Supporting Information

Barshis et al. 10.1073/pnas.1210224110

SI Materials and Methods

mRNA Extraction. Total RNA was extracted from each sample using a modified TRIzol (GibcoBRL/Invitrogen) protocol. Approximately 150–200 mg of coral tissue and skeleton was placed in 1 mL of TRIzol and homogenized for 2 min by vortexing with ∼100 μL of 0.5-mm Zirconia/Silica Beads (BioSpec Products). Resulting tissue/TRIzol slurry was removed by centrifugation, and the standard TRIzol extraction was performed according to manufacturer's specifications with the replacement of 250 μL of 100% (vol/vol) isopropanol with 250 μL of high-salt buffer (0.8 M Na citrate, 1.2 M NaCl) during the final precipitation step. Resulting RNA pellet was resuspended in 12 μL of diethylpyrocarbonate (DEPC)-treated H_2O. mRNA was isolated from total RNA using the Micro-FastTrack mRNA isolation kit (Invitrogen) and an overnight precipitation at −80 °C. Between 40 ng and 1 μg of mRNA was used in Illumina library construction as in Beck et al. (1), but random hexamer primers were used to increase transcriptome coverage.

mRNA-Sequencing Lengths. Seven of the 31 libraries were sequenced by Illumina with a 76-bp paired-end sequencing length (152 bp per sequence total), four libraries were sequenced using single-end sequencing and a length of 36 bp in the laboratory of Arend Sidow (Stanford University, Palo Alto, CA), and the remaining 20 libraries were sequenced by Eureka Genomics. The latter 20 libraries were all done with single-end sequencing; three were 72-bp reads, and the rest were 36 bp. An additional 36-bp paired-end lane was run for four of these libraries at Stanford University's functional genomics facility. These four additional lanes generated few reads because of concentration problems, but are still incorporated in the analyses.

Coral de Novo Transcriptome Assembly and Annotation. Poor-quality portions (quality score, <20) were trimmed from the ends of the raw sequences using the *FASTX-Toolkit*, and any reads of <20 bp were discarded (fastq_quality_trimmer -t 20 -l 20; http://hannonlab.cshl.edu/fastx_toolkit). The *FASTX-Toolkit* was also used to remove any potential adapter sequences (fastx_clipper -l 20 -n). After quality and adaptor trimming, all lanes of Illumina sequence data (n = 35) were assembled, de novo, using CLC Genomics Workbench (Version 4; CLC Bio) with the following parameters: mismatch cost of 1, insertion and deletion costs of 2, length fraction of 0.27, similarity of 0.8, paired-end distance range of 1–750, single-end limit of 5, voting conflict resolution, random assignment of nonspecific matches and minimum contig length of 200. Putative coral sequences were identified through nucleotide sequence similarity (BLASTN) to a wide array of Cnidarian cDNA databases: *Acropora hyacinthus* and *Acropora millepora* larval ESTs from the laboratory of M. V. Matz (The University of Texas, Austin, TX) (2), *A. millepora* larval ESTs from the laboratory of David John Miller (James Cook University, Townsville, Australia) (3), predicted transcripts from the *Acropora digitifera* genome ("adi_v1.0.1.cdna.fa.gz" downloaded from maringenomics.oist.jp) (4) and predicted transcripts from the *Nematostella vectensis* genome ("transcripts. Nemve1FilteredModels1" downloaded from http://genome.jgi-psf.org) (5). Contigs were identified as putatively coral if they matched one or more of these databases with a hit of ≥100 bp and with ≥85% sequence identity to *Acropora* or ≥75% identity to *Nematostella*. Ribosomal RNA contamination was then removed based on significant-nucleotide similarity (BLASTN, e value, $\leq 1e^{-8}$) to the SILVA rRNA database ("lsu-parc.fasta" and "ssu-parc.fasta" downloaded on September 14, 2011 from www.arb-silva.de), and potential *Symbiodinium* contamination was removed based significant nucleotide similarity (BLASTN, ≥100 bp and ≥70% identity) to ESTs from *Symbiodinium sp*. KB8 (clade A) and *Symbiodinium sp*. MF1.04b (clade B) (6) and the related dinoflagellate *Polarella glacialis*. Finally, the resulting contigs were compared (via BLASTX) to the NCBI nonredundant protein database (nr; downloaded on June 7, 2011 from www.ncbi.nlm.nih.gov). The nonredundant (nr) results were used to remove any additional sequences likely to be noncoral, based on similarity to alveolates, fungi, bacteria, or Archaea, as determined using the metagenome analyzer (MEGAN) Version 4 (min. support = 1, min. score = 200, top percent = 20) (7).

The remaining, putatively coral, contigs were then "meta-assembled" using CAP3 (≥95% match over ≥50 bp, <20 bp clipping range) (8), followed by several custom python scripts. Our custom scripts join contigs together based on top BLASTN hits to a reference sequence database. This joining was done in three steps: (*i*) contigs were joined if they had "good" top blast hits (BLASTN, ≥100 bp and ≥85% identity) to the same *A. hyacinthus* EST, if they were overlapping or directly adjacent and if they were ≥95% identical in the region where they overlap; (*ii*) contigs were joined identically to step 1 but with the predicted *A. digitifera* transcripts as references; and (*iii*) nonoverlapping/adjacent contigs were joined together if they had "very good" top blast hits (BLASTN, ≥100 bp and ≥90% identity) to the same *A. digitifera* transcript and if none of the contigs with top hits to that transcript are overlapping. In this last case, contigs were joined together by "N"s, with the number of Ns corresponding to the number of reference bases separating the two BLASTN hits.

The final list of putatively coral, metaassembled contigs were aligned with BLASTX against the NCBI nonredundant protein database (nr; downloaded on June, 7, 2011 from www.ncbi.nlm.nih.gov) and against the Swiss-Prot and TrEMBL protein databases from Uniprot (release 2011_03; www.uniprot.org). A custom annotation pipeline constructed in python was used to annotated each sequence with: (*i*) the top nr BLASTX match with an e value of $\leq 1e^{-4}$; (*ii*) when possible, a more informative, lower ranking nr match avoiding a set of keywords (i.e., "predicted," "unknown," "hypothetical"); and (*iii*) functional annotation information for each match to Uniprot with an e value of $\leq 1e^{-4}$ (first using matches to Swiss-Prot and then subsequently TrEMBL if no Swiss-Prot match was found). Uniprot annotation information was extracted from flat files (downloaded from www.uniprot.org) and Uniprot keywords, gene ontology (GO) categories, and Kyoto encyclopedia of genes and genomes (KEGG) pathways were extracted from each flat file for the top BLASTX hit.

One challenge with building a de novo assembly using Illumina, short-read data are that read lengths are substantially shorter than longer sequencing technologies (e.g., Roche 454 GS-Flx), and resulting contigs may, thus, be shorter. Following initial annotation, we developed a decision framework to select the best available annotation for each of our contigs based upon: the above annotation information, similarly constructed annotation tables for the four Cnidarian reference databases used during coral sequence identification (the three larval *Acropora* EST libraries and the *A. digitifera* genome predicted transcripts), and the best BLASTN matches among our contigs and sequences within each of the four Cnidarian references. We ultimately used the annotation information from the longest coral sequence available (i.e., the longest BLASTN match among our contigs and the other *Acropora* references) with a good BLASTX-based

annotation. For example, if one of our contigs had no BLASTX match but matched a longer *A. digitifera* transcript that did have a BLASTX match, we used the *A. digitifera* match as the annotation for that contig. We designed this combined, association-based annotation approach to generate the best available annotation information based upon a comprehensive set of publicly available coral sequence databases for related congeneric species making the assumption that annotation reliability was increased with longer sequence lengths. We used a data cutoff of September 2011 for the purposes of this study; hence, these analyses do not include new coral transcriptome/genome sequences released after the cutoff date.

Read Mapping and Symbiodinium Genotyping. Data from each of the 31 libraries was combined (when multiple lanes were sequences from the same libraries). To account for the different lengths of individual sequencing runs, all longer reads were trimmed to 36 bp, and only the forward sequences were mapped from paired-end lanes. Data were mapped using the Burrow's-Wheeler aligner (BWA) (aln -n 0.005 -k 5 –I) (9) to a combined reference assembly of the above coral sequences as well as a similar de novo assembly that was constructed for *Symbiodinium.* This combined assembly was used as a mapping reference to avoid taxonomical miss-assignment of individual reads (i.e., because multiply mapping reads were excluded from expression analyses, any reads that could not be definitively assigned to either coral or *Symbiodinium* were not counted). Duplicate reads were identified using *Picard* Version 1.43 (MarkDuplicates.jar; http://picard.sourceforge.net) and read counts for each contig were compiled from .sam files for uniquely, well-mapped, nonduplicate reads (≥20 bp; mapping quality, ≥20) using a custom python script.

We estimated the proportion of each clade of *Symbiodinium* at the individual sample level (i.e., mRNA library level) by counting the abundance of clade-specific reads at three loci that are known to be highly divergent between clades: internal transcribed spacer regions 1 and 2 (ITS1, ITS2) and chloroplast 23S rRNA (cp23S). Clade C and D ITS1 and cp23S sequences are from Oliver and Palumbi (10). Clade-specific ITS2 sequences were mined from a preliminary de novo assembly of the data based on nucleotide similarity to the reported sequences from GenBank. This resulted in two ITS2 sequences with best hits to type C3k (100% match) and type D2 (1 bp different). Genotypes will be referred to as clade C and clade D throughout for simplicity.

Each library was mapped to these six clade-specific sequences using BWA (aln -n 0.005 -k 5 –I) (9). For paired-end lanes, only the forward sequences were mapped. Duplicate reads were identified using Picard Version 1.43 (MarkDuplicates.jar) (http://picard.sourceforge.net), and clade proportions at each locus were calculated based on the number of well-mapped, nonduplicate reads (≥25 bp; mapping quality, ≥30) to each of the clade-specific sequences, controlling for slight differences in sequence lengths between clades.

Gene Expression Analyses. Preliminary analyses revealed a strong batch effect for the 11 libraries that were constructed in the laboratory of Dr. John Pringle (Stanford University, Palo Alto, CA) and sequenced separately (by Illumina and the Sidow Lab); hence, these data were only used for de novo assembly and were subsequently excluded from all gene expression analyses. Data were normalized for variation in sequencing depth and differentially expressed genes were identified based on the negative binomial distribution (using the estimateSizeFactors and nbinomTest functions in DESeq, respectively) (11). Low-expression (average normalized expression, <5) contiguous sequences (i.e., "contigs") were excluded from analyses to avoid potential artifact caused by assembly and/or sequencing errors, and high interindividual variability contigs (within-group mean < 1 SD) were also excluded, so that statistical comparisons would not be overly influenced by outlier individuals. The false-discovery rate (FDR) was controlled at 5% according to the method of Benjamini and Hochberg (ref. 12; p.adjust in R). Principal components analysis (PCA) was conducted using the princomp function in R and a correlation matrix. Hierarchical clustering was performed using the pvclust function in R (13). Statistically overrepresented GO categories were determined using default statistical tests and multiple-testing adjustments in GOEAST (14).

SI Results

Transcriptome Sequencing, Assembly, and Annotation. mRNA was isolated from 31 individual coral fragments (16 individual coral colonies) that had been subjected to either 72 h of ambient (termed "Control") or elevated temperature (termed "Heated") exposure. mRNA was converted to cDNA and sequenced on the Illumina GA-II platform in 35 individual lanes. This sequencing effort yielded ∼528 million reads, after quality processing, for a total of 23.9 Gb (Table S2). A total of 220,233 individual contigs were assembled from the data, incorporating 64.71% of the filtered sequences (Table S2). Of these contigs, 41,709 (18.9%) were putatively identified as coral in origin via nucleotide similarity to known Cnidarian sequence resources (larval *Acropora* ESTs and sequenced Cnidarian genomes) and subsequently metaassembled into our final reference transcriptome of 33,496 contigs (N50 = 529 totaling 14.9 Mb; Table S2). Approximately half (49.95%) of these contigs had significant BLASTX matches to known protein sequences in the NCBI nonredundant database (nr), with an almost equivalent number of matches to the Uniprot Swiss-Prot and TrEMBL databases (49.6%; Table S2). Following the first round of initial annotation, our combined annotation pipeline using linked sequences from the other *Acropora* resources (primarily the *Acropora digitifera* predicted transcripts) greatly improved the annotation, with an additional 8,248 sequences with significant BLASTX matches to nr and 7,777 additional matches to Uniprot (Table S2). In total, the 24,980 matches to the nr database are comprised of 12,152 unique NCBI records, whereas the 24,394 Uniprot matches consist of 11,669 unique Uniprot records, all of which had at least one accompanying GO category (7,470 unique GO categories).

Read Mapping and Symbiodinium Genotyping. Alignment of 395.93 million sequences from 31 samples (16 control and 15 heat stressed corals; n = 16 individuals; range: 1.98–22.35 million reads per sample) produced 53.96 million (13.63%) unambiguously aligned coral sequences. Following duplicate removal, an average of 604,881 (5.45% of total) singly aligned, quality (≥20 mapping quality over ≥20 bp), nonduplicate reads per individual were used in subsequent expression analysis (range: 84,289–1.296 million).

A total of 20–704 nonduplicate reads (mean, 348) mapped to our three *Symbiodinium* clade-specific markers per library, and the estimated clade proportions were highly consistent across the three markers (average SD across markers, 2.8%). Three control and four heated samples from the MV pool had an average estimated proportion of clade C of ≥95%, with the remaining two control samples showing mixed assemblages (77% and 81% clade C; Fig. S4). In contrast, all HV samples appeared to be dominated by clade D *Symbiodinium* (10 of 11 samples were ≥95% D; 1 was 94.3%; Fig. S4).

Functional Enrichment Analysis: Control Versus Heated. Functional enrichment analysis across all controls versus heated corals showed 88 biological process (BP), 18 cellular component (CC), and 23 molecular function (MF) categories overrepresented in the up-regulated contigs compared with the entire contig set (Dataset S2*A*). The most significantly enriched of these categories were processes involved in regulation of multicellular organismal process (GO:0051239, GO:0032501), and cell/biological adhesion (GO:0007155, GO:0022610; Dataset S2*A*). There were 13 BP, 2 CC,

and 20 MF categories overrepresented in the down-regulated contigs in the control versus heated comparison, the most significant of which were symporter activity (GO:0015293) and secondary active transmembrane transporter activity (GO:0015291).

1. Beck A, et al. (2010) 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* 5(1):e8768.
2. Meyer E, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10:219.
3. Grasso L, et al. (2008) Microaray analysis identifies candidate genes for key roles in coral development. *BMC Genomics* 9(540):1–18.
4. Shinzato C, et al. (2011) Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* 476(7360):320–323.
5. Putnam NH, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317(5834):86–94.
6. Bayer T, et al. (2012) *Symbiodinium* transcriptomes: Genome insights into the dinoflagellate symbionts of reef-building corals. *PLoS ONE* 7(4):e35269.
7. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21(9):1552–1560.
8. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9(9):868–877.
9. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
10. Oliver TA, Palumbi SR (2011) Do fluctuating temperature environments elevate coral thermal tolerance? *Coral Reefs* 30(2):429–440.
11. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
12. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc, B* 57(1):289–300.
13. Suzuki R, Shimodaira H (2006) Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12):1540–1542.
14. Zheng Q, Wang XJ (2008) GOEAST: A web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res* 36(Web Server issue):W358-63.

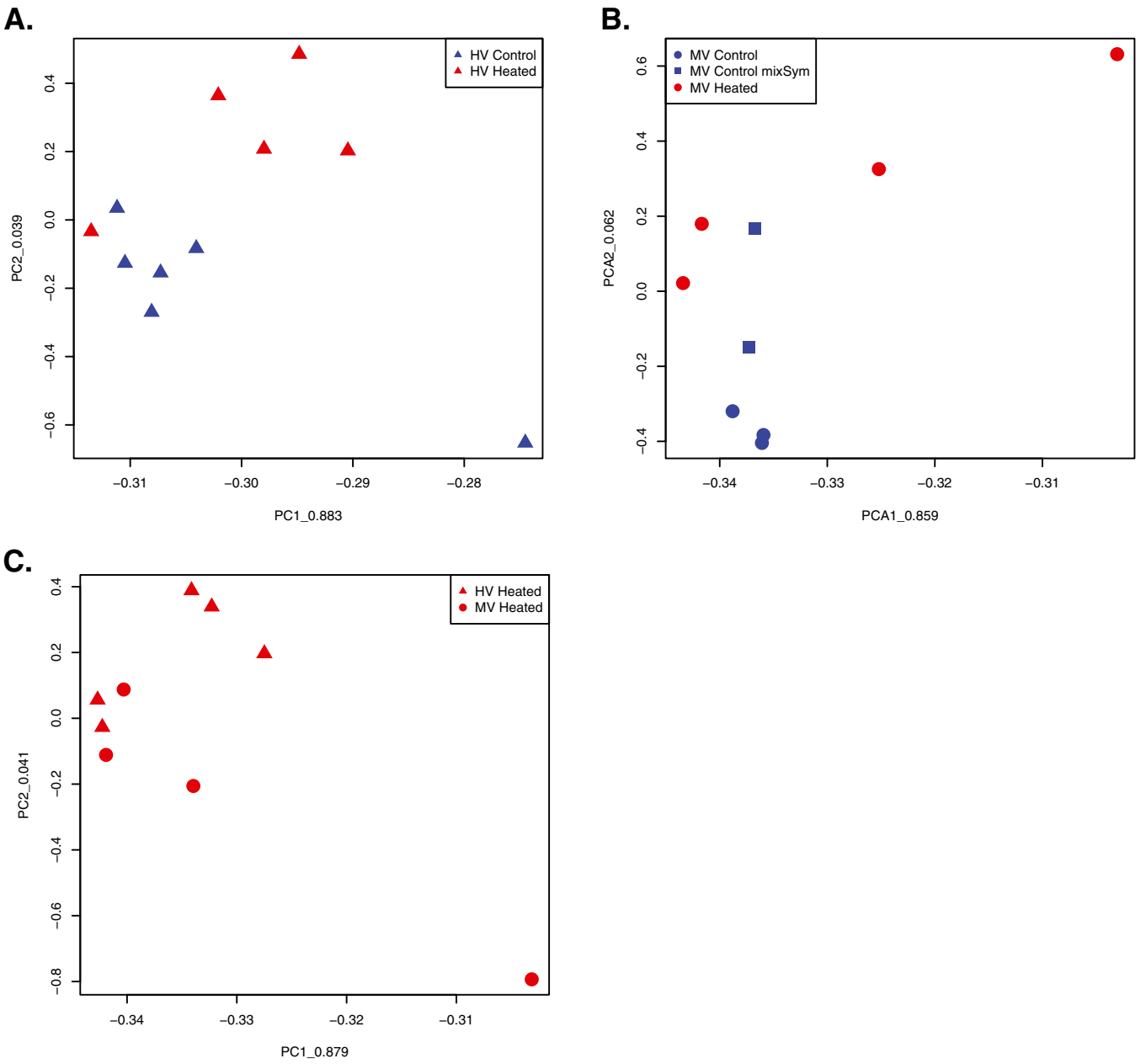

Fig. S1. PCA components 1 and 2 (PC1, PC2) of expression values for all 33,496 contigs in the reference assembly for HV control and HV heateds (*A*), MV control and MV heateds (*B*), and all heated coral samples (*C*). The numbers next to each axis label represent the proportion of variance explained by that principal component (PC). Specific colors reflect treatments and shapes reflect sample populations as shown in each legend. PCA was computed in R using the princomp function and a correlation matrix.

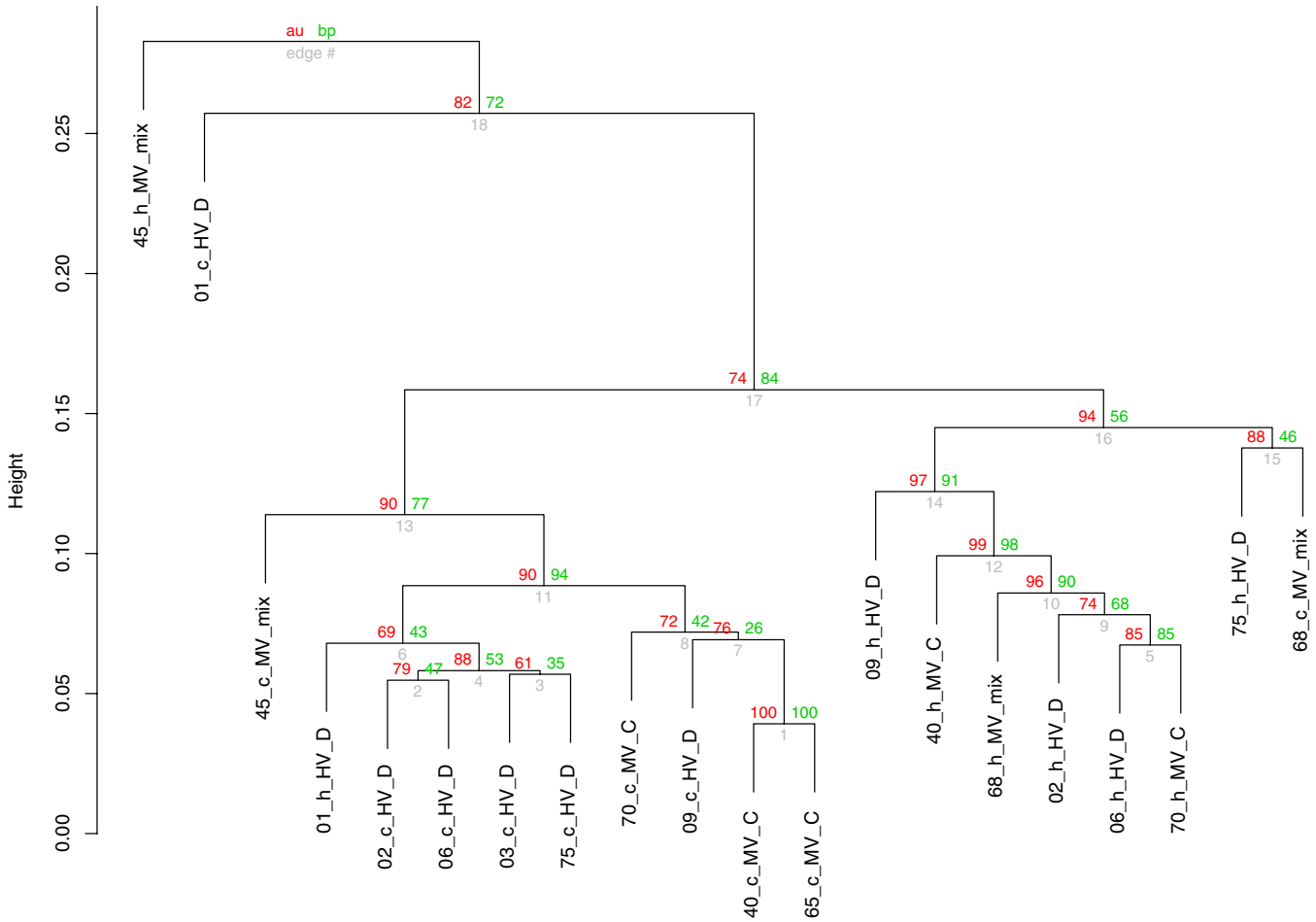

Fig. S2. Hierarchical clustering analysis of expression values for all 33,496 contigs in the reference assembly as calculated by pvclust in R with 10,000 replicates. Numbers in red represent approximately unbiased (AU) *P* values, whereas numbers in green represent bootstrap probability (BP) values. Sample numbers are coded as follows: coral colony number followed by treatment (c for control, h for heated), followed by population (HV or MV), followed by dominant *Symbiodinium* clade (C or D; e.g., "01_c_HV_D" is colony 1, control treatment, HV population, clade D). Distance method: correlation; cluster method: average.

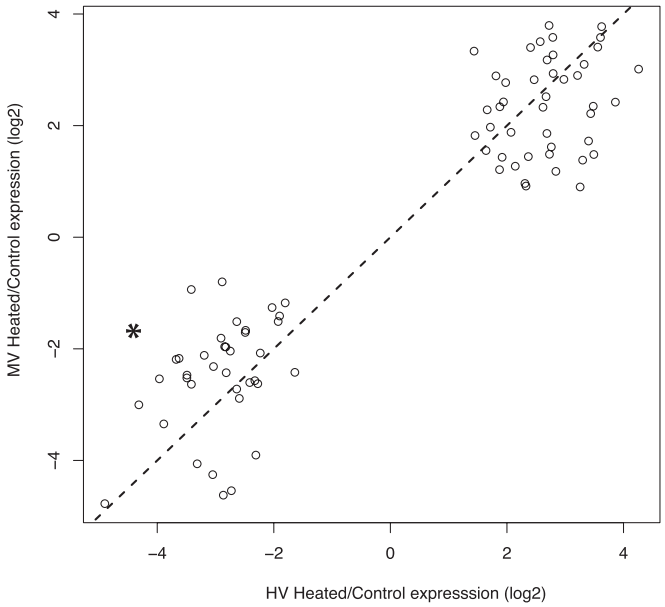

Fig. S3. Scatterplot of the log2 fold changes in gene expression in response to heat stress in the HV corals vs. the MV corals for the 81 differentially expressed genes that were unique to the HV control vs. heated comparison. Each open circle represents an individual contig, the dashed line is a 1:1 line, * denotes a significant departure ($P = 0.013$ for down-regulated contigs) from a 50/50 null expectation of distribution around the 1:1 line (χ^2 test for goodness of fit).

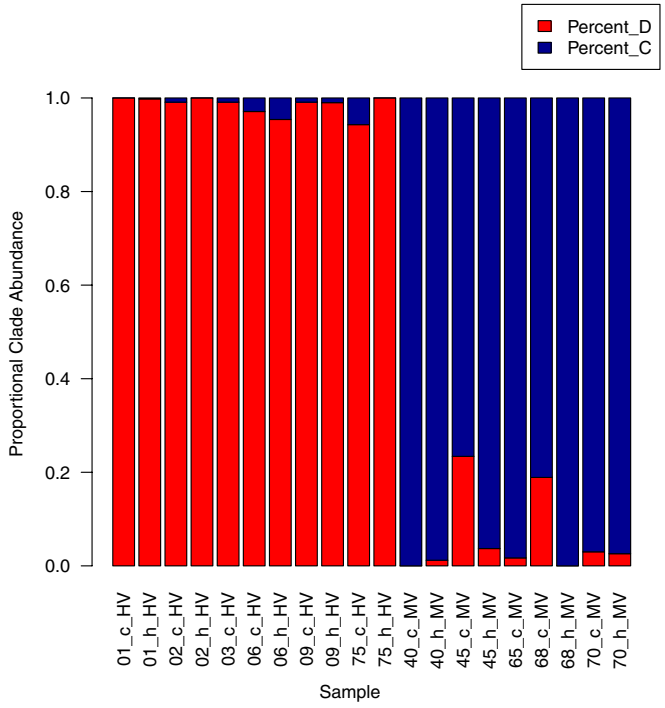

Fig. S4. Average proportional clade abundance of *Symbiodinium* ITS2 types C3k and D2 (referred to as clade C or D for simplicity) as estimated from the number of reads that uniquely mapped to three diagnostic reference loci (ITS1, ITS2, and CP23S). Proportions of clade D and clade C are shown in red and blue, respectively. Sample numbers are coded as in Fig. S2 excepting the clade type.

Table S1. Summary of differentially expressed genes grouped by functional category

	All controls vs. heateds		HV controls vs. heateds		MV controls vs. heateds		
Category name	No. of contigs	Average fold change	No. of contigs	Average fold change	No. of contigs	Average fold change	Total no. of contigs
Membrane_component	109/59	3.56/−4.65	33/29	5.79/−6.56	66/16	4.84/−6.74	312
No_useful_annotation	63/29	4.23/−4.09	28/11	6.88/−8.05	38/13	6.03/−8.05	182
Calcium_ion_homeostasis	42/22	3.03/−4.94	10/12	4.72/−8.87	34/5	4.33/−7.65	125
Transcription_related	39/22	3.59/−4.46	9/8	6.40/−6.38	20/9	4.82/−7.73	107
DNA-replication	32/22	3.64/−4.04	13/4	6.25/−6.22	15/8	6.24/−6.76	94
Miscellaneous	38/15	3.46/−3.37	11/3	6.44/−7.61	26/6	4.69/−5.81	99
Zinc-finger_related	29/11	4.06/−4.90	9/3	6.97/−6.71	16/2	5.63/−8.85	70
Sensory_systems	24/15	3.38/−4.73	5/10	5.95/−7.97	15/7	5.03/−8.05	76
CellAdhesion	25/12	3.10/−4.43	5/11	5.13/−8.84	14/4	4.69/−6.14	71
Cell_differentiation	20/9	3.29/−5.69	4/8	5.13/−5.98	8/5	3.43/−7.69	54
Apoptosis	16/8	4.27/−3.80	5/6	6.53/−6.17	8/2	5.90/−5.65	45
Actin-Cytoskeleton	16/7	3.39/−3.38	3/4	7.25/−11.35	10/1	5.58/−6.46	41
InnateImmunity	14/8	3.02/−7.53	2/2	3.81/−9.18	7/3	4.00/−10.24	36
Lipid_metabolism	15/4	3.48/−4.47	4/1	5.77/−6.03	9/1	4.52/−3.52	34
G protein-coupled	16/2	4.10/−3.23	6/1	7.90/−3.27	4/1	7.16/−4.43	30
RNA-binding	15/2	3.78/−4.13	9/0	6.77/0.00	7/0	5.64/0.00	33
Oxidative_Stress	13/4	3.92/−9.13	3/1	8.33/−29.85	13/2	4.48/−10.06	36
GTP-binding	10/6	3.58/−4.86	3/1	5.40/−5.75	8/1	4.76/−6.46	29
Ubiquitin	7/7	5.78/−3.63	2/2	11.48/−4.86	4/1	9.12/−4.84	23
ECM-Cell_structure	6/8	2.49/−4.56	4/5	5.45/−5.57	6/4	4.35/−4.68	33
Cell_Cycle	8/4	3.81/−4.29	1/1	6.81/−7.39	6/1	5.61/−8.87	21
MAP_kinase-activity	9/0	2.73/0.00	2/0	4.48/0.00	4/0	4.19/0.00	15
Transposon	7/1	3.99/−2.95	6/0	8.32/0.00	3/1	5.97/−5.60	18
Heat-Shock	6/2	3.05/−4.37	3/1	5.71/−5.75	8/1	7.02/−6.46	21
Calcification	1/4	2.52/−4.04	0/3	0.00/−12.59	0/3	0.00/−6.20	11
Protease_inhibition	3/1	3.97/−4.22	3/0	4.85/0.00	1/0	5.65/0.00	8
Ribosomal_proteins	3/0	3.08/0.00	1/0	11.24/0.00	1/0	3.62/0.00	5
Cysteine-rich_peptide_activity	2/1	2.69/−5.38	0/1	0.00/−4.15	2/1	2.92/−7.25	7
Total no. of unique contigs	484		159		247		574*

Categories were chosen using GO classifications, additional information from BLAST matches to the NCBI nr database and literature. Numbers and fold changes are for up-regulated and down-regulated genes, respectively. The specific genes in each category can be found in Dataset S1*F*. Note: categories are not mutually exclusive; hence, a single contig may be represented in more than one category.
*Comprising 404 unique Uniprot matches.

Table S2. Summary of sequencing, de novo assembly, and annotation

	Sequences	Bases (Mb)	N50	Maximum length (bp)
No. of reads after quality filtering	527,721,439	23,912	—	—
No. of reads assembled (64.71%)	341,470,944	16,912	—	—
Total contigs	220,233	96.3	500	8,757
Total coral contigs	33,496	14.9	529	9,382
Total contigs with BLAST matches				
to NCBI nr	16,732	9.76	753	9,382
to Uniprot	16,617	9.72	757	9,382
Total contigs with BLAST matches after coral genome/EST association				
to NCBI nr	24,980 (12,152 unique matches)*	12.4	621	9,382
to Uniprot	24,394 (11,669 unique matches)*	12.1	623	9,382

—, not applicable.
*Numbers in parenthesis represent unique NR and Uniprot matches.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)