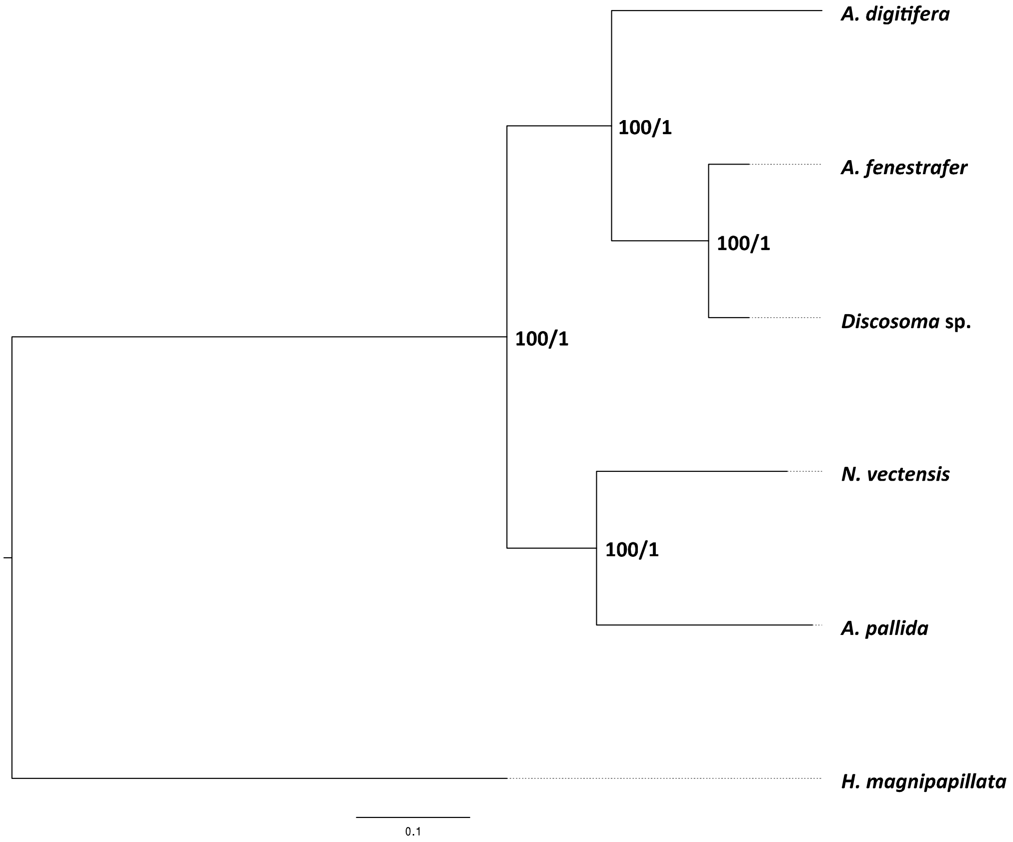

Fig. 3 Phylogenetic tree based on 696 single copy orthologs of five hexacorallian genomes and the non-hexacorallian cnidarian *Hydra magnipapillata* as outgroup. The numbers on the nodes correspond to the bootstrap values estimated by RAxML and posterior probabilities estimated by MrBayes respectively.

Table 1. Genome and annotation statistics for *A. fenestrafer* and *Discosoma* sp.

		A. fenestrafer	*Discosoma* sp.
Genome	Estimated genome size (Mbp)	350	428
	Number of scaffolds	5,631	6,449
	Scaffold length (bp)	370,076,467	444,930,386
	Scaffold N50 (bp)	510,298	769,877
	Number of contigs	30,993	39,613
	Contig length (bp)	305,666,947	364,304,111
	Contig N50 (bp)	20,058	18,746
	GC content (%)	39.30	39.97
Gene	Number of gene model	21,372	23,199
	Average gene length (bp)	7,194	6,907
	Gene density (genes per 100kbp)	5.78	5.22
	Number of exons	139,123	138,376
	Average exon length (bp)	218	226
	Number of introns	117,751	115,177
	Average intron length (bp)	1,047	1,119
	Overall GC content (%)	36.70	37.09
Repeat	Total repeat length (Mbp)	113.5	167.9
	Repeat proportion (%)	30.7	37.8
Nocoding	Total tRNAs	2,936	4,775
	Total rRNAs	752	1,066

Table 2. Functional annotation of gene models from five cnidarian species.

		A. fenestrafer	*Discosoma* sp.	*A. digitifera*	*A. pallida*	*N. vectensis*
Total gene number		21,372	23,199	23,668	29,269	27,273
	SwissProt	12,867	13,408	16,025	18,876	20,125
	TrEMBL	3,684	4,233	2,519	4,117	3,496
	nr	1,534	1,677	1,428	2,052	3,652
	No hit	3,279	3,826	3,551	4,224	0
Pfam		16,045	16,966	15,102	19,499	18,171
PANTHER		13,276	13,888	15,813	19,582	20,001
GO		16,552	17,642	18,546	22,997	23,621
KO		7,356	7,592	9,310	10,619	11,402