Supplementary Methods

Sequence Data and Computational Tools:

Transcriptomic data from 17 species of Scleractinia (stony corals) and 3 species of Actiniaria (anemones) were downloaded from the web (supplemental table S4, Supplementary Material online; Schwarz et al. 2008; Sunagawa et al. 2009; Polato et al. 2011; Shinzato et al. 2011; Moya et al. 2012; Kenkel et al. 2013; Lubinski and Granger 2013; Sun et al. 2013; Maor-Landaw et al. 2014; Nordberg et al. 2014; Willette et al. 2014; Kitchen et al. 2015; Davies et al. Forthcoming). Instructions, scripts, and example output files for computational methods used in this study are available on GitHub (https://github.com/grovesdixon/metaTranscriptomes). Gene Ontology and KOG annotations were applied as described in (Dixon et al. 2015). Instructions and scripts for the gene annotation pipeline are available on GitHub (https://github.com/z0on/annotatingTranscriptomes). Significance for enrichment of KOG terms across MBD-scores was tested using Mann-Whitney U tests implemented in the R package KOGMWU as in Dixon et al. (2015).

Ortholog Identification and alignment

Orthologs were identified based on reciprocal best Blast hits between extracted protein sequences. First, coding and amino acid sequences were extracted from each transcriptome based on alignments (e-value cutoff = 1e-5) to a reference proteome using BlastX (Altschul et al. 1997) and a custom Perl script CDS_extractor_v2.pl (https://github.com/z0on/annotatingTranscriptomes), which identifies and corrects frame shift

mutations within the BlastX-aligned sequences. The reference proteome was a concatenation of the *Nematostella vectensis* (Nordberg et al. 2014) and *Acropora digitifera* (Shinzato et al. 2011) reference proteomes. The protein sequences for all pairs of species were reciprocally blasted using BlastP (Altschul et al. 1990). Because our MBD-seq dataset was generated from *A. millepora*, we used its sequences as anchors for orthologous groups. First an initial set of candidate orthologs was compiled based on reciprocal best hits between *A. millepora* and each other species. Only hits with alignment lengths >75% of the subject sequence and an e-value < 1e-5 were retained. This initial set was then refined to include only sequences that were reciprocal best hits with >= 50% of other candidate orthologs within the group (supplementary fig. S10, Supplementary Material online). Orthologous groups with fewer than three (15%) representative species were excluded. For building the species tree, a separate, highly conserved set of orthologs was assembled with percent amino acid identity > 75%. These were further filtered by retaining only orthologs with representative sequences from > 80% of species. As a final filter, we used cluster analysis of dS values to identify likely paralogs and spurious orthologs. For each species a three component Gaussian mixture model was fit to the pairwise dS estimates with *A. millepora*. The first two components were assumed to capture the true orthologs, the third component was assumed to have captured false positive orthologs (supplemental fig. S11, Supplemental Material online). Mean dS for the third component was on average 60 times higher than the second highest component. On average 10% of ortholog calls were flagged as false positives and removed. Amino acid sequences for each ortholog were aligned with MAFFT (Katoh and Standley 2013) using the 'localpair' algorithm. The protein

alignments were then reverse translated into codon sequences using Pal2Nal (Suyama et al. 2006).

Substitution rate analyses

To estimate substitution rates (dS and dN) we used codeml in the software package PAML (Yang 2007). Substitution rates were estimated using pair-wise comparisons between *A. millepora* and each other species that had representative sequences for each ortholog. Example codeml control files for the pair-wise comparisons and the branch-sites models are available on GitHub (https://github.com/grovesdixon/metaTranscriptomes).

Building species tree

Based on a highly conserved set of ortholog sequences we constructed species tree using RAxML (Stamatakis 2014). For phylogenetic construction we ran RAxML rapid bootstrapping algorithm using the GTRGAMMA model and 1000 iterations. We decided to use putative orthologs with representative sequences in > 80% of taxa through iterations of tree building. The best trees from ortholog sets using 40%, 50% and 60% cutoffs all gave the same topology. The best tree using the 80% cutoff was chosen for because it had highest bipartition bootstrap values.

Library preparation for MBD-seq

To quantify gbM in *Acropora millepora* we used methyl-CpG binding domain protein-enriched sequencing (MBD-seq). Enrichment reactions were performed using the MethylCap kit (Diagenode Cat. No. C02020010). Seven enrichment reactions were performed. Input DNA for all reactions was isolated from a single colony of *A. millepora* sampled from the Central Great

Barrier Reef (Great Barrier Reef Marine Park Permit G09/29894.1). DNA was diluted to 0.1 µg/µl then sheared with a Misonix Sonicator 3000 for nine or ten minutes using 15 second cycles at ~30W. Sheared fragments ranged from ~100 to 800 bp spanning the range 300 – 500 bp recommended by the manufacturer. The manufacturer's protocol recommended an input of 12 µl of sheared DNA diluted in 130 µl of buffer. We found that our yields were higher when we used 18, 24 or 48 µl of sheared DNA (1.5x, 2x and 4x concentrated). As the kit is intended for mammalian DNA, lower genome wide methylation levels in our system could explain why higher input concentrations worked better in our case. Flow-through from initial capture of methylated DNA was retained for sequencing. The methylated fraction was eluted from the capture beads in a single step using High Elution Buffer. Electrophoresis gels were used to assess the size and quality of each elution. For sequencing, product from enrichment replicates 1-4 and 5-7 were pooled. Final concentrations of these pooled libraries were 6 and 4 ng/µl measured with a Nanodrop Spectrophotometer. Similarly, the flow-through components from the same replicates were pooled. Concentrations of the flow-through pools were 34 and 36 ng/µl. Adapter ligation using a NEBnext kit (New England Biolabs®), library quality assessment using a Bioanalyzer (Agilent Technologies), and sequencing on a HiSeq 2500 platform (Illumina®) were performed by the University of Texas Genome Sequencing and Analysis Facility.

Analysis of gene body methylation

Raw reads from the MBD-sequencing libraries were trimmed using cutadapt (Martin 2011) and quality filtered using Fastx toolkit

(http://cancan.cshl.edu/labmembers/gordon/fastx_toolkit/). Reads were then aligned to coding sequences extracted from the *A. millepora* reference transcriptome (Moya et al. 2012), as described above. DESeq2 (Love et al. 2014) was used to calculate the $\log_2$ fold difference between the MBD-enriched and flow-through libraries. We used this $\log_2$ fold difference, which we refer to as MBD-score, as our quantification of the strength of gbM for each gene. Negative values indicate weak methylation and positive values indicate strong methylation. To examine the distribution of MBD-scores we used the R package Mclust (Fraley and Raftery 2007). We first assessed the optimal mixture model and number of components based on Bayesian Information Criterion (BIC). The optimal number of components was greater than one with little change in BIC beyond two components (supplemental fig. S1A, Supplemental Material online). Based on this result we fitted a two-component mixture model to the MBD-scores (supplemental fig. S1B, Supplemental Material online).

Because of the hypermutability of 5mC, genes that are strongly methylated in the germline become deficient in CpG dinucleotides over evolutionary time (Sved and Bird 1990). As a result, normalized CpG content (CpGo/e) can be used to estimate historical germline methylation. This metric has been shown to correlate closely with direct measures of gbM (Zemach et al. 2010; Sarda et al. 2012). To corroborate that our measure of gbM also correlated with CpGo/e we calculated it for the *A. millepora* coding regions as described in Dixon et al. (2014). To control for effects on gene length, CpGo/e was calculated based on the first 1000 bases of each sequence.

Gene expression datasets

To test for correlations between MBD-score and transcriptional variation we used gene expression data from two previous experiments. Both datasets were generated using Tag-based RNA-seq (Meyer et al. 2011) from samples of *A. millepora* taken from the central Great Barrier Reef, Australia. The current laboratory and bioinformatics protocols for analysis of Tag-based RNA-seq are available on GitHub (https://github.com/z0on/tag-based_RNAseq). The first dataset was a subset of that described in (Dixon et al. 2015), including 12 adult samples (triplicate samples from 2 genotypes from Princess Charlotte Bay and 2 from Orpheus Island: Great Barrier Reef Marine Park Authority permit G38062.1 exposed to 28°C) and 30 samples of their larval offspring (10 genetic families, reared for five days at 28°C in triplicate). Variation in gene expression between adults and larvae was analyzed using DESeq2 (Love et al. 2014). Comparisons between MBD-score and transcript abundance were based on counts from adult samples transformed to a $\log_2$ scale using the rlogTransformation function. Mean expression levels from this dataset were also used to calculate indices of codon bias described below. The second dataset described in (Dixon et al. 2014) included 56 colony fragments reciprocally transplanted between two environmentally distinct reefs: Keppel and Orpheus Island (Keppel: 23°09S 150°54E and Orpheus 18°37S 146°29E: Great Barrier Reef Marine Park Authority permit G09/29894.1). Expression profiles from these samples were analyzed with respect to the transplantation site to examine variation in gene expression due to environmental conditions.

Codon bias

We tested for relationships between MBD-score and synonymous codon usage using four metrics: relative synonymous codon usage (RSCU), frequency of optimal codons (Fop), codon

adaptation index (CAI), and the effective number of codons (Nc). RSCU was calculated as the ratio of the observed number of occurrences of a particular codon to the excepted number of occurrences if codon usage was neutral (Sharp et al. 1986):

$$RSCU_{ij} = \frac{x_{ij}}{1/n_i \sum_{j=1}^{n_i} x_{ij}}$$

Where X_{ij} is the number of occurrences of the *j*th codon for the *i*th amino acid and n_i is the number of synonymous codons for the *i*th amino acid. This measure quantifies relative codon usage while controlling for variation in amino acid composition between proteins. Fop is intended to measure the degree of selection for optimal codon usage in a particular coding sequence. It was originally defined as the ratio of optimal codons to the total number of codons in a gene, with optimal codons identified based on the cellular content of isoaccepting tRNAs and the nature of codon-anticodon interactions (Ikemura 1981). Optimal codons are also inferred based on relative usage in a set of highly expressed genes such as ribosomal proteins (Behura and Severson 2013). To estimate Fop for *A. millepora* coding regions we used the software package CodonW (Peden 1999)(http://codonw.sourceforge.net/). CodonW uses correspondence analysis of codon usage to derive a set of optimal codons and then estimates their usage for each sequence. CAI is similar to Fop, and is intended to quantify the strength of selection on codon usage. For a given gene, CAI is equal to the geometric mean of the relative adaptiveness (*W*) of all codons within that gene. The relative adaptiveness W_{ij} of codon *i* that codes for amino acid *j* is equal to the ratio its relative synonymous codon usage to that of the most abundant synonymous codon in a set of highly expressed genes (Sharp and Li 1987a):

$$W_{ij} = RSCU_{ij}/RSCU_{imax}$$

Relative adaptiveness (based on the top 5% most highly expressed genes) and CAI were calculated using custom python scripts. Unlike CAI and Fop, Nc does not depend on a set of preferred codons, and provides an estimate of a gene's departure from random use of synonymous codons based solely on codon usage. The measure is analogous to the 'effective number of alleles' in population genetics (treating amino acids as loci and codons as alleles) summed the values across the 20 amino acids. It and is bounded between 20 (completely biased) to 61 (neutral)(Wright 1990). Nc was calculated using CodonW (Peden 1999)(http://codonw.sourceforge.net//culong.html).

Statistical Analyses

Statistical analyses of the relationship between MBD-score and other gene characteristics were performed in R (R Core Team 2015). Significance for correlations was established using Spearman's rank-order correlation test. Significance tests for differences in counts between the strongly methylated and weakly methylated classes were performed using Fisher's exact tests (Fisher 1922). Principal component analysis was performed using prcomp function in R.

References:

Altschul S, Gish W, Miller W. 1990. Basic Local Alignment Search Tool. J Mol Biol. 215:403–410.

Altschul SF, Madden TL, Schäffer a a, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Behura SK, Severson DW. 2013. Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. Biol. Rev. 88:49–61.

Davies SW, Marchetti A, Ries JB, Castillo KD. Thermal and pCO2 stress elicit divergent transcriptomic responses in a resilient coral. Scientific Reports. Front. Mar. Sci.

Dixon GB, Bay LK, Matz M V. 2014. Bimodal signatures of germline methylation are linked with gene expression plasticity in the coral Acropora millepora. BMC Genomics 15:1109.

Dixon GB, Davies SW, Aglyamova G V, Meyer E, Bay LK, Matz M V. 2015. Genomic determinants of coral heat tolerance across latitudes. Science 348:1460–1462.

Fisher RA. 1922. On the Interpretation of χ2 from Contingency Tables , and the Calculation of P. J. R. Stat. Soc. 85:87–94.

Fraley C, Raftery AE. 2007. Model-based Methods of Classification : Using the mclust Software in Chemometrics. J. Stat. Softw. 18:1–13.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151:389–409.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kenkel CD, Meyer E, Matz M V. 2013. Gene expression under chronic heat stress in populations of the mustard hill coral (Porites astreoides) from different thermal environments. Mol. Ecol. 22:4322–4334.

Kitchen SA, Crowder CM, Poole AZ, Weis VM, Meyer E. 2015. De novo assembly and characterization of four anthozoan (phylum Cnidaria) transcriptomes. G3 Genes Genomes Genet. 5:2441–2452.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 15:1–21.

Lubinski T, Granger B. 2013. PocilloporaBase. Available from: http://cnidarians.bu.edu/PocilloporaBase/cgi-bin/index.cgi

Maor-Landaw K, Karako-Lampert S, Ben-Asher HW, Goffredo S, Falini G, Dubinsky Z, Levy O. 2014. Gene expression profiles during short-term heat stress in the red sea coral Stylophora pistillata. Glob. Chang. Biol. 20:3026–3035.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:10–12.

Meyer E, Aglyamova G V, Matz M V. 2011. Profiling gene expression responses of coral larvae

(Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Mol. Ecol20:3599–3616.

Moya A, Huisman L, Ball EE, Hayward DC, Grasso LC, Chua CM, Woo HN, Gattuso J-P, Forêt S, Miller DJ. 2012. Whole transcriptome analysis of the coral Acropora millepora reveals complex responses to CO_2-driven acidification during the initiation of calcification. Mol. Ecol. 21:2440–2454.

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev I V., Dubchak I. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42:26–31.

Peden JF. 1999. Analysis of codon usage. Thesis Submitt. to Univ. Nottingham:50–90.

Polato NR, Vera JC, Baums IB. 2011. Gene discovery in the threatened elkhorn coral: 454 sequencing of the acropora palmata transcriptome. PLoS One 6:e28634.

R Core Team. 2015. R: a language and environment for statistical computing. Available from: http://www.r-project.org/

Sarda S, Zeng J, Hunt BG, Yi S V. 2012. The evolution of invertebrate gene body methylation. Mol. Biol. Evol. 29:1907–1916.

Schwarz J a, Brokstein PB, Voolstra C, Terry AY, Manohar CF, Miller DJ, Szmant AM, Coffroth MA, Medina M. 2008. Coral life history and symbiosis: functional genomic resources for two reef building Caribbean corals, Acropora palmata and Montastraea faveolata. BMC Genomics 9:97.

Sharp PM, Li WH. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.

Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, Fujie M, Fujiwara M, Koyanagi R, Ikuta T, et al. 2011. Using the Acropora digitifera genome to understand coral responses to environmental change. Nature 476:320–323.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Sun J, Chen Q, Lun JCY, Xu J, Qiu JW. 2013. PcarnBase: Development of a Transcriptomic Database for the Brain Coral Platygyra carnosus. Mar. Biotechnol. 15:244–251.

Sunagawa S, Wilson EC, Thaler M, Smith ML, Caruso C, Pringle JR, Weis VM, Medina M, Schwarz J a. 2009. Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone Aiptasia pallida and its dinoflagellate endosymbiont. BMC Genomics 10:258.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:609–612.

Sved J, Bird a. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. U. S. A. 87:4692–4696.

Willette D a., Allendorf FW, Barber PH, Barshis DJ, Carpenter KE, Crandall ED, Cresko W a., Fernandez-Silva I, Matz M V., Meyer E, et al. 2014. So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. Bull. Mar. Sci. 90:79–122.

Wright F. 1990. The “effective number of codons” used in a gene. Gene 87:23–29.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916–919.