OPEN ACCESS Freely available online

PLoS one

Gene Discovery in the Threatened Elkhorn Coral: 454 Sequencing of the *Acropora palmata* Transcriptome

Nicholas R. Polato, J. Cristobal Vera, Iliana B. Baums*

Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America

Abstract

Background: Cnidarians, including corals and anemones, offer unique insights into metazoan evolution because they harbor genetic similarities with vertebrates beyond that found in model invertebrates and retain genes known only from non-metazoans. Cataloging genes expressed in *Acropora palmata*, a foundation-species of reefs in the Caribbean and western Atlantic, will advance our understanding of the genetic basis of ecologically important traits in corals and comes at a time when sequencing efforts in other cnidarians allow for multi-species comparisons.

Results: A cDNA library from a sample enriched for symbiont free larval tissue was sequenced on the 454 GS-FLX platform. Over 960,000 reads were obtained and assembled into 42,630 contigs. Annotation data was acquired for 57% of the assembled sequences. Analysis of the assembled sequences indicated that 83–100% of all *A. palmata* transcripts were tagged, and provided a rough estimate of the total number genes expressed in our samples (~18,000–20,000). The coral annotation data contained many of the same molecular components as in the Bilateria, particularly in pathways associated with oxidative stress and DNA damage repair, and provided evidence that homologs of p53, a key player in DNA repair pathways, has experienced selection along the branch separating Cnidaria and Bilateria. Transcriptome wide screens of paralog groups and transition/transversion ratios highlighted genes including: green fluorescent proteins, carbonic anhydrase, and oxidative stress proteins; and functional groups involved in protein and nucleic acid metabolism, and the formation of structural molecules. These results provide a starting point for study of adaptive evolution in corals.

Conclusions: Currently available transcriptome data now make comparative studies of the mechanisms underlying coral's evolutionary success possible. Here we identified candidate genes that enable corals to maintain genomic integrity despite considerable exposure to genotoxic stress over long life spans, and showed conservation of important physiological pathways between corals and bilaterians.

Citation: Polato NR, Vera JC, Baums IB (2011) Gene Discovery in the Threatened Elkhorn Coral: 454 Sequencing of the *Acropora palmata* Transcriptome. PLoS ONE 6(12): e28634. doi:10.1371/journal.pone.0028634

Editor: Timothy Ravasi, King Abdullah University of Science and Technology, Saudi Arabia

Received June 22, 2011; **Accepted** November 12, 2011; **Published** December 28, 2011

Funding: This work was made possible with funding from a National Science Foundation (NSF) Graduate Research Fellowship Program grant to NP, National Oceanic and Atmospheric Administration Coral Ree462 to IB, NSF grant OCE – 0825979 to IB and JV was supported by NSF grant IOS-0950416. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: baums@psu.edu

Introduction

Cnidarians are valuable for understanding the evolution of metazoan genomes because they comprise a sister group to the Bilateria. From a developmental perspective, comparisons among these groups have yielded much information regarding the evolution of body plans in complex multicellular animals [1]. The two cnidarian genomes sequenced to date (*Hydra magnipapillata and Nematostella vectensis*) have revealed that basal metazoans retain genes known only from non-metazoans [2], and possess genetic pathways found in vertebrates but lacking in worm and fly models [3].

Corals in particular offer insights into important biological processes (including calcification and symbiosis) that are responsible for the formation of reefs worldwide [4]. These processes contribute to the fitness of coral genets. Successful genets can live to be hundreds of years old [5,6], and because generation times for *A. palmata* are short by comparison (on the scale of 2 to 5 years [7]), a single successful genet can contribute an enormous number of offspring to future generations over its lifetime. Thus, chronic exposure to high levels of UV irradiation and radical oxygen species (ROS) produced as a byproduct of symbiont photosynthesis, is likely to have driven the evolution of powerful genoprotective mechanisms and DNA repair pathways in corals [8].

Progress in sequencing technology, chemistry, and bioinformatics now enable the sequencing and *de novo* assembly of whole transcriptomes from non-model organisms [9,10,11]. Such genomic resources are needed for species of conservation concern because they allow researchers to target functional variation in wild populations [12,13,14], track adaptive responses to environmental stress through space and time [15], and inform breeding programs and reintroduction efforts [16].

Despite a population reduction of >80% throughout its range in recent decades [17], the Elkhorn Coral (*Acropora palmata*) remains an ecologically important community member [18,19], providing the three dimensional structure of many Caribbean reefs and a substantial contribution to reef primary productivity [18,19]. Further, *A. palmata* is one of only a few Caribbean species for

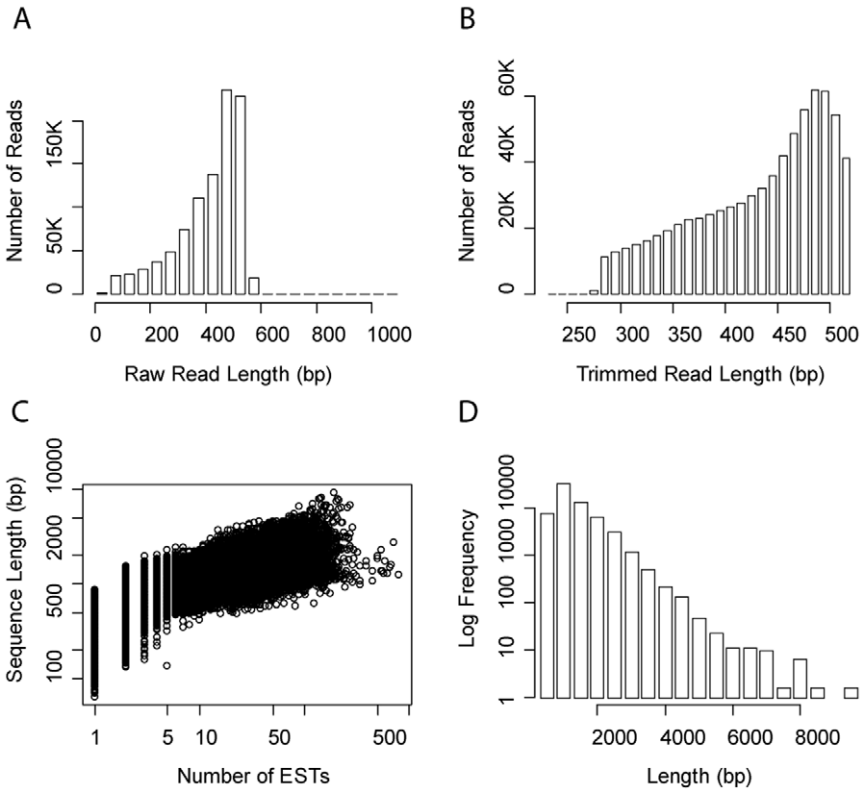

Figure 1. Summary of sequencing read length distributions from the *A. palmata* transcriptome pre and post assembly. A) Unassembled raw sequencing read lengths (in basepairs, bp) prior to trimming. B) Size distribution of unassembled sequences (in bp) following quality trimming. C) Plot of the relationship between the length of assembled contigs (in bp) and the depth of coverage in terms of the number of raw ESTs they include (note log-log axes). D) Frequency histogram of contig lengths (in bp).
doi:10.1371/journal.pone.0028634.g001

which population genetic data is available [20,21,22]. Microsatellite markers have revealed patterns of population differentiation and the contribution of asexual reproduction throughout *A. palmata's* range [23,24] and as such, a baseline for genetic and genotypic diversity within this species exists.

Transcriptome and EST data are now available from several coral species including *Acropora millepora*, *Acropora hyacinthus*, *Montastraea faveolata*, *Pocillopora damicornis*, *and Porites astreoides* [9,25,26,27,28, Matz and Traylor-Knowles pers. comm.] enabling comparative genomics within the scleractinia [28,29]. Likewise the genomes of two cnidarians, *Nematostella vectensis* and *Hydra magnipapillata* were recently completed [30,31] permitting evolutionary analysis within the phylum.

To develop genome scale tools for *A. palmata* we present here its transcriptome sequenced with 454 GS-FLX technology, and survey the assembled and annotated sequences for genetic markers, and functional enrichment of important biological processes (with a particular focus on stress response and DNA repair pathways). The genetic resources presented here are a major advance for studies of this species and will promote the study of adaptive trait variation in wild populations of *A. palmata* as well as comparative genomics among basal metazoans.

Results

Biological material

RNA was acquired from genetically diverse larvae collected over a broad geographic range previously identified as comprising two divergent populations [23]. Larvae were exposed to three temperature and two CO_2 treatments, and sampled over the course of development from fertilization to the planula stage. A small amount of adult tissue was also included to represent adult transcripts. We thus expected to find developmental and stress related genes derived from the larval temperature and CO_2 treatments, as well as genes related to symbiosis and calcification from the adult tissue.

Sequencing and Assembly

One plate of a normalized cDNA library generated from a pooled *A. palmata* sample following the library preparation methods of *Meyer et al.* [9] was sequenced on the 454 GS-FLX platform using Titanium chemistry. This run yielded 964,519 raw reads with an average length of 398 bp ($\sigma = 118$), totaling of 384 Mb (Fig. 1A; Table 1). After trimming for size, quality, and primer sequence 741,271 reads remained averaging 432 bp

Table 1. Summary of sequencing and assembly of the *Acropora palmata* transcriptome.

	N sequences	Total length [Mb]	Avg. length (sd) [bp]
raw reads	964,519	384	398 (118)
trimmed reads	741,271	320	432 (64)
contigs	42,630	44	1030 (623)
singletons	45,390	20	439 (94)
total	88,020	64	

doi:10.1371/journal.pone.0028634.t001

($\sigma = 64$) in length and totaling 320 Mb. The majority of trimmed reads (98%) were over 400 bp in length (Fig. 1B) contributing to the high quality of the assembly.

Trimmed reads were assembled with a set of publicly available *A. palmata* ESTs (n = 36,236; SymBioSys database: sequoia.ucmerced.edu/SymBioSys/). Assembly yielded 42,630 contigs averaging 1,030 bp long ($\sigma = 623$). Contig sizes ranged from 132 to 9,066 bp. The mean depth of coverage was 5.6 sequences (Fig. 1C). The size distribution of contig lengths showed an abundance of large contigs resulting from the long read lengths (Fig. 1D). Over 88% (37,819) of the contigs were greater than 500 bp long, and over 38% (16,274) were greater than 1,000 bp. After assembly, 45,390 singletons remained that could not be incorporated into any contig. Due to their long average length (439 bp; $\sigma = 94$), and the fact that a substantial proportion of singletons are likely to represent low abundance transcripts [9], these sequences were included with the contig sequences in BLAST searches [32] for annotation information.

Transcriptome Completeness

Estimation of transcriptome completeness based on two different methods suggest that our single sequencing run tagged between 83–100% of the genes expressed in *A. palmata*. A BLAST query of the assembled *A. palmata* sequences against a set of 119 orthologs conserved across metazoans and found to be single copy in cnidarians (see methods), showed high quality hits (e-value $\leq 2\times10^{-38}$, bitscore ≥ 130) to all 119 genes. Comparison of the distributions of coverage depth between all *A. palmata* transcripts and those with hits to the 119 single copy orthologs showed that the mean depth of coverage for contigs with hits to these 119 genes (9.7) is greater than that for contigs in general (5.6). This analysis indicated that depth of coverage was high even for single copy genes and that the set of 119 orthologs can serve as a representative sample of the *A. palmata* transcriptome.

Secondly, by comparing our assembled data to the published *N. vectensis* transcriptome we found hits to 83% of all *N. vectensis* transcripts. The lengths of the *A. palmata* sequences were generally greater than those of *N. vectensis*, suggesting that full length transcripts were obtained for the majority of the *A. palmata* genes (Information S1). The distribution of log length ratios was shifted to the right of zero due to the fact that *A. palmata* sequences were longer on average than the corresponding *N. vectensis* sequence. The modal value of the length ratio distribution (106%) reflected the longer average lengths of the *A. palmata* transcripts, and taken together with the high proportion of hits to the *N. vectensis* transcriptome data indicated very high coverage of the *A. palmata* transcriptome.

Annotation and Transcriptome Size

All contigs and singletons (n = 88,020 sequences) were used to search against the UniProt protein database. Searches of both Swiss-Prot and TrEMBL resulted in hits for 50,118 (57%) of the queried sequences, of which 32,114 (36%) represented unique subject names. 31,888 of these hits corresponded to Gene Ontology (GO) [33] annotations representing 109,664 terms (4,617 unique). Of the GO terms identified, 46% were molecular functions, 32% biological processes, and 22% cellular components.

To estimate annotation efficiency, the description terms associated with each of the 119 single copy genes were compared to those in the *A. palmata* dataset (derived from UniProt). A clear match was found for 88% of the 119 annotation descriptions. An additional 9% of matching *A. palmata* transcripts were described with uninformative species specific codes from UniProt, but functional matches were easily found by looking up the codes in the NCBI UniGene database. Only three of the 119 genes did not match with the *A. palmata* annotation data, and in all three cases the *A. palmata* annotation was derived from blast hits to the Florida lancelet or the green puffer, species with limited annotation data. Overall this suggests high (97%) annotation efficiency.

Assuming that gene duplication is not excessive (a reasonable expectation based on the results of the paralog analysis below), the number of assembled sequences matching to known single copy genes can give an estimate of the level of residual gene fragmentation (either from sequencing gaps or splice variants), and/or assembly difficulty. BLAST results from the 119 orthologs were used to examine the average number of hits to each putatively singe copy gene. Results of this test showed that 63% of the 119 single copy genes had multiple sequence matches (range: 1–19; mean = 2.2; $\sigma = 1.95$; Information S1) in the *A. palmata* data, suggesting a moderate amount of fragmentation in our data, with some outliers that may be related to gene size and/or species specific duplications.

The degree of fragmentation in our assembly was also used to generate a rough estimate of the total number of transcripts expressed in the *A. palmata* sample. The mean value of 2.2 contigs per transcript was used as an estimate of the redundancy in our dataset. This yielded an estimate of ~19,377 genes (42,630 contigs ÷2.2) expressed in our *A. palmata* sample. Finally, based on the BLAST results with *N. vectensis* we found that 43,036 of the *A. palmtata* sequences matched to a non-redundant set of 17,988 proteins from *N. vectensis*. Thus we predict that the number of transcripts expressed in *A. palmata* is in the range of ~18,000–20,000 genes.

Taxonomic Annotation

The majority of taxonomic associations of the annotated transcripts matched to metazoan invertebrates and vertebrates (44% and 48% respectively). A small proportion of sequences matched to bacteria (3%), protists (2%), plants and algae (2%). All other groups including fungi, archaea and viruses accounted for 1% or less of matches. Stony coral (Scleractinian) sequences accounted for only 1% of the annotated sequences primarily due to a paucity of annotated coral sequences present in current protein databases.

Because BLAST annotation associates a given sequence with its closest match in the database, results are limited to previously identified genes. Thus, the vast majority of annotated transcripts in this dataset showed matches to the Starlet Sea Anemone, *Nematostella vectensis* (32,465 using best hit criterion), the most closely related organism with a sequenced genome [31]. Similarly, an abundance of cepheolochordate BLAST hits reflected the sequencing efforts on the Florida Lancet, *Branchiostoma floridae* [34]. Because annotation information is limited for both of these genomes we considered the next most informative BLAST hits in our annotation database when the best sequence match lacked any functionally relevant data. Viewed this way the number of *A. palmata* sequences with best matches to *N. vectensis* was reduced from 36% to 12%. Thus, despite slightly higher sequence similarity of hits to *N. vectensis*, the search for useful functional annotations was greatly improved by exploring more distantly related hits.

Functional Annotation

To guide functional interpretation of *A. palmata* transcripts, we placed *A. palmata* homologs in the context of known biological pathways using IPA software (Ingenuity Systems; www.ingenuity.com). Information in the IPA database reflects the level of interest certain pathways have received. Thus, comparison of *A. palmata* transcripts to the IPA database identified well annotated metazoan

pathways for which a high proportion of pathway components were present in *A. palmata*. These pathways were then used as guides for further analysis. Note that this approach did not make inferences about expression levels of transcripts. Among the most highly represented pathways were many that are expected to be important to heat stressed embryos and who are thus of interest to the current study including protein ubiquitination, cell cycle control, NRF2 mediated oxidative stress response, p53 signaling and DNA double stranded break repair (Fig. 2).

DNA damage and oxidative stress. Genes involved in the response to DNA damage and oxidative stress (Figs 3 & 4) included homologs of the transcription factors HIF1α and NRF2 responsible for promoting expression of numerous proteins with oxidoreductase activity [35,36]. Additionally, a homolog of KEAP1, the primary regulator of NRF2, [37] was identified along with other transcripts involved in the detoxification, repair and removal of damaged proteins. A multitude of heat stress response genes were found, including 13 HSP variants, numerous HSP binding proteins, and associated transcription factors.

Selection in p53 family genes. Next, we placed genes involved in the response to DNA damage and oxidative (and irradiant) stress in a phylogenetic context to test for signatures of natural selection. 498 bp of the conserved DNA binding domain from homologs of two *N. vectensis* p53 family members (pVS53a, and p63), were chosen to test for positive selection with the program PAML [38]. A third homolog (pEC53a) was also identified but the sequence did not include the conserved DNA binding domain and could not be included in the alignment. Results of a likelihood ratio test ($p = 0.02$; $2\Delta\ln L = 11.4$, d.f. = 4,) comparing a "nearly neutral" model with two site classes ($dN/dS = 0$ and $dN/dS = 1$), and a "positive selection" model that included a third class with a dN/dS ratio >1, supported the hypothesis that positive selection has occurred on the branch of the tree separating cnidarian and bilaterian p53 family members (Fig. 5; Information S2).

Calcification. Calcification is crucial for coral growth and survival, and numerous genes involved in coral calcification were found in the *A. palmata* transcriptome. These included 68 transmembrane ion transporters, including 10 sodium driven bicarbonate exchangers which may play an important role in supplying bicarbonate ions to the reaction site of calcification. Another enzyme, carbonic anhydrase found in the coral calicodermis (a cell layer at the interface of the polyp and skeleton that secretes organic molecules to promote biomineralization) [39], catalyzes the hydration of carbon dioxide leading to hydrogen and bicarbonate ions, which in turn react with calcium ions to form the calcium carbonate skeleton. Twenty-four hits to carbonic anhydrase enzymes were found in our dataset including a homolog of α-CA recently cloned from *Stylophora pistillata* [40].

Paralog Group Analysis

Identification of orthologous and paralogous transcripts in a species pair provides information regarding gene duplication prior to and following speciation, and may point out gene families of particular importance to certain species. Between *N. vectensis* and *A. palmata*, 7,754 ortholog pairs were detected, with 9,604 in-paralogs

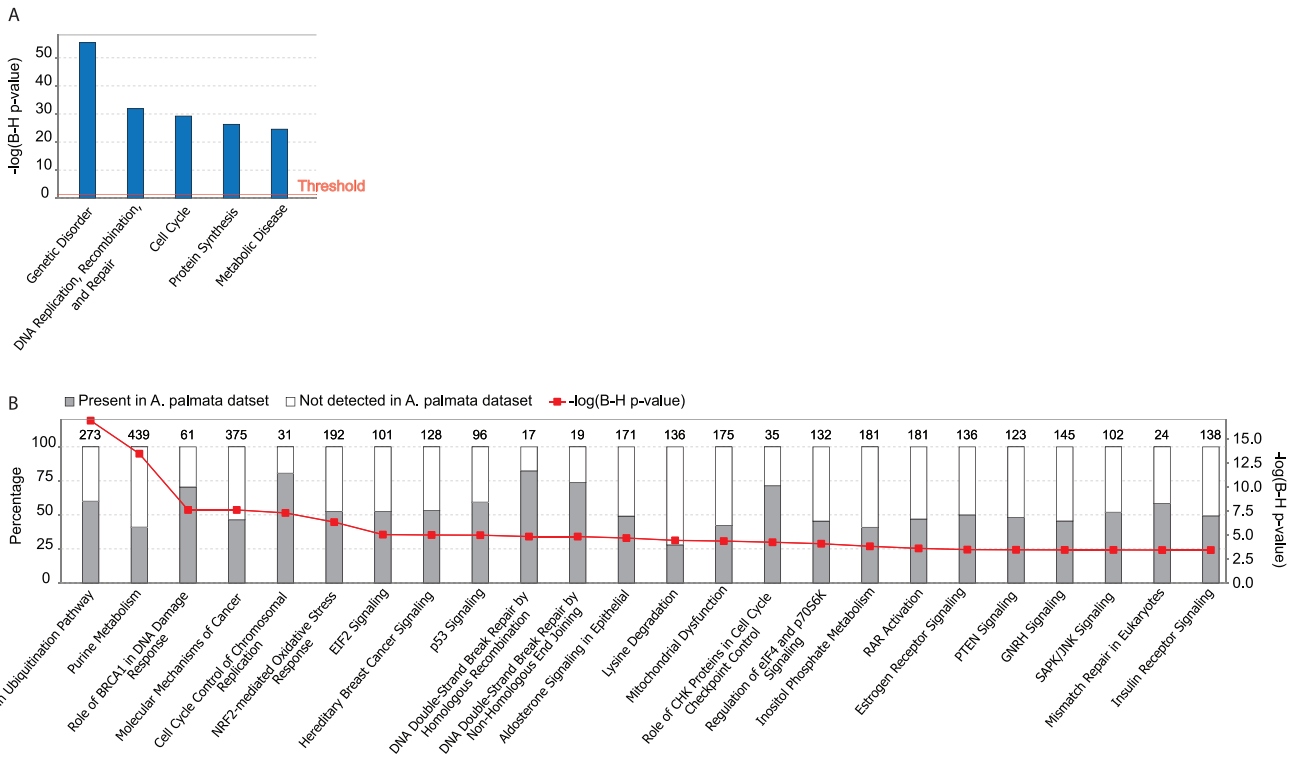

Figure 2. Enrichment analysis highlighted well annotated pathways that are of interest in heat stressed embryos. A) The five most highly enriched functional categories in the *A. palmata* dataset (the orange line represents the threshold for significance at $p<0.05$ with FDR adjustment for multiple testing. B) The top canonical pathways from the IPA library of pathways that were most significant to the dataset. The left Y axis shows the percentage of proteins in the pathway that were identified in the *A. palmata* data (grey bars, note that this is independent of expression levels). The right Y axis shows the corrected –log (p-values) for Fisher's exact test of the probability that the association between the pathway and the data is explained by chance.
doi:10.1371/journal.pone.0028634.g002

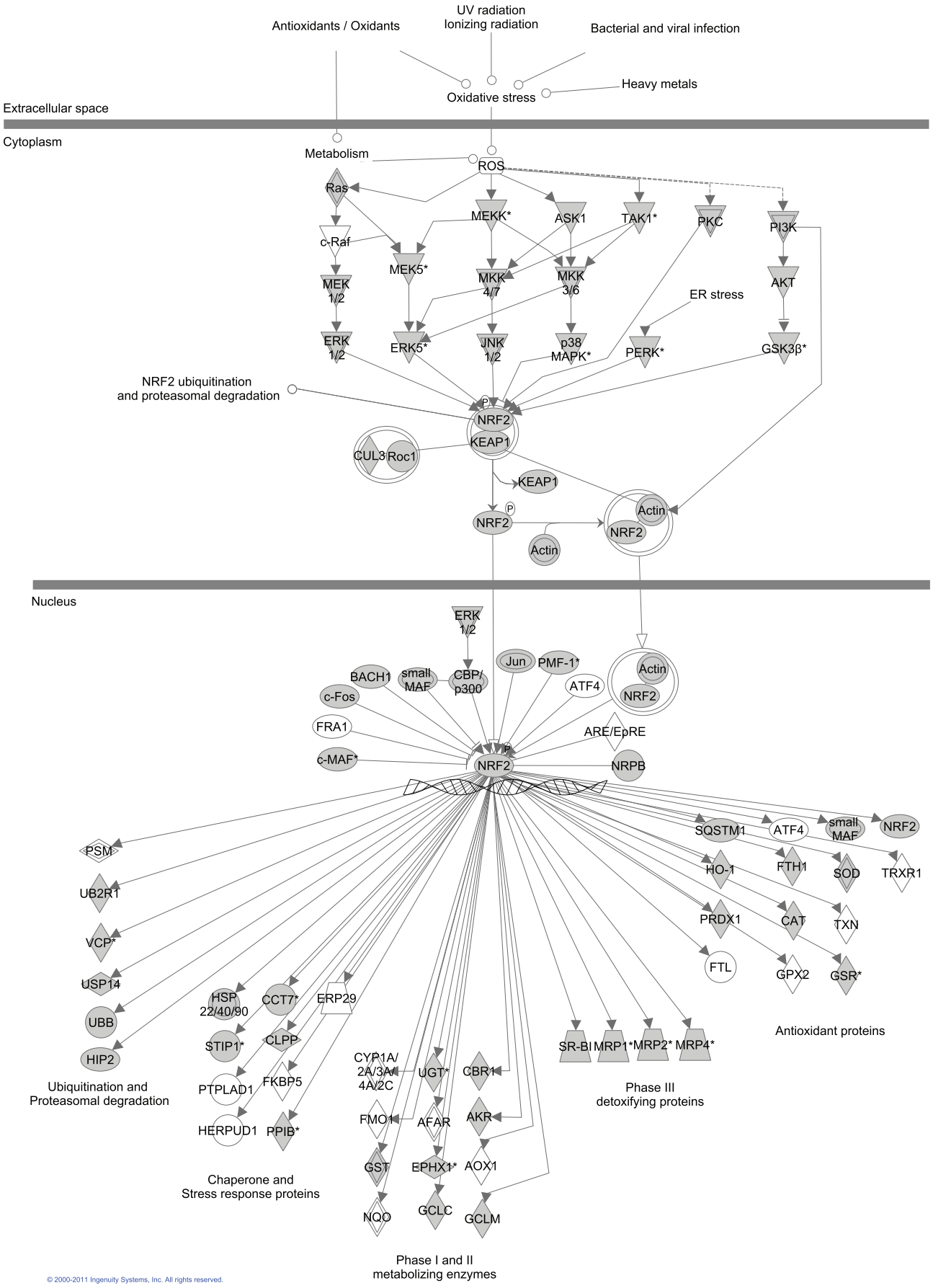
Antioxidants / Oxidants
UV radiation
Ionizing radiation
Bacterial and viral infection
Heavy metals
Oxidative stress
Extracellular space
Cytoplasm
Metabolism
ROS
Ras
c-Raf
MEK 1/2
ERK 1/2
MEK5*
ERK5*
MEKK*
MKK 4/7
JNK 1/2
ASK1
MKK 3/6
p38 MAPK*
TAK1*
PERK*
PKC
ER stress
PI3K
AKT
GSK3β*
NRF2 ubiquitination and proteasomal degradation
P
NRF2
KEAP1
CUL3
Roc1
KEAP1
P
NRF2
Actin
NRF2
Actin
Nucleus
ERK 1/2
BACH1
small MAF
CBP/ p300
Jun
PMF-1*
ATF4
Actin
NRF2
c-Fos
FRA1
c-MAF*
ARE/EpRE
P
NRF2
NRPB
PSM
UB2R1
VCP*
USP14
UBB
HIP2
Ubiquitination and Proteasomal degradation
HSP 22/40/90
CCT7*
ERP29
STIP1*
CLPP
PTPLAD1
FKBP5
HERPUD1
PPIB*
Chaperone and Stress response proteins
CYP1A/ 2A/3A/ 4A/2C
UGT*
CBR1
FMO1
AFAR
AKR
GST
EPHX1*
AOX1
NQO
GCLC
GCLM
Phase I and II metabolizing enzymes
SR-BI
MRP1*
MRP2*
MRP4*
Phase III detoxifying proteins
SQSTM1
ATF4
small MAF
NRF2
HO-1
FTH1
SOD
TRXR1
PRDX1
CAT
TXN
FTL
GPX2
GSR*
Antioxidant proteins
© 2000-2011 Ingenuity Systems, Inc. All rights reserved.

Figure 3. Homologs involved in canonical oxidative stress response pathways identified in the *A. palmata* transcriptome. Filled symbols indicate proteins with homologs that were detected in the *A. palmata* dataset. Unfilled symbols indicate proteins that were not detected in our data, but are present in model vertebrate pathways.
doi:10.1371/journal.pone.0028634.g003

in *N. vectensis* and 9,365 in-paralogs in *A. palmata*. The size distribution of paralog groups showed that the number of groups detected declines with increasing group size, and groups of 5 or more were rare (Fig. 6). Functional annotation of genes with multiple paralogs (group size >5) in *A. palmata* revealed a number of transcripts of known importance to corals, including green fluorescent proteins, carbonic anhydrase, and the oxidative stress response gene ferritin (Table 2). Interestingly among the largest groups identified were homologs of the immunoglobulin superfamily proteins in the IgLON family (Immunoglobulin superfamily containing LAMP, OBCAM, and neurotrimin), and several tumor necrosis factor receptor-associated factors (TRAFs).

Microsatellite and SNP discovery

A total of 333 microsatellites (defined as repeats with motifs between 2 and 6 bp) were found, with a large majority consisting of trinucleotide repeats (n = 170; Fig. 7). Within the trinucleotides, AAN type repeats were most abundant, with the highest number being AAC repeats, followed by AAG, and AAT (44, 35, and 20 respectively).

A total of 72,605 candidate SNPs were identified from 13,803 contigs spanning 19.8 Mb of sequence. The overall SNP frequency was 1 per 272 bp, with 71% transitions (Ts) and 29% transversions (Tv). Frequencies of different transitions types were similar, as were frequencies of different transversion types (Fig. 8). The transcriptome wide Ts/Tv ratio was 2.4, and while most contigs (55%; n = 3,823) had ratios of 2 or greater, many contigs had ratios that were substantially higher or lower (Fig. 9). Contigs with especially high and low ratios were analyzed to identify possible genes under selection and associated functional groups (Table 3). These contigs included genes involved in TGF-β, Hedgehog, and WNT signaling pathways; DNA and protein repair; lipid, protein and energy metabolism; HSP binding; oxidation-reduction; and regulation of transcription (Information S3).

To date, 18 SNPs from 6 contigs have been targeted for amplification using PCR, and the products were sequenced on an ABI3700 fragment analyzer. Of these 18 SNPs, 8 were observed in the sample population (success rate 44%), which consisted of 11 individuals from Florida, 8 from Curacao, and 7 from Puerto Rico. The SNPs detected were in genes for NFkB, melanopsin, Cyano Fluorescent Protein, NRF2, galaxin, and LITAF (Information S4). Ongoing work will develop a panel of SNPs for surveys of functional variation across the species' range.

Discussion

Coral species are the foundation of reefs because of their association with intra-cellular photosymbionts and ability to calcify. Comparative functional genomic analyses are now possible to elucidate the mechanisms underlying coral's evolutionary success. The *Acropora palmata* transcriptome presented here joins transcriptomes and EST data available for several other scleractinian corals, including that of the Pacific congener *A. millepora* (Matz and Traylor-Knowles pers. comm.) [9,25,26,27,28]. The distribution of transcript sizes observed in *A. palmata* (mean = 1,030 bp; median = 830 bp; σ = 623 bp) was comparable to that of *N. vectensis* (mean = 1,091 bp; median = 806 bp; σ = 1073 bp), and the estimated number of genes expressed in this species (~18,000–20,000) was within the range estimated for *A. millepora* (~11,000) and *N. vectensis* (27,273).

The high quality and comprehensiveness of the assembled data enabled functional analysis of pathways shared among metazoans as well as those of particular importance to the Cnidaria. *A. palmata* possesses many of the molecular components present in the more

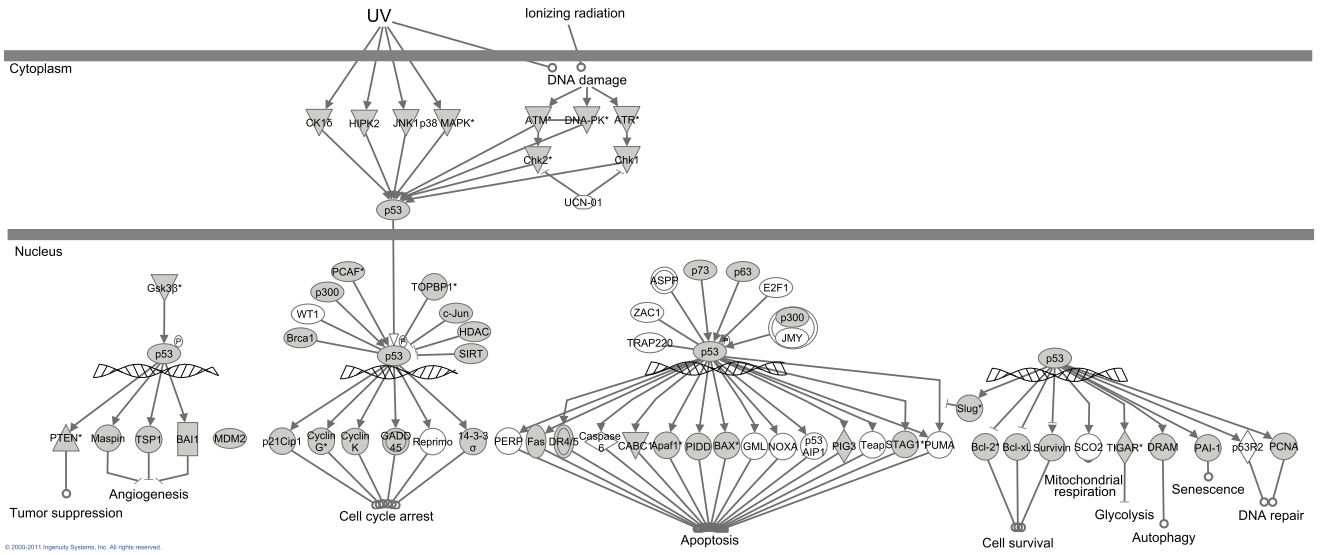

Figure 4. Genes involved in the various p53 mediated responses to DNA damage. Specific pathway details are likely to differ somewhat in the Cnidaria, but the presence of these components in *A. palmata* suggested the capacity for similar responses in corals. Filled symbols indicate proteins with homologs that were detected in the *A. palmata* dataset. Unfilled symbols indicate proteins that were not detected in our data, but are present in model vertebrate pathways.
doi:10.1371/journal.pone.0028634.g004

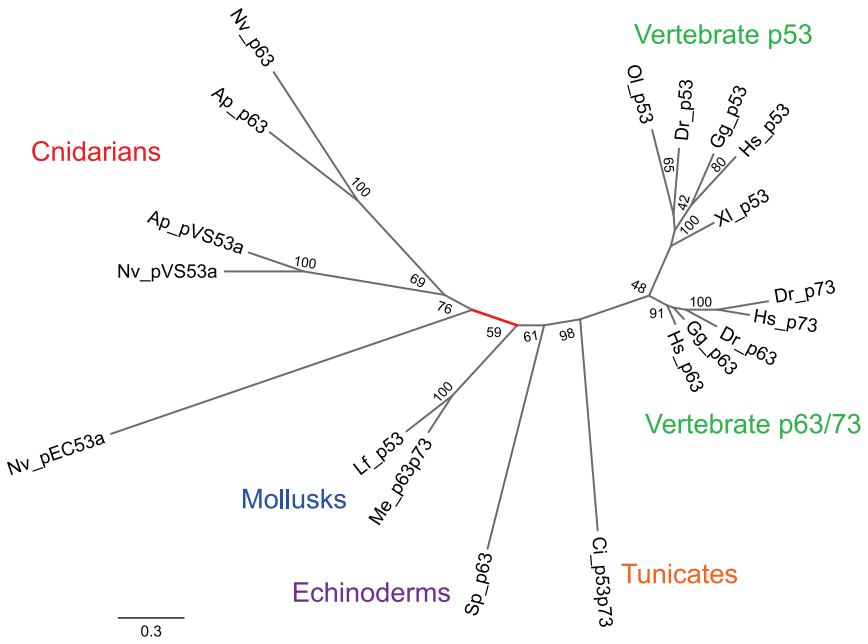

Figure 5. p53 gene family tree. Tests were performed to detect evidence of natural selection along the branch leading to the Cnidaria with the program PAML (highlighted in red). The sequences in the tree include a subset of those from Rutkowski (2010). The tree was generated in GARLI using the TIM2+G model.
doi:10.1371/journal.pone.0028634.g005

complex Bilateria (particularly in pathways associated with oxidative stress and DNA damage repair) and homologs of p53, a key player in several DNA repair pathways, showed evidence of natural selection along the lineage separating Cnidaria and Bilateria.

Taxonomic Annotation

Xenobiotic sequences from intracellular parasites are commonly found in transcriptome data of wild species [10]. Many of the non-coral sequences identified here may represent symbiotic or pathogenic organisms associated with *A. palmata*. Indeed, approximately half of the sequences with BLAST hits in the 'protists' category (555) were associated with the superphylum which includes symbiotic zooxanthellae.

Hits to sequences from scleractinians were rare because of the small number of coral genes in current databases (0.01% of total). As such, the genes identified are likely to be highly conserved among the scleractinians, with remaining unannotated transcripts (43% of all assembled sequences) representing a combination of coral specific genes [41] and genes that are too divergent from those in existing databases to identify using BLAST with reasonable cutoff parameters.

Functional Annotation

Oxidative Stress. Corals live in shallow, warm waters and house intra-cellular photosymbionts resulting in exposure to high levels of UV radiation and oxidative stress [8]. Excessive oxidative stress and UV exposure can damage nucleic acids [42,43] and these conditions may have selected for powerful and efficient DNA repair mechanisms in corals. Homologs of a majority of genes involved in the canonical pathway for oxidative stress response (as identified in model organisms) were present in the *A. palmata* transcriptome (Fig. 3) suggesting that corals may deal with high oxidative stress loads using many of the same biochemical mechanisms as the Bilateria.

DNA damage repair. Our data includes a coral homolog of the mutS gene which is involved in mismatch repair and is thought to be responsible for the low mtDNA mutation rate observed

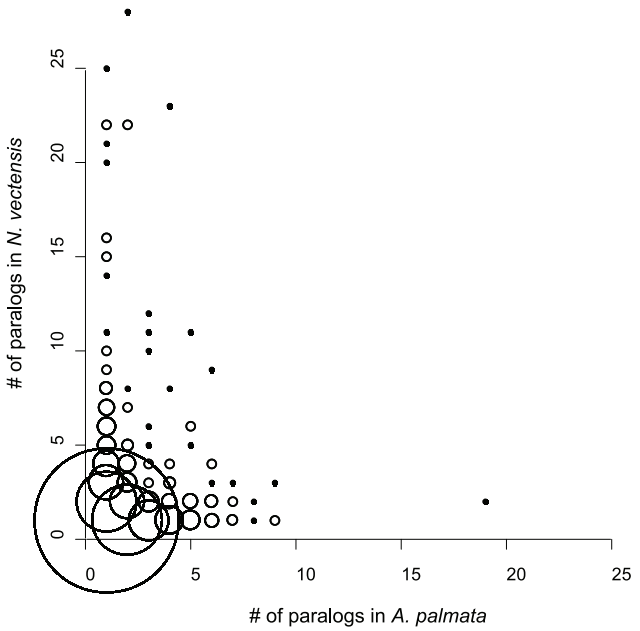

Figure 6. Distribution of paralog groups identified between *A. palmata* and *N. vectensis* by INPARANOID. Plotting on a gene by gene basis showed that genes with large numbers of paralogs (>5) differed between species. Groups with a single member from each species were most common, and Groups with >5 members each accounted for less than 0.5%. Only a single group with >9 members was found in *A. palmata*. Circle size is proportional to the number of groups of a given size class.
doi:10.1371/journal.pone.0028634.g006

Table 2. Paralog group size and identity for paralog groups identified in *Acropora palmata* as compared to the *Nematostella vectensis* proteome.

# of Paralogs	Descripton	Gene Function
19	IgLON family member	Immunoglobulin domain; Cell Adhesion Molecule
9	TNF receptor-associated factor 2/6	Signal transduction; Apoptosis
8	Egg protein	
7	Ferritin	Iron; Ironstorage; Metal-binding
7	Egg protein	
6	Triglyceride lipase	Hydrolase
6	GPI-linked carbonic anhydrase	Lyase
6	Chromoprotein	3D-structure; Chromophore; Luminescence; Photoprotein
6	Neuronal pentraxin-1	Calcium; Metal-binding
6	G-protein coupled receptor	Receptor
6	Myosin VIIA	ATP-binding; Motor protein; Myosin; Nucleotide-binding
6	Cupin family protein	
6	Dopamine beta-monooxygenase-like protein	Monooxygenase
6	TNF receptor-associated factor 3/6	Metal-binding; Signal transduction; Apoptosis
6	Egg protein	EGF-like domain
6	SH3 and PX domain- containing protein 2A/B	Cell projection
5	TNF receptor-associated factor 6	Metal-binding; Ubl conjugation; Signaling; Apoptosis
5	mab-21-like 1 (MAB21L1)	Cell proliferation
5	Sulfotransferase	Transferase
5	EGF like domain containing protein	
5	LPS-induced TNF-alpha factor	Transcription regulation; Apoptosis
5	Fucose binding lectin	Lectin
5	Putative SAM-dependent methyltransferase	Methyltransferase
5	CCAAT/Enhancer binding protein beta	DNA-binding
5	Skeletrophin, putative	Hydrolase; Metal-binding
5	Rho guanine nucleotide exchange factor 7	Guanine-nucleotide releasing factor; Phosphoprotein
5	Chromobox protein homolog 7	Chromatin regulator; Transcription regulation
5	Neuroblast differentiation associated protein	

doi:10.1371/journal.pone.0028634.t002

among cnidarians [44,45]. In addition we identified nearly 500 transcripts with GO terms relating to DNA repair mechanisms including MGMT, ATR/ATM, Rad52 family proteins, XP family genes, TFIIH and RPA complex members, as well as numerous transcription factors, polymerases and ligases involved in the various DNA repair pathways (Figs 3 & 4).

p53 Homologs. p53 proteins play a central role in the response to DNA damage in vertebrates by initiating pathways leading to apoptosis, or cell cycle arrest and DNA repair [46]. Canonical p53 is restricted to vertebrates, but ancestral homologs (p63/73) are found throughout the metazoans (Fig. 5), and some of the DNA repair related functions of p53 are also associated with p63 [46,47]. Surveys of the *N. vectensis* genome reveal three p53 family members (NVp63, pVS53a, and pEC53) [48] and homologs of all three variants were identified in *A. palmata*.

Divergence time between the Actinaria (anemones) and the Scleractinia (hard corals) is approximately 225 mya and divergence between the Cnidaria and Bilateria is estimated at 650–1000 mya [49]. During this time p53 family proteins have been duplicated and diversified yet their role in protecting the cell from genomic instability has been maintained [50]. p63's role in mediating apoptosis in germ line cells following DNA damage highlights its importance in mitigating genotoxic stress (Fig. 8) [48]. This function is important in long lived organisms exposed to high levels of UV radiation. Both *N. vectensis* and *A. palmata* are potentially very long lived, as genets can reproduce asexually more or less indefinitely. Thus, minimizing the impact of DNA damage to somatic and germ line cells is critical to survival and fitness. Indeed, our analysis suggests that p53 family proteins have experienced positive selection along the lineage separating Cnidaria from Bilateria (Fig. 5; Information S2). Additionally, the majority of genes involved in the canonical p53 mediated stress response (as characterized in model organisms) were present in *A. palmata* (Fig. 4).

Paralog Group Analysis

Paralog group analyses have proven informative in identifying gene and genome duplication events and how they relate to speciation in a variety of organisms [51,52]. We employed a conservative reciprocal best hit approach [53] to screen for duplicated genes in the *A. palmata* transcriptome. We found similar patterns in the group size distributions of paralogs in both the *N. vectensis* genome and the *A. palmata* transcriptome. While it is possible that additional paralogous copies of some genes were not

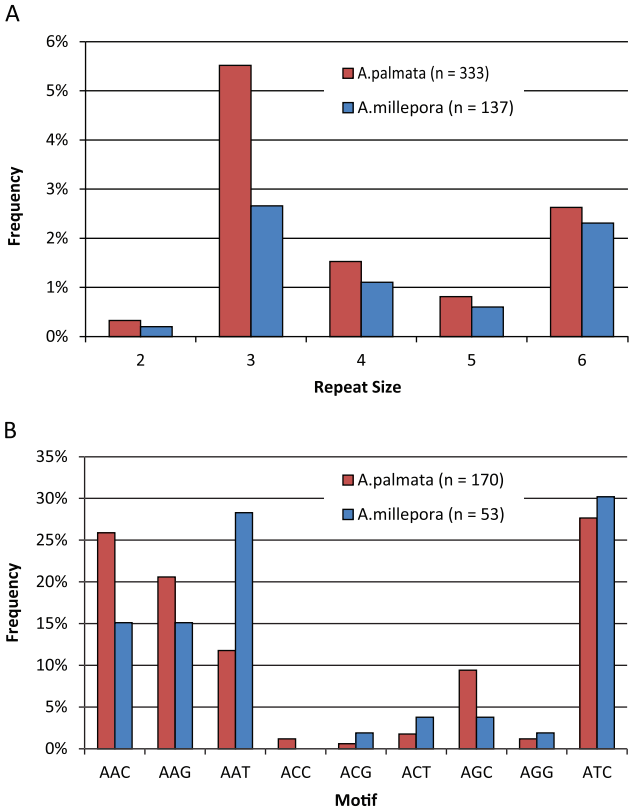

Figure 7. Microsatellite repeat distributions show similar patterns in *A. palmata* (red) and *A. millepora* (blue). A) Microsatellite frequency distributions showed an abundance of tri and hexanucleotide repeats in both species. B) Trinucleotide motifs were dominated by ATC repeats in both species, followed by varying proportions of AAN type repeats. GC rich repeats were rare, and no CCG repeats were found.
doi:10.1371/journal.pone.0028634.g007

sequenced, such missing transcripts add additional conservatism to this analysis.

Given the rarity of large paralog groups, members of such groups must serve important functions to justify the maintenance of multiple copies. The largest paralog group in *A. palmata* consisted of immunoglobulin superfamily proteins in the IgLON family involved in the organization of neuronal connections in the developing nervous system in other metazoans [54,55]. Interestingly, four paralogs of another group of Ig superfamily genes involved in nervous system development, the NCAMs (neural cell adhesion molecule like genes), have been identified in *N. vectensis* and are expressed in developing larvae and planulae [56]. Further investigation of these gene families in cnidarians is of interest because of their role in the developing nervous system [57].

The second most abundant paralog group consisted of homologs of the TRAF proteins. Several TRAF homologs showed multiple paralogs in *A. palmata*, including TRAF 2, 3, 6 and a lipopolysaccharide induced TNF alpha factor. In mammals TRAF proteins are important signal transducers mediating innate immune response, apoptosis, bone metabolism and response to stress (including DNA damage) [58,59].

These annotations were determined solely by bioinformatic means, thus further verification is required, but it is of interest to note that the six TRAF proteins found in vertebrates are thought to be the result of an evolutionarily recent diversification because few TRAF homologs have been detected in *Drosophila* and

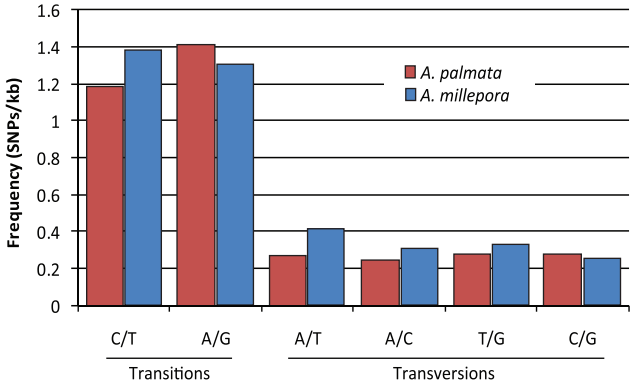

Figure 8. Frequency of various single nucleotide polymorphism (SNP) types in *A. palmata* (red) and *A. millepora* (blue). Frequencies are given per 1000 bp. Transitions are in red and transversions in blue. The overall frequency of SNPs in *A. palmata* was 1 per 272 bp.
doi:10.1371/journal.pone.0028634.g008

Caenorhabditis (n = 3 & 1 respectively) [60]. Thus the presence of multiple paralogs of the TRAF proteins in cnidarians may indicate more ancestral diversity in this protein family than previously acknowledged.

Genetic resources for *A. palmata*

Microsatellite Frequency Distribution. Microsatellite repeats commonly occur in portions of transcribed mRNA including 5′ and 3′ untranslated regions [61,62]. In the *A. palmata* transcriptome numerous microsatellite repeats were observed (n = 333), with tri and hexanucleotide motifs representing the most common types (Fig. 7A). This is true in coding sequences of many organisms [63,64,65] including *A. millepora* (pers. obs.) [66], because mutations in other size classes can result in frame shifts.

Among trinucleotide repeats, motifs with high AT content were most frequent in both acroporid transcriptomes (Fig. 7B). Specifically, ATC motifs were most numerous, followed by AAN type motifs. This is surprising because variants of some of these high frequency motif classes (ATC and AAT) can code for stop codons. There was a complete lack of CG and CCG variants and trinucleotides containing more than one G or C were low (<10%). This contrasts with findings from crop plants and vertebrates where motifs with higher GC content were frequent, but agrees with findings from other land plants, and fungi [63,64,65].

SNP Characterization. The frequency of SNPs in *A. palmata* was 1 per 272 bp (n = 33,433 SNPs in 14,613 contigs). The frequency in *A. millepora* was slightly higher (1 per 207 bp) [9] despite the greater depth of coverage and broader geographical origins of the *A. palmata* samples. Only a small set of SNPs were chosen for preliminary validation here, but their wide distribution and the presence of frequency differences in even a small sample size is encouraging (Information S4). Upon further validation and marker development, these SNPs will become a vital resource for studies of population structure and adaptive variation in *A. palmata* [67,68].

Transition Transversion Ratios. The transition bias is a phenomenon observed in vertebrates and invertebrates where despite a twofold higher probability of occurrence, transversions are far less frequent than transitions. The distribution of SNPs in *A. palmata* and *A. millepora* species was similar in terms of the proportion of transitions to transversions with some variation in the proportion of transition and transversion types (Fig. 8). While Ts/Tv ratios around 2 are considered common, this has only been

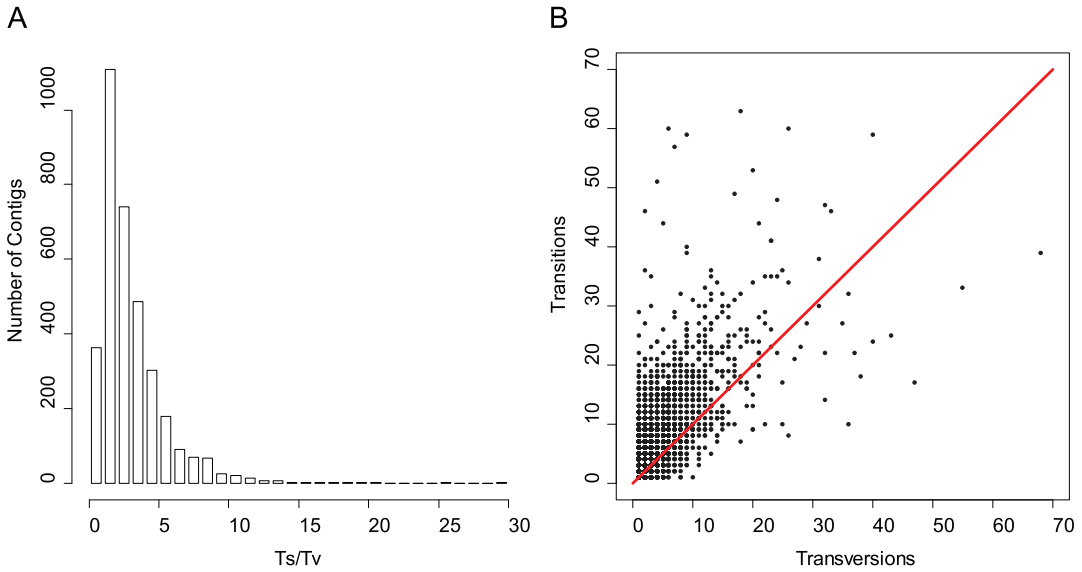

Figure 9. Variation in transition to transversion ratios (Ts/Tv) among contigs. A) The histogram of Ts/Tv ratios showed a number of contigs with ratios <1. B) Plotting the number of transitions (Ts) against the number of transversions (Tv) showed genes that have appreciable numbers of both SNP types. The diagonal line represents a ratio of 1, points below the line have a ratio <1 that may indicate a history of positive selection.
doi:10.1371/journal.pone.0028634.g009

evaluated in a small number of model organisms [69,70]. Recent findings suggest that some non-model species may not display such transition biases [71]. The transcriptome wide Ts/Tv ratio of 2.4 in *A. palmata* suggests that the transition bias does exist in this species.

Transversions are more likely to be removed by selection because there is a greater probability that they will result in an amino acid altering change. Thus deviations from a ratio of two may act as a rough test for the presence of selection [72,73]. Transcripts with the highest 5% of ratios (Ts/Tv>5), a potential indicator of positive selection, were enriched for GO terms associated with nucleic acid metabolism and pigmentation. Transcripts with the lowest 5% of ratios, a potential indicator of purifying selection, were enriched for carbohydrate and protein metabolism functions as well as the formation of structural molecules (Table 3; Information S3). While this is only a rough test, these transcripts may serve as a useful starting point for investigations of targets of selection in corals. Indeed, the processes highlighted here are of primary importance when considering the unique metabolic adaptations of corals to their dinoflagellate symbionts. Specifically, the coral host metabolizes not only the life sustaining carbohydrates, but also the damaging oxygen radicals that are released by the photosynthesizing algae [74]. These processes occur in conjunction with exposure to high light and UV levels [75] necessitating a simultaneous response to light and heat stress which likely involves pigmentation molecules and protein chaperones in addition to a suite of metabolic and regulatory genes.

Conclusion

Our analyses highlighted many processes conserved between corals and more complex bilaterian animals that contribute to corals' ability to maintain genomic integrity despite exposure to high levels of UV, ROS and prolonged life spans. Enrichment of these functional categories in the *A. palmata* transcriptome underscores their importance to coral survival and fitness.

This dataset provides a comprehensive look at the genes expressed in *A. palmata*. The resources developed from this dataset provide tools for coral researchers and will help generate insights into functional diversity within and among wild populations. More importantly perhaps, this sequence data enables scans for selection to identify functional variation in coral genes that contribute to their ability to adapt to changing environmental conditions.

Methods

Collection and CITES export permits (where applicable) were obtained from the local authorities (Puerto Rico Department of Natural and Environmental Resources, the Florida Keys National Marine Sanctuary, the Florida Fish and Wildlife Commission, Caribbean Research & Management of Biodiversity in Curacao) for all samples used in this study.

Larval Experimental Treatments and Sampling

Acropora palmata adults contain intracellular symbionts known as zooxanthellae. Like many other corals, symbionts are taken up after settlement and metamorphosis of the planula larvae [26]. To enrich for host genetic material and avoid a large contribution of symbiont transcripts larval tissues were targeted for RNA extraction.

To obtain larval tissue, gametes were collected from adult colonies as they were released during annual spawning events in 2008. Larvae were acquired from colonies at multiple locations in the upper Florida Keys, Puerto Rico, and Curacao to minimize ascertainment bias [76,77]. To generate batch crosses, gametes from multiple genets (as determined by microsatellite genotyping following Baums *et al.* [23]) were combined and incubated for one hour to ensure fertilization then rinsed in filtered sea water. The resulting embryos were incubated in aquaria simulating environmental stress conditions including three temperatures and elevated CO_2 levels.

Low, mean, and elevated temperatures (25, 27, and 30°C respectively) were used to stimulate the expression of thermal stress response genes. Larval batches were housed in 1 L plastic containers with mesh sides, suspended in four separate 45 L plastic bins filled with filtered sea water at each treatment

Table 3. Significantly enriched GO terms associated with transcripts showing especially low (<1) Ts/Tv ratios (Upper) and high (>5) Ts/Tv ratios (Lower).

Ts/Tv Class	Primary GO Level	n (K<1)	n (total)	% (K<1)	% (total)	p value	GO ID	GO term
Low (K<1)	Biological Process:	64	681	7.6	9.8	0.05	GO:0006807	nitrogen compound metabolic process
		56	609	6.7	8.7	0.04	GO:0006139	nucleic acid metabolic process
		124	1296	14.8	18.6	0.01	GO:0044237	cellular metabolic process
		69	774	8.2	11.1	0.01	GO:0043473	pigmentation
		69	748	8.2	10.7	0.03	GO:0050794	regulation of cellular process
		69	773	8.2	11.1	0.01	GO:0050789	regulation of biological process
	Cellular Component:	211	2124	25.1	30.4	0.00	GO:0044464	cell part
		127	1423	15.1	20.4	0.00	GO:0044424	intracellular part
		81	963	9.7	13.8	0.00	GO:0043231	intracellular membrane-bounded organelle
		10	177	1.2	2.5	0.02	GO:0044428	nuclear part
		74	820	8.8	11.7	0.01	GO:0005737	cytoplasm
	Molecular Function:	250	2511	29.8	36	0.00	GO:0005488	binding
		75	806	8.9	11.5	0.02	GO:0005515	protein binding
Ts/Tv Class	**Primary GO Level**	**n (K>5)**	**n (total)**	**% (K>5)**	**% (total)**	**p value**	**GO ID**	**GO term**
High (K>5)	Biological Process:	72	649	11.8	9.3	0.05	GO:0009058	biosynthetic process
		61	514	10	7.4	0.02	GO:0010467	gene expression
		108	1010	17.6	14.5	0.03	GO:0044260	cellular macromolecule metabolic process
		134	1296	21.9	18.6	0.04	GO:0044237	cellular metabolic process
		3	106	0.5	1.5	0.04	GO:0044262	cellular carbohydrate metabolic process
		96	875	15.7	12.5	0.03	GO:0019538	protein metabolic process
		37	301	6	4.3	0.05	GO:0080090	regulation of primary metabolic process
	Cellular Component:	152	1423	24.8	20.4	0.01	GO:0044424	intracellular part
		123	1178	20.1	16.9	0.04	GO:0043226	organelle
		65	521	10.6	7.5	0.01	GO:0044444	cytoplasmic part
	Molecular Function:	17	110	2.8	1.6	0.03	GO:0005198	structural molecule activity
		88	806	14.4	11.5	0.04	GO:0005515	protein binding

Low Ts/Tv ratios may be a rough signal of positive selection, while elevated ratios could indicate purifying selection.
doi:10.1371/journal.pone.0028634.t003

temperature. Water was circulated with an aquarium pump and changed daily with filtered sea water preheated to the target temperature. Temperatures were maintained within $\pm 1^{\circ}C$ with aquarium heaters and chillers, and were monitored with HOBO data loggers (Onset Computer Corp., MA).

Larval samples were collected throughout development, from immediately following fertilization until day five, to include a full range of developmental genes in the sequencing results. Approximately 100 larvae from each container at each time-point were incubated in RNA later (Ambion, TX) then frozen in liquid nitrogen following manufacturer's recommendations. Samples were stored at $-80^{\circ}C$ prior to extraction. Additional samples from adult colonies were also collected to include a component of post metamorphosis transcripts.

For the CO_2 treatments, larvae from Florida were raised in the Climate Change Laboratory at the Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL. Water in the treatment aquaria was held at $28^{\circ}C$ and CO_2 concentration was manipulated by bubbling with CO_2-enriched air to maintain control and high CO_2 treatment conditions (400 and 800 µatm respectively) as described in Albright et al. [78]. Larvae were collected at 12 hour intervals from 36 to 84 hours of development.

RNA extraction and Sequencing

Purification of RNA from all samples was performed using a modified Trizol extraction protocol. Samples were removed from the preservative and immediately submerged in 1 ml of Trizol for 5 minutes. Following the addition of 0.2 ml of Chloroform and incubation for 3 minutes, the samples were centrifuged at 12,000 rpm for 15 minutes at $4^{\circ}C$ to separate the phases. The upper aqueous phase was collected and mixed with an equal volume of 70% ethanol, then applied to a Qiagen RNeasy mini spin column (Qiagen, CA) and purified following manufacturer's instructions. Sample concentration was determined using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, FL)

and RNA quality was assessed with an Agilent Bioanalyzer (Agilent Technologies, CA).

A total of 71 samples, including 1 to 5 replicates each of 13 larval developmental time-points from all 3 locations, 10 CO_2 treatment samples, and 4 adult samples from Puerto Rico and Curacao were pooled into a single batch sample such that the contribution of genetic material from each geographic location was equivalent (~17 ug each). An aliquot of the resulting pool was submitted to the Indiana University Center for Genomics and Bioinformatics (Bloomington, IN) for cDNA preparation, normalization and sequencing on the 454-GS FLX using Titanium chemistry following previously published methods [9,79].

Assembly and Annotation

The resulting sequencing reads were preprocessed and annotated using a custom PERL pipeline (PIPEMETA) following the methods of Vera *et al.* [10]. Briefly, 454 sequences were quality trimmed to filter out reads of dubious quality (>1 N, average quality <20, length <280 or >530) and screened for primer sequence using the SMARTSCREENER script in the PIPEMETA package (SMART "CAP" primer 5′-AAGCAGTGGTATCAACGCAGAGT-3′), then entered into SEQMAN PRO v8 (DNAstar) for additional quality filtering and assembly. Default quality filtering parameters for 454 data were used, and assembly was performed in SEQMAN using parameters suggested by the manufacturer for short read data. Prior to assembly an additional 36,239 high quality EST Sanger reads from a pool of *A. palmata* samples from different life stages and stress conditions were included to promote the assembly of long contigs. Sanger sequences are available from the SymBioSys database (sequoia.ucmerced.edu/SymBioSys/). 454 sequences and quality scores generated in this study are available from the NCBI short read archive (Accession #: SRP006958; http://www.ncbi.nlm.nih.gov/sra), and the assembled annotated sequences can be downloaded from http://main.g2.bx.psu.edu/u/nickpolato/h/apalmataassembly.

The resulting contigs and remaining singletons were aligned with BLASTX to the complete Uniprot database (release 2010_05; www.uniprot.org/downloads). Acceptable hits were determined by a bitscore of >45 and a corresponding e-value of $<1^{-5}$. Because two of the most closely related genomes in publicly available databases (*Nematostella vectensis* and *Branchiostoma floridae*) are largely unannotated, a large proportion of the highest ranking BLAST hits were uninformative. To maximize the information content, results were augmented with lower ranking (but more informative) annotations using a PERL script and a list of keywords to avoid (i.e. 'uncharacterized protein' and 'predicted protein'). Contig and singleton sequences were also submitted to the KEGG automated annotation server (www.genome.ad.jp/tools/kaas/) for further functional annotation.

Transcriptome Completeness

A BLAST search was performed comparing all *A. palmata* transcriptome data against a subset of transcripts from the *N. vectensis* genome project ("*transcripts_Nemve1FilteredModels1*" downloaded from genome.jgi-psf.org). The *N. vectensis* sequences were first compared to themselves to eliminate multiple copies of sequences with >90% similarity (i.e. multi copy genes). Only BLAST hits between *A. palmata* and *N. vectensis* sequences with a bitscore ≥45 were considered. The length of the *A. palmata* sequence was then divided by the length of the *N. vectensis* transcript to obtain a percent coverage estimate for each unigene (Information S1), using only the best unique hit for each query and subject. This method provides a reasonable estimate of coverage because if only partial transcript sequences were obtained for the majority of *A. palmata* genes one would expect an abundance of low length ratios. Additionally, a set of 119 orthologs conserved across the metazoans and single copy in the cnidarians (*Hydra magnipapillata* and *N. vectensis*) was downloaded from OrthoDB (cegg.unige.ch/orthodb4) [80]. BLAST was used to identify homologs of these genes in the *A. palmata* data.

Functional Analysis

Pathway analysis was performed in the Ingenuity Pathway Analysis (IPA) software. IPA is a web-based application that performs functional enrichment analyses to determine the probability that a given gene set is associated with pre-defined reference pathways beyond what would be expected by random chance. Pathway data in IPA is based on information in the Ingenuity Knowledge Base which is a manually reviewed database of pathways and relationships taken directly from the primary literature and public sources including GO, KEGG and EntrezGene. While this software is capable of other functions that consider gene expression levels based on microarray or RNAseq data, our analyses used only presence or absence of pathway components in the *A. palmata* dataset to compare with functional pathway relationships known in other model organisms. IPA was then used to test whether some pathways were more highly represented then others considering the number of sequences from *A. palmata* that mapped to a given pathway and the pathways' size. This application of IPA did not depend on expression level (i.e. the number of sequencing reads observed per gene). This is appropriate here because the *A. palmata* transcriptome was generated from a normalized library. IPA bases its statistical analyses on a Fisher's Exact Test (corrected for multiple testing) which is comparable to other well-known enrichment analysis methods [81]. However experimental evidence has shown that it performs better on large datasets than other similar methods [82].

A canonical pathway analysis identified those pathways in the Ingenuity Knowledge Base that were the most well represented in the *A. palmata* dataset. Our analyses compared successfully mapped *A. palmata* transcripts to the IPA reference set consisting of all molecules present in the Ingenuity Knowledge Base. Significance of the association between the *A. palmata* data set and a reference pathway was measured in two ways: 1) A ratio was calculated by dividing the number of *A. palmata* transcripts that map to a given reference pathway by the total number of molecules associated with that pathway in the Ingenuity Knowledge Base, and 2) A Fisher's exact test determined the probability that there was a significant enrichment of *A. palmata* transcripts mapping to a given pathway beyond what was expected by chance alone given the total number of genes involved in that reference pathway. Because some pathways are better characterized than others these results should be interpreted as a guide for selecting conserved and well annotated functional metazoan pathways present in the *A. palmata* transcriptome.

Testing for Natural Selection in p53 family genes. Tests of natural selection in p53 family genes were performed with the CODEML program in PAML [38]. 498 bp of the conserved DNA binding domain from two *A. palmata* homologs of *N. vectensis* genes (pVS53a and p63) were aligned with 17 other sequences representing a variety of taxa from anemones to humans (see Information S2 for gen bank accession numbers) using the global homology strategy in MAFFT (mafft.cbrc.jp/alignment/server/index.html). JMODELTEST [83] was used to determine the appropriate substitution matrix and a maximum likelihood phylogenetic tree was generated using GARLI (www.molecularevolution.org/software/phylogenetics/garli) with the

TIM2+G model. One hundred bootstrap replicates were performed and bootstrap values were matched to the maximum likelihood tree using SUMTREES [84]. The branch-site model in PAML [85] was used to test for evidence of selection along the branch separating cnidarian and bilaterian sequences. Tests considered different dN/dS ratios (i.e. the ratio of non-synonymous substitutions per non-synonymous site (dN) to synonymous substitutions per synonymous site (dS)) across the phylogenetic tree. A likelihood ratio test ($2\Delta\ln L = 2(\ln L_{h0} - \ln L_{h1})$) of the difference between the log likelihood of the null model ($\ln L_{h0}$; NSsites = 1a) relative to that of the alternative model ($\ln L_{h1}$; NSsites = 2a), that incorporated positive selection along the branch leading to the Cnidaria by allowing the dN/dS ratio to be >1 along that branch, was performed and the result compared to a χ^2 distribution to assess statistical significance.

Paralog Analysis

The identification of in-paralog groups in *A. palmata* and *N. vectensis* was assessed using INPARANOID v4.1 [53]. A set of 27,273 filtered protein models from *N. vectensis* (downloaded from genome.jgi-psf.org), and 43,081 ORFs from the *A. palmata* transcriptome were used in the analysis. Acceptable hits were determined by a bitscore of >40 using the BLOSUM62 scoring matrix.

Development of Genetic Resources

Potential SNPs were detected in contigs with sufficient depth of coverage using the SNPHUNTER script in PIPEMETA [10]. The resulting annotation and SNP data were archived in a searchable MYSQL database and will be made available to the public via the Dryad data archive (http://datadryad.org/) along with sequences containing potential microsatellite markers.

To confirm the results obtained by SNPHUNTER SNPs were validated using PCR and the amplified products were sequenced on an ABI3700. Primers were designed using PRIMER3 (frodo.wi.mit.edu/primer3/; Information S4) to amplify portions of six contigs containing putative SNPs in a sample population which consisted of 11 individuals from Florida, 8 from Curacao, and 7 from Puerto Rico. Equal volumes (0.2 µl) of forward and reverse primers (5 uM each) were combined with 20–200 ng of template DNA, 1× NH_4 Reaction Buffer (Bioline, MA), 2 mM of $MgCl_2$, 0.2 mM of dNTPs, and 2 U of Biolase polymerase (Bioline, MA). PCR was carried out in an Eppendorf Mastercyler Gradient with an initial denaturation step of 95°C for 5 min followed by 35 cycles of 95°C for 20 s; annealing at 56°C for 20 s; and 72°C for 30 s, and a final extension of 30 min at 72°C.

Transition and transversion ratios were determined based on the number of SNPs in each class identified by SNPHUNTER for each contig. Enrichment analysis of GO terms associated with transcripts showing the top and bottom 5% of all Ts/Tv ratios (<1 or >5) was carried out using WEGO (http://wego.genomics.org.cn/cgi-bin/wego/index.pl) [86].

To detect repeat elements, the full set of sequences was uploaded to the Tandem Repeats Database (http://tandem.bu.edu/cgi-bin/trdb/trdb.exe) [87] and searched using TANDEM REPEATS FINDER v4.04 [88]. Repeats between 2 and 6 bp were considered and comparisons were made with microsatellite frequencies from the publicly available *A. millepora* transcriptome (www.bio.utexas.edu/research/matz_lab/matzlab/Data.html) [9].

Supporting Information

Information S1 Distributions of log length ratios for *A. palmata* and *N. vectensis* transcripts, and fragmentation levels as compared to a set of 119 single copy orthologs.
(XLSX)

Information S2 Results of tests for natural selection using the program CODEML.
(XLSX)

Information S3 Top and bottom 5% of Ts/Tv ratios.
(XLS)

Information S4 SNP validation summary and primer sequences.
(XLSX)

Acknowledgments

We thank M. Medina and the members of her Lab at UC Merced for sharing their collection of *A. palmata* ESTs to include in our assembly. C. Praul and colleagues at the Penn State Genomic Core facility provided valuable advice regarding sample preparation and bioinformatic analysis. Transcriptome sequencing was performed at the Center for Genomics and Bioinformatics (CGB) at Indiana University, under the direction of J. Colbourne and K. Mockaitis. Assembly and bioinformatics support was provided by H. Tang and J.H. Choi. The CGB is supported in part by the METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. Thanks also to M. Matz and E. Meyer for providing *A. millepora* sequence data for comparison purposes. We thank M. Miller and A. Valdivia of NOAA and the members of SECORE for assistance with spawn collection and rearing, and R. Albright, B. Mason and C. Langdon for sharing samples raised under CO_2 treatment conditions. Thanks also to the members of the Baums Lab at Penn State for assistance in the field and lab.

Author Contributions

Conceived and designed the experiments: NP IB. Performed the experiments: NP IB. Analyzed the data: NP JV. Contributed reagents/materials/analysis tools: NP IB. Wrote the paper: NP. Designed the software used in analysis: JV.

References

1. Martindale MQ, Hejnol A (2009) A developmental perspective: changes in the position of the blastopore during bilaterian evolution. Developmental cell 17: 162–174.
2. Technau U, Rudd S, Maxwell P, Gordon PM, Saina M, et al. (2005) Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet 21: 633–639.
3. Miller DJ, Ball EE, Technau U (2005) Cnidarians and ancestral genetic complexity in the animal kingdom. Trends in Genetics 21: 536–539.
4. Allemand D, Ferrier-Pagès C, Furla P, Houlbrèque F, Puverel S, et al. (2004) Biomineralisation in reef-building corals: from molecular mechanisms to environmental control. Comptes Rendus Palevol 3: 453–467.
5. Bessat F, Buigues D (2001) Two centuries of variation in coral growth in a massive Porites colony from Moorea (French Polynesia): a response of ocean-atmosphere variability from south central Pacific. Palaeogeography, Palaeoclimatology, Palaeoecology 175: 381–392.
6. De'ath G, Lough JM, Fabricius KE (2009) Declining coral calcification on the Great Barrier Reef. Science 323: 116.
7. Wallace C (1985) Reproduction, recruitment and fragmentation in nine sympatric species of the coral genus Acropora. Marine Biology 88: 217–233.
8. Shick J, Lesser MP, Jokiel PL (1996) Effects of ultraviolet radiation on corals and other coral reef organisms. Global Change Biology 2: 527–545.
9. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC genomics 10: 219.
10. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology 17: 1636–1647.
11. Ekblom R, Galindo J (2010) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity;doi:10.1038/hdy.2010.152.

12. Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. Trends in Ecology & Evolution 21: 629–637.
13. Romanov M, Tuttle E, Houck M, Modi W, Chemnick L, et al. (2009) The value of avian genomics to the conservation of wildlife. BMC genomics 10: S10.
14. Bernatchez L, Renaut S, Whiteley AR, Derome N, Jeukens J, et al. (2010) On the origin of species: insights from the ecological genomics of lake whitefish. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 1783.
15. Chapman RW (2001) EcoGenomics–a consilience for comparative immunology? Developmental and comparative immunology 25: 549.
16. Baums IB (2008) A restoration genetics guide for coral reef conservation. Molecular ecology 17: 2796–2811.
17. Aronson R, Bruckner A, Moore J, Precht B, Weil E (2008) Acropora palmata. in: IUCN 2010 IUCN Red List of Threatened Species Version 2010.4.
18. Hatcher BG (1988) Coral reef primary productivity: A beggar's banquet. Trends in Ecology & Evolution 3: 106.
19. Rogers CS, Salesky NH (1981) Productivity of Acropora palmata (Lamarck), macroscopic algae, and algal turf from Tague Bay reef, St. Croix, US Virgin Islands. Journal of Experimental Marine Biology and Ecology 49: 179–187.
20. Vollmer SV, Palumbi SR (2006) Restricted gene flow in the Caribbean staghorn coral Acropora cervicornis: Implications for the recovery of endangered reefs. Journal of Heredity.
21. Foster NL, Baums IB, Mumby PJ (2007) Sexual versus asexual reproduction in an ecosystem engineer: the massive coral Montastraea annularis. Journal of Animal Ecology 76: 384–391.
22. Foster N, Paris C, Kool J, Baums I, Stevens J, et al. (in review) Complementary insights into coral connectivity from empirical and modelled gene flow. Molecular Ecology.
23. Baums IB, Miller MW, Hellberg ME (2005) Regionally isolated populations of an imperiled Caribbean coral, Acropora palmata. Molecular Ecology 14: 1377–1390.
24. Baums IB, Miller MW, Hellberg ME (2006) Geographic variation in clonal structure in a reef building Caribbean coral, Acropora palmata. Ecological Monographs 76: 503–519.
25. Kortschak RD, Samuel G, Saint R, Miller DJ (2003) EST Analysis of the Cnidarian Acropora millepora Reveals Extensive Gene Loss and Rapid Sequence Divergence in the Model Invertebrates. Current Biology 13: 2190–2195.
26. Schwarz JA, Brokstein PB, Voolstra C, Terry AY, Manohar CF, et al. (2008) Coral life history and symbiosis: functional genomic resources for two reef building Caribbean corals, Acropora palmata and Montastraea faveolata. BMC Genomics 9: 97.
27. Voolstra CR, Schwarz JA, Schnetzer J, Sunagawa S, Desalvo MK, et al. (2009) The host transcriptome remains unaltered during the establishment of coral-algal symbioses. Mol Ecol 18: 1823–1833.
28. Voolstra CR, Sunagawa S, Matz MV, Bayer T, Aranda M, et al. (2011) Rapid Evolution of Coral Proteins Responsible for Interaction with the Environment. PloS one 6: e20392.
29. Iguchi A, Shinzato C, Forêt S, Miller DJ (2011) Identification of Fast-Evolving Genes in the Scleractinian Coral Acropora Using Comparative EST Analysis. PloS one 6: e20140.
30. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, et al. (2010) The dynamic genome of Hydra. Nature 464: 592–596.
31. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317: 86–94.
32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389.
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nature genetics 25: 25–29.
34. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071.
35. Lushchak VI (2011) Environmentally induced oxidative stress in aquatic animals. Aquatic Toxicology 101: 13–30.
36. Lewis KN, Mele J, Hayes JD, Buffenstein R (2010) Nrf2, a Guardian of Healthspan and Gatekeeper of Species Longevity. Integrative and comparative biology 50: 829.
37. Nguyen T, Nioi P, Pickett CB (2009) The Nrf2-antioxidant response element signaling pathway and its activation by oxidative stress. Journal of Biological Chemistry 284: 13291.
38. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
39. Allemand D, Tambutté É, Zoccola D, Tambutté S (2011) Coral Calcification, Cells to Reefs. Coral Reefs: An Ecosystem in Transition. pp 119–150.
40. Moya A, Tambutté S, Bertucci A, Tambutté E, Lotto S, et al. (2008) Carbonic Anhydrase in the Scleractinian Coral Stylophora pistillata. Journal of Biological Chemistry 283: 25475.
41. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG (2009) More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics 25: 404–413.
42. Baruch R, Avishai N, Rabinowitz C (2005) UV incites diverse levels of DNA breaks in different cellular compartments of a branching coral species. Journal of experimental biology 208: 843.
43. Nesa B, Hidaka M (2008) Thermal stress increases oxidative DNA damage in coral cell aggregates. Proceedings of the 11th International Coral Reef Symposium 1: 149–151.
44. Pont-Kingdon GA, Okada NA, Macfarlane JL, Beagley CT, Wolstenholme DR, et al. (1995) A coral mitochondrial mutS gene. Nature 375: 109.
45. Hellberg ME (2006) No variation and low synonymous substitution rates in coral mtDNA despite high nuclear variation. BMC Evolutionary Biology 6: 24.
46. Rutkowski R, Hofmann K, Gartner A (2010) Phylogeny and function of the invertebrate p53 superfamily. Cold Spring Harbor Perspectives in Biology 2.
47. Suh EK, Yang A, Kettenbach A, Bamberger C, Michaelis AH, et al. (2006) p63 protects the female germ line during meiotic arrest. Nature 444: 624–628.
48. Pankow S, Bamberger C (2007) The p53 tumor suppressor-like protein nvp63 mediates selective germ cell death in the sea anemone Nematostella vectensis. PLoS One 2: e782.
49. Darling JA, Reitzel AR, Burton PM, Mazza ME, Ryan JF, et al. (2005) Rising starlet: the starlet sea anemone, Nematostella vectensis. Bioessays 27: 211–221.
50. Belyi VA, Ak P, Markert E, Wang H, Hu W, et al. (2010) The origins and evolution of the p53 family of genes. Cold Spring Harbor Perspectives in Biology 2.
51. Koonin EV (2005) Orthologs, Paralogs, and Evolutionary Genomics. Annu Rev Genet 39: 309–338.
52. Smith JJ, Putta S, Zhu W, Pao GM, Verma IM, et al. (2009) Genic regions of a large salamander genome contain long introns and novel genes. BMC genomics 10: 19.
53. Remm M, Storm CEV, Sonnhammer ELL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. Journal of molecular biology 314: 1041–1052.
54. Fusaoka E, Inoue T, Mineta K, Agata K, Takeuchi K (2006) Structure and function of primitive immunoglobulin superfamily neural cell adhesion molecules: a lesson from studies on planarian. Genes to Cells 11: 541–555.
55. Walmod PS, Pedersen MV, Berezin V, Bock E (2007) Cell Adhesion Molecules of the Immunoglobulin Superfamily in the Nervous System. Handbook of neurochemistry and molecular neurobiology: Neural protein metabolism and function. 35 p.
56. Marlow HQ, Srivastava M, Matus DQ, Rokhsar D, Martindale MQ (2009) Anatomy and development of the nervous system of Nematostella vectensis, an anthozoan cnidarian. Developmental neurobiology 69: 235–254.
57. Ingber DE (2003) Tensegrity II. How structural networks influence cellular information processing networks. Journal of Cell Science 116: 1397–1408.
58. Bradley JR, Pober JS (2001) Tumor necrosis factor receptor-associated factors (TRAFs). Oncogene 20: 6482.
59. Chung JY, Park YC, Ye H, Wu H (2002) All TRAFs are not created equal: common and distinct molecular mechanisms of TRAF-mediated signal transduction. Journal of cell science 115: 679–688.
60. Grech A, Quinn R, Srinivasan D, Badoux X, Brink R (2000) Complete structural characterisation of the mammalian and Drosophila TRAF genes: implications for TRAF evolution and the role of RING finger splice variants. Molecular Immunology 37: 721–734.
61. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual Review of Genetics 44: 445–477.
62. Goldstein DB, Schlötterer C (1999) Microsatellites: evolution and applications. Oxford, UK. 11–14: Oxford Press.
63. Young ET, Sloan JS, Van Riper K (2000) Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. Genetics 154: 1053.
64. Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research 10: 967.
65. Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular and Molecular Biology Letters 7: 537–546.
66. Wang S, Zhang L, Matz M (2009) Microsatellite characterization and marker development from public EST and WGS databases in the reef-building coral Acropora millepora (Cnidaria, Anthozoa, Scleractinia). Journal of Heredity 100: 329.
67. Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology 14: 671–688.
68. Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19: 208–216.
69. Petrov DA, Hartl DL (1999) Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America 96: 1475.
70. Denver DR, Morris K, Lynch M, Thomas WK (2004) High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. Nature 430: 679–682.
71. Keller I, Bensasson D, Nichols RA (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3: e22.
72. Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon(Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery. BMC genomics 10: 203.

73. Zhang J (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. Journal of molecular evolution 50: 56–68.
74. DeSalvo MK, Voolstra CR, Sunagawa S, Schwarz JA, Stillman JH, et al. (2008) Differential gene expression during thermal stress and bleaching in the Caribbean coral Montastraea faveolata. Mol Ecol 17: 3952–3971.
75. Levy O, Achituv Y, Yacobi Y, Dubinsky Z, Stambler N (2006) Diel 'tuning' of coral metabolism: physiological responses to light cues. Journal of experimental biology 209: 273.
76. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome research 15: 1496.
77. Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168: 2373.
78. Albright R, Mason B, Miller M, Langdon C (2010) Ocean acidification compromises recruitment success of the threatened Caribbean coral Acropora palmata. PNAS 107: 20400–20404.
79. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
80. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic acids research 39: D283.
81. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research 37: 1.
82. Hong MG, Pawitan Y, Magnusson PKE, Prince JA (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Human genetics 126: 289–301.
83. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25: 1253–1256.
84. Sukumaran J, Holder MT (2010) DendroPy: A Python library for phylogenetic computing. Bioinformatics 26: 1569–1571.
85. Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19: 49–57.
86. Ye J, Fang L, Zheng H, Zhang Y, Chen J, et al. (2006) WEGO: a web tool for plotting GO annotations. Nucleic acids research 34: W293.
87. Gelfand Y, Rodriguez A, Benson G (2006) TRDB–The Tandem Repeats Database. Nucleic Acids Research.
88. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 27: 573.