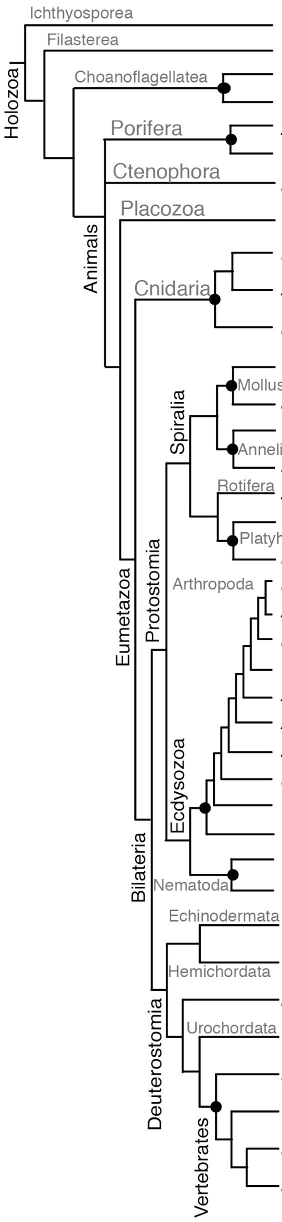
Convergent evolution of a vertebrate-like methylome in a marine sponge

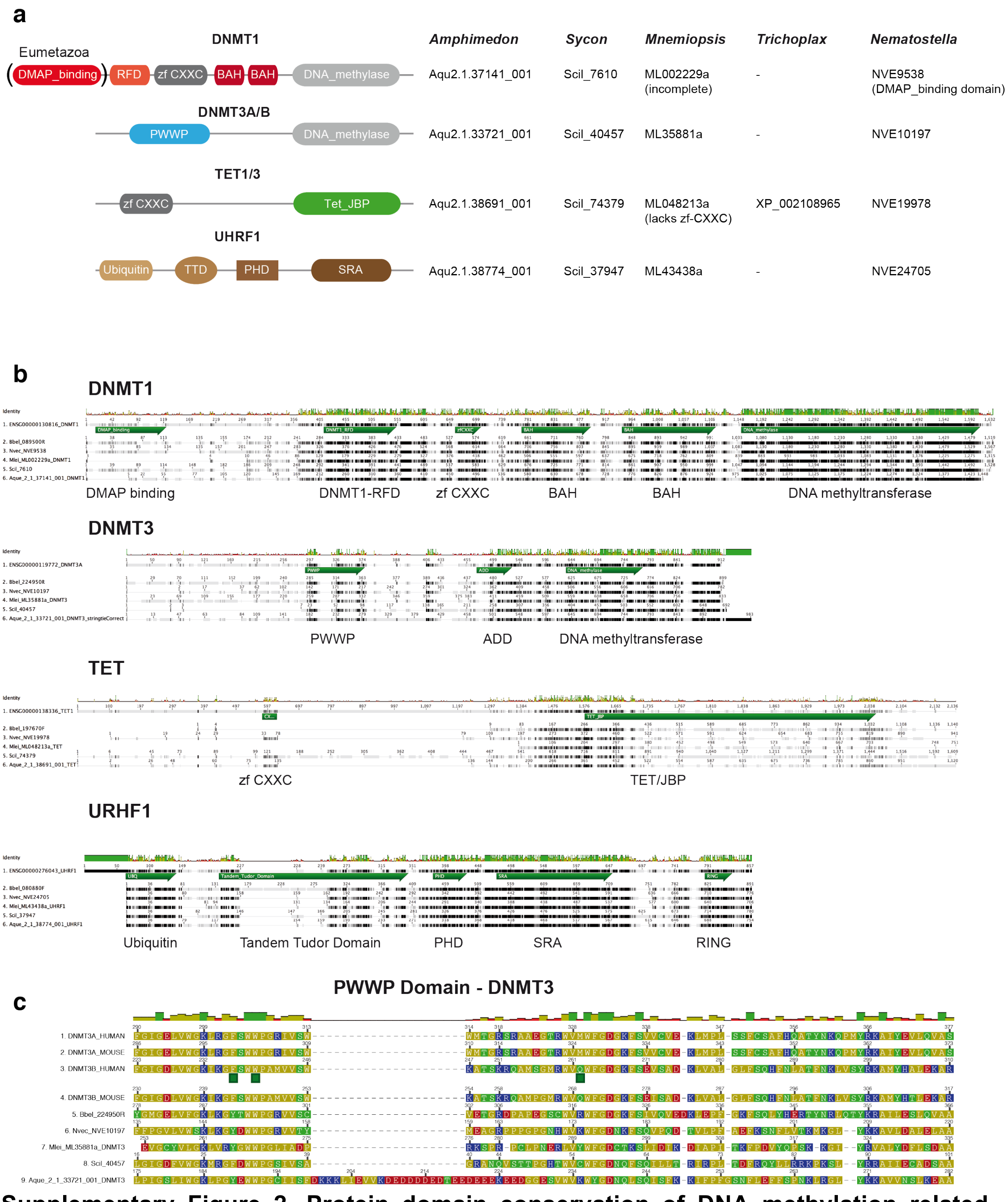
Alex de Mendoza, William Hatleberg, Kevin Pang, Sven Leininger, Ozren Bogdanovic, Jahnvi Pflueger, Sam Buckberry, Ulrich Technau, Andreas Hejnol, Maja Adamska, Bernard M Degnan, Sandie M Degnan, Ryan Lister

Supplementary Material.



						MBD-containing proteins						zf-CXXC proteins			
	DNMT1	DNMT3(A/B/L)	UHRF1	SRA unknown	TET (1/2/3)	MBD4/MeCP2	MBD1/2/3	BAZ2A/B	SETDB1/2	MBD-Fbox	MBD-R2	KDM2A/B	KMT2A/B	CXXC1	CXXC4/5
Creolimax fragrantissima	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
Caspaspora owczarzaki	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
Salpingoeca rosetta	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Monosiga brevicollis	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Amphimedon queenslandica	1	1	1	0	1	0	1	1	0	0	0	**1**	**1**	**2**	0
Sycon ciliatum	1	1	1	0	1	0	1	1	1	0	0	**4**	1	**2**	0
Mnemiopsis leidyi	1	1	1	0	1	0	1	1	0	1	0	**2**	1	2	0
Trichoplax adhaerens	0	0	0	0	1	0	0	1	0	0	0	1	1	**1**	0
Nematostella vectensis	1	1	1	1	1	1	1	1	1	0	0	**3**	**4**	**2**	0
Acropora digitifera	1	1	0?	0	1	0	1	0	0	0	0	**3**	1	**8**	0
Hydra magnipapillata	1	1	0?	2	1	0	1	1	0	0	0	**1**	1	**2**	0
Crassostrea gigas	1	1	1	1	1	1	1	1	1	0	0	**1**	**1**	**1**	0
Lottia gigantea	1	1	1	1	1	1	1	0	1	1	0	**2**	1	**1**	0
Capitella teleta	2	1	1	1	2	1	1	2	1	1	0	**1**	2	**2**	0
Helobdella robusta	1	1	2	0	1	0	1	0	0	1	0	**4**	2	**1**	0
Adineta vaga	0	0	0	0	0	0	0	4	2	0	0	**5**	0	2	0
Schistosoma mansoni	0	0	0	0	0	0	1	0	1	0	0	0	0	**2**	0
Echinoccocus multilocularis	0	0	0	0	0	0	1	1	1	0	0	0	1	**2**	0
Drosophila melanogaster	0	0	0	0	1	0	1	1	1	0	1	**1**	2	**3**	0
Aedes aegypti	0	0	0	0	1	0	1	1	1	0	1	**1**	1	**2**	0
Bombyx mori	1	0	1	0	1	0	1	1	1	1	1	**2**	1	**1**	0
Tribolium castaneum	1	0	1	0	1	0	1	1	1	1	1	**1**	1	**1**	0
Apis mellifera	2	1	0?	1	1	0	1	1	1	1	1	**1**	1	**1**	0
Acyrthosiphon pisum	3	1	2	0	1	0	0	0	1	9	0	**1**	1	**1**	0
Zootermopsis nevadensis	0?	2	0?	0	1	1	1	1	1	1	0	**1**	1	**1**	0
Daphnia pulex	1	1	1	4	1	1	1	1	2	2	1	1	1	**1**	0
Strigamia maritima	4	1	1	0	1	1	1	1	2	1	0	**1**	**1**	**1**	0
Tetranychus urticae	1	1	1	0	1	0	1	1	3	0	0	**2**	0	5	0
Caenorhabditis elegans	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0
Trichinella spiralis	0	1	0	0	0	0	0	1	0	0	0	0	0	**1**	0
Strongylocentrotus purpuratus	1	1	1	1	1	1	1	1	0	0	0	**2**	**1**	**4**	0
Saccoglossus kowalevskii	1	0	1	1	1	1	1	1	1	0	0	**1**	**1**	**1**	0
Branchiostoma belcheri	1	1	1	1	1	1	1	1	1	0	0	**1**	**2**	**1**	0
Ciona intestinalis	1	2	0?	0	1	0	1	0	1	0	0	**3**	1	**1**	0
Petromyzon marinus	1	2	1	0	0?	1	1	1	1	0	0	**1**	1	0	0
Callorhinchus milii	0?	2	2	0	3	2	0	2	0	0	0	**2**	**1**	0	2
Danio rerio	1	6	3	0	3	1	**4**	3	2	0	0	**5**	**3**	**2**	4
Homo sapiens	1	3	2	0	3	2	**3**	2	2	0	0	**3**	**2**	**1**	2

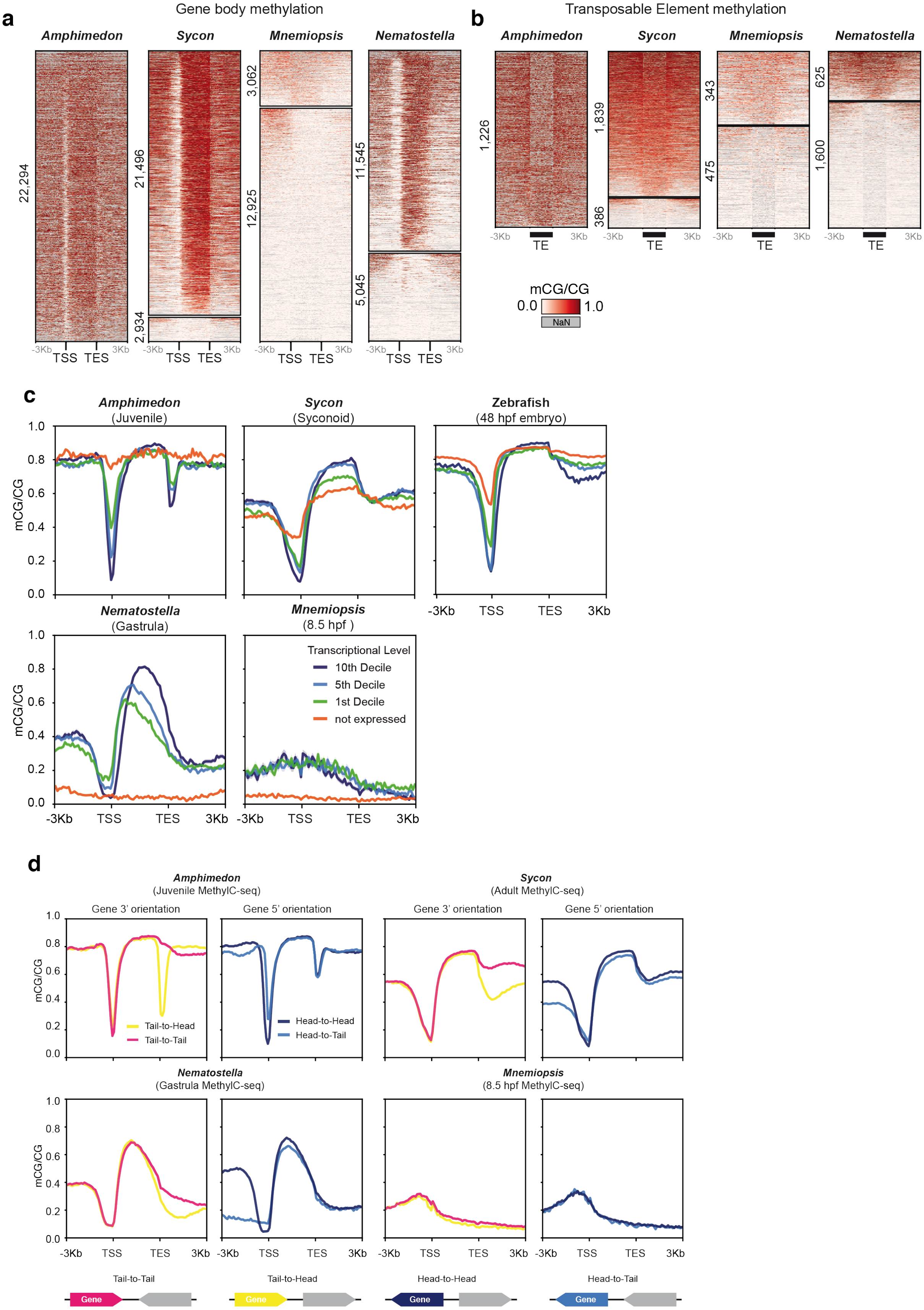
Supplementary Figure 1. Distribution of DNA methylation related genes in metazoan genomes. The indicated gene number was determined using phylogenetic reconstruction (Maximum likelihood, see Methods section) of each gene family (DNMT, SRA, TET, MBD, zinc finger CXXC) and mapped on the current animal phylogeny [77,78]. A question mark is indicated on gene absences that might be due to incomplete genome annotations or assemblies, based on its incongruence with the presence of other genes of the pathway (e.g. lacking UHRF1 when encoding DNMT1). Grey shading indicates separate lineages. Bold red numbers indicate the presence of a zinc finger CXXC in any of the given orthologues of a gene family found in a given species. Presence or absence of genes has been determined by phylogenetic reconstruction of each family, see Gene annotation section in Methods.



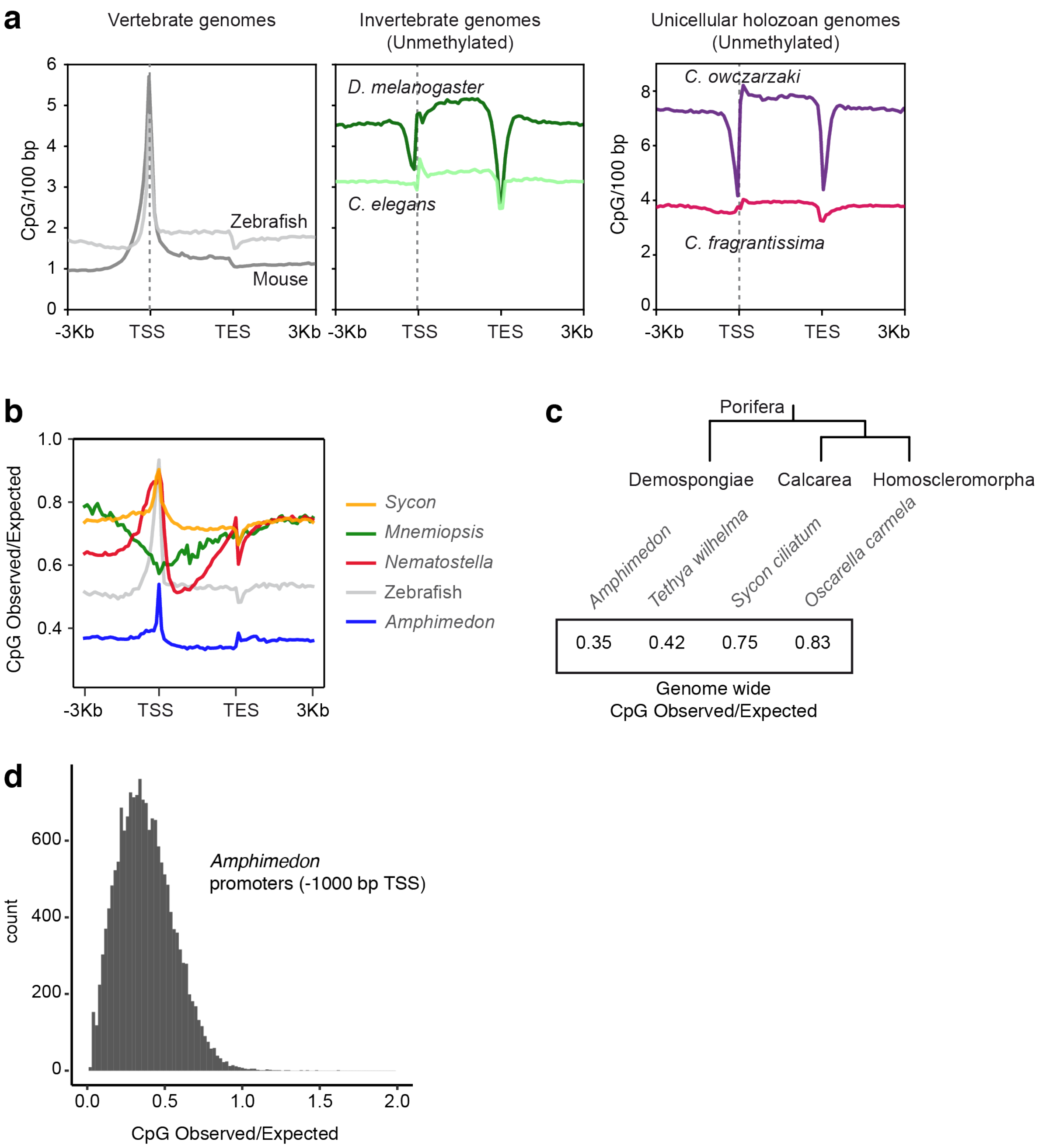
	Amphimedon	*Sycon*	*Mnemiopsis*	*Trichoplax*	*Nematostella*
DNMT1	Aqu2.1.37141_001	Scil_7610	ML002229a (incomplete)	-	NVE9538 (DMAP_binding domain)
DNMT3A/B	Aqu2.1.33721_001	Scil_40457	ML35881a	-	NVE10197
TET1/3	Aqu2.1.38691_001	Scil_74379	ML048213a (lacks zf-CXXC)	XP_002108965	NVE19978
UHRF1	Aqu2.1.38774_001	Scil_37947	ML43438a	-	NVE24705

Supplementary Figure 2. Protein domain conservation of DNA methylation related genes. (**a**) Protein Domain architecture of DNMT1, DNMT3, TET and UHRF1 in metazoans and the corresponding gene IDs in *Amphimedon*, *Sycon, Mnemiopsis, Trichoplax* and *Nematostella*. Protein domains are defined as in the PFAM database (PF06464 DMAP_binding, PF12047 RFD, PF02008 zf-CXXC, PF01426 BAH, PF00145 DNA_methylase, PF00855 PWWP, PF12851 Tet_JBP, PF00240 Ubiquitin, PF12148 TTD, PF00628 PHD, PF02182 SRA). The *Mnemiopsis* DNMT1 gene model is incomplete in the genome annotation, but can be fully recovered using a transcriptome assembly. In the case of the *Mnemiopsis* TET ortholog, both the genome annotation and the transcriptome assembly lack a zf-CXXC domain. DMAP_binding domain, responsible for binding to the transcriptional co-repressor DMAP1, is only found in *Nematostella* DNMT1 and in some bilaterians (including vertebrates), but is absent in sponges, placozoans and ctenophores.

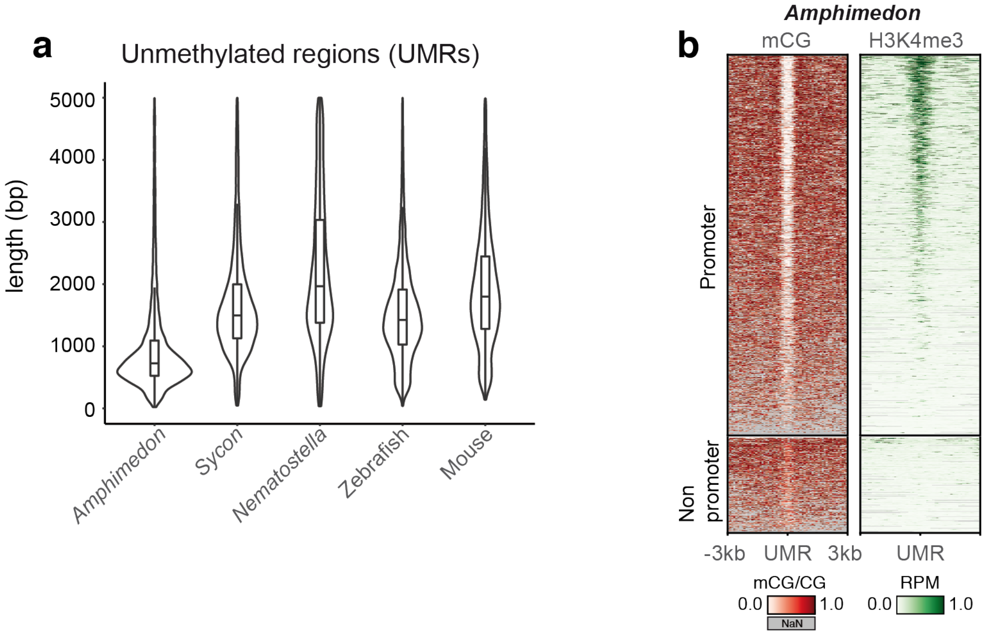
(**b**) Amino acid multi-sequence alignment of DNMT1, DNMT3, TET and UHRF1 orthologs in 5 species. Top track shows level of identity in each given aligned position, and color of the amino acids is determined by similarity, where black is identical and white is more dissimilar as computed by Geneious software. Domains are highlighted with green arrows on the human orthologues. (**c**) Alignment focusing on the PWWP domain from DNMT3 orthologues. Amino acids coloured by polarity (using Geneious). Green squares indicate amino acid positions that diminish DNMT3B preference for H3K26me3 when mutated [44].



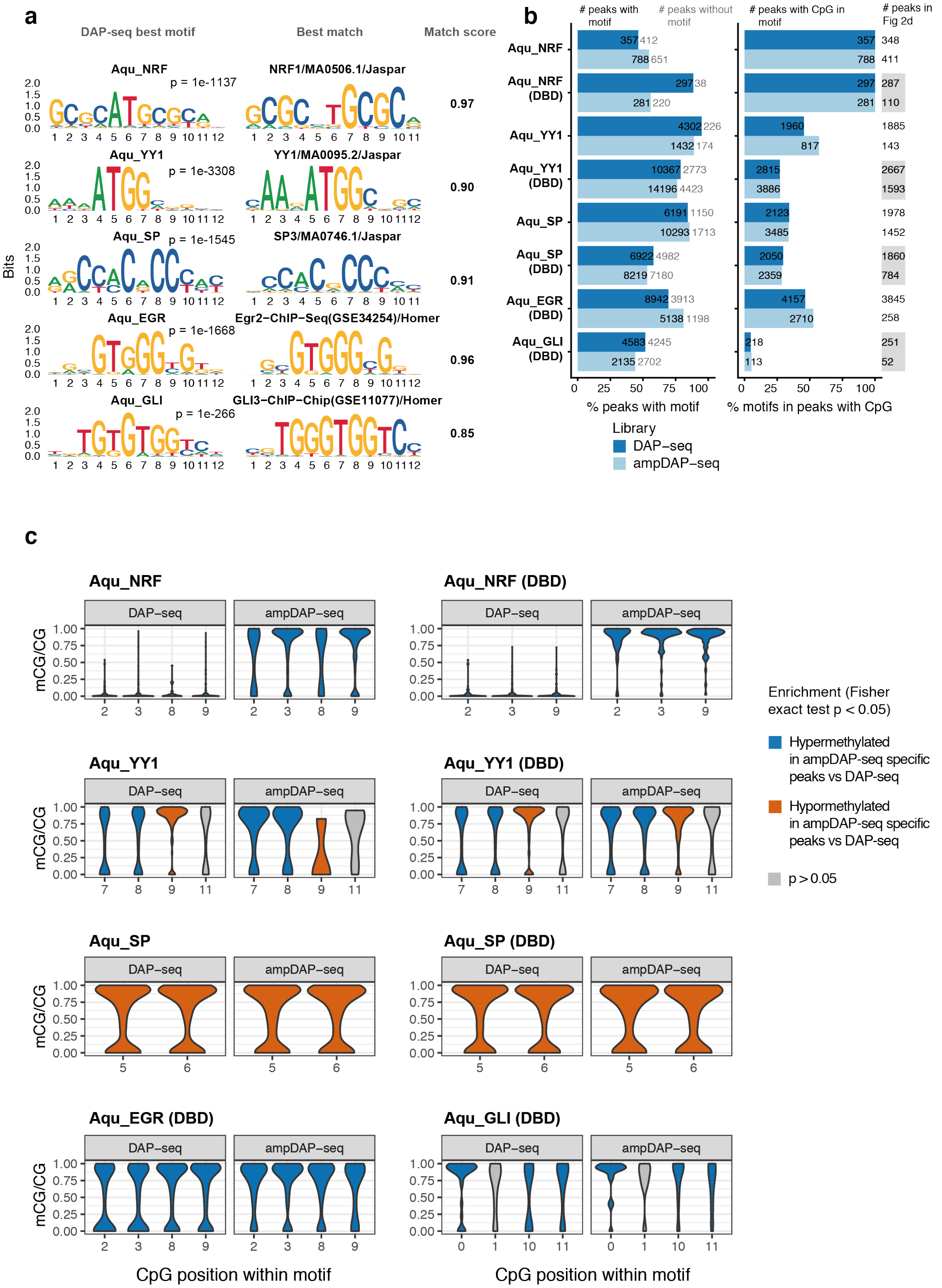
Supplementary Figure 3. *Amphimedon* shows widespread methylation on gene bodies and transposable elements. (**a**) Heatmap showing methylation levels on gene bodies of four non-bilaterian species. Genes have been classified as unmethylated based on having an overall gene body mCG level (mCG/CG) < 0.1. Colour legend on the right hand side, left, missing data (lack of coverage or lack of CpGs in window) is shown in grey. (**b**) Heatmap showing methylation levels on transposable elements. (**c**) Profile showing the mean methylation level on gene bodies classified by deciles of expression in stage-matched RNA-seq datasets. “Not expressed” genes are defined as genes encoding at least a PFAM domain (not associated with transposable elements) and with RNA-seq TPM < 1, however some of these genes might be in fact pseudo-genes misspredicted by *ab initio* gene prediction algorithms. Very low methylation levels on these “non expressed” genes in *Mnemiopsis* and *Nematostella* might suggest that either these genes are pseudogenes or that genes that are expressed in a very specific cell-type or time-point (and thus not-detected by bulk RNA-seq) reside in gene-poor regions. (**d**) Mean methylation level on gene bodies divided by gene orientation. Orientation of genes as shown in legend. Gene 3’ end orientation influences the methylation profiles of *Amphimedon, Sycon* and *Nematostella,* while *Mnemiopsis* is not affected. 5’ orientation shows increased depletion of upstream methylation in *Sycon* and *Nematostella* non-bidirectional promoters, while it does not affect *Mnemiopsis* genes. In the case of *Amphimedon*, bi-directional promoters show a wider unmethylated region than unidirectional promoters. For each species, the specific MethylC-seq library used to obtain the methylation levels is indicated below the species genus. Abbreviations: TSS (Transcriptional Start Site), TES (Transcriptional End Site).



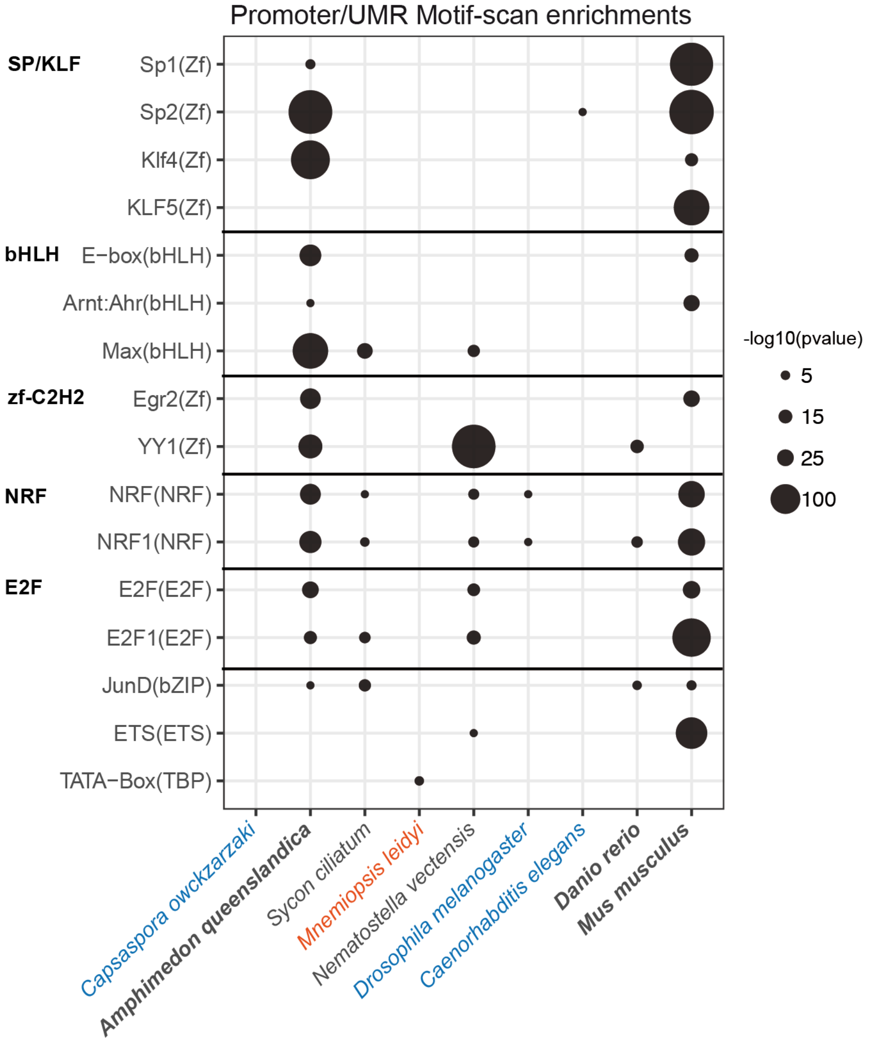
Supplementary Figure 4. CpG density at TSS for selected metazoan and holozoan species. (**a**) Mean CpG density on gene bodies of two vertebrates (*D. rerio*, *M. musculus*), two invertebrates lacking DNA methylation (*D. melanogaster, C. elegans*), and two unicellular holozoans lacking DNA methylation (*C. owczarzaki, C. fragrantissima*). The unicellular holozoans were selected for having gene annotations with curated untranslated regions (UTRs) based on strand-specific RNA-seq. (**b**) CpG density mean profile corrected by the CpG Observed versus Expected ratio on methylated gene bodies of the four non-bilaterian species and zebrafish. (**c**) Genomic CpG Observed versus Expected ratios of four sponge genomes. The cladogram represents the established phylogenetic relationships between sponge classes. (**d**) CpG Observed versus Expected ratio distribution on *Amphimedon* promoters.



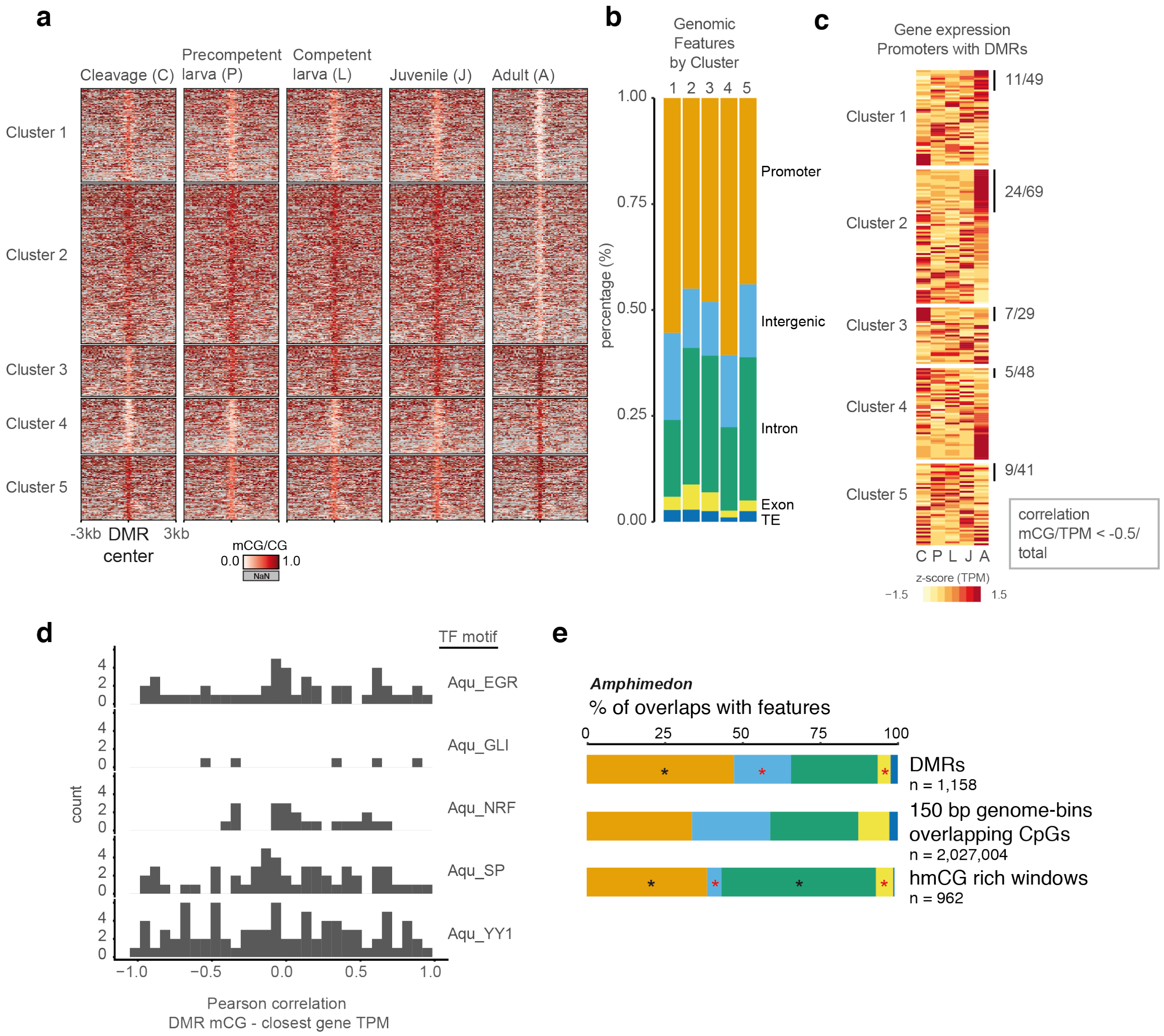
Supplementary Figure 5. Unmethylated region (UMR) characteristics in different species. (**a**) Size distribution of UMRs in several species as defined by MethylSeeker [59]. Boxplot centre lines are medians, box limits are quartiles 1 (Q1) and 3 (Q3), whiskers are 1.5 × interquartile range (IQR) and points are outliers. (**b**) Heatmap of methylation levels and H3K4me3 ChIP-seq signal (Reads Per Million, RPM) centered on UMRs for *Amphimedon* adult [28].



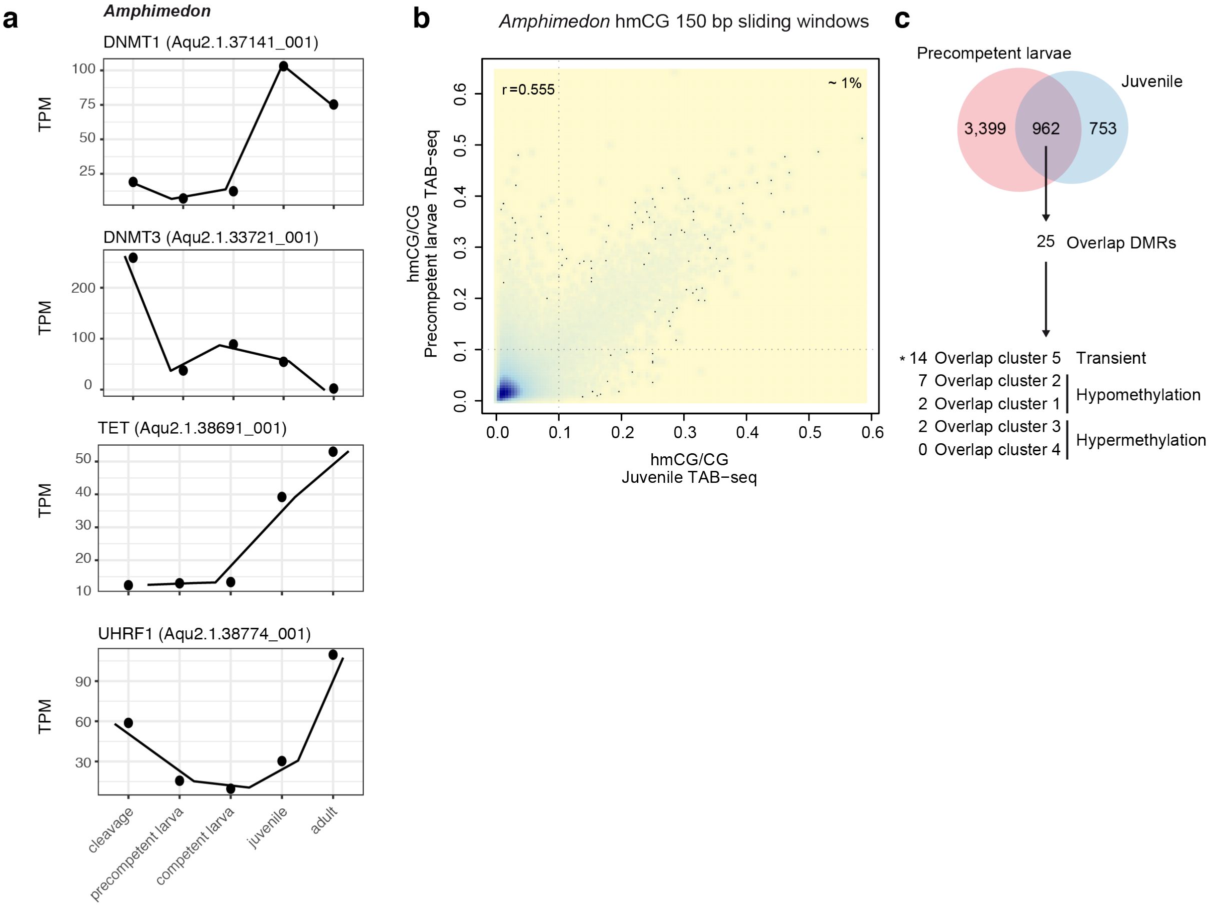
Supplementary Figure 6. DAP-seq and ampDAP-seq show conservation of binding motifs and methyl-sensitivity for *Amphimedon* transcription factors. (**a**) Left panel displays sequence logos showing the top *de novo* motif enrichment at a union of DAP-seq and ampDAP-seq peaks. Enrichment levels as defined by HOMER p-values [70]. Right panel shows the sequence logo of the best match motif in the HOMER database. (**b**) Number of peaks for each DAP-seq and ampDAP-seq sample, the percentage of peaks with the top enriched motif, the number of peaks with a motif spanning a CpG, and the number of peaks analysed in Figure 2d containing a CpG in a consensus position within the motif. (**c**) Distribution of mCG/CG values on distinct positions of the motif of each transcription factor as defined in panel (**a**). The values for the ampDAP-seq peaks are only those that do not overlap with the DAP-seq peaks. Colour code indicates the enrichment trend when comparing methylated calls versus unmethylated calls in DAP-seq peaks versus ampDAP-seq specific peaks as represented in the legend on the right hand side.



Supplementary Figure 7. Convergent motif enrichment in hypermethylated genomes. Motif scan enrichments in unmethylated promoters for species with mC, and H3K4me3 marked promoters for species lacking mC (shown in blue). *Mnemiopsis* is an exception as it shows DNA methylation in promoters unlike any other metazoan. P-value as obtained by a one-sided binomial test. For promoter definition, see Methods section.



Supplementary Figure 8. Promoter methylation usually does not anti-correlate with transcription in *Amphimedon*. (**a**) Heatmap showing methylation levels in *Amphimedon* DMRs divided by cluster of developmental trajectories. (**b**) Intersection of DMRs with genomic features in *Amphimedon* divided by cluster. (**c**) Heatmap showing expression levels of genes with DMRs in their promoter regions divided by cluster. (**d**) Distribution of pearson correlation values between DMR mCG level and the transcript abundance (TPM) of the associated gene(s) for each developmental stage for DMRs that overlap a methyl-sensitive transcription factor motif. (**e**) Intersection of DMRs and hydroxymethylated regions from **Figures 3b** and **Figure 4e** with genomic features in *Amphimedon* compared to background distribution, obtained through intersecting 150 bp bins (covering the whole genome, overlapping at least 1 CpG) with genomic features. Asterisks indicate two-sided Fisher's exact test p < 0.05 enrichment against expected, black asterisks indicate enrichment against background whereas red asterisks indicate depletion.



Supplementary Figure 9. *TET* transcriptional dynamics and conservation of hydroxymethylated cytosine across mid-developmental stages. (**a**) Transcriptional dynamics of *DNMT1*, *DNMT3*, *UHRF1* and *TET* in *Amphimedon*. (**b**) Scatter plot showing hydroxymethylation levels of 150 sliding windows harbouring at least 3 CpGs and mean coverage of 4x in both precompetent larva and juvenile *Amphimedon* samples. (**c**) Overlap of collapsed windows that have a hmCG level > 0.1 in both samples and its relationships with all DMRs (from figure 3a). Asterisk indicates two-sided Fisher's exact test $p < 0.05$ enrichment between the DMR cluster 5 overlap against the rest of DMRs.

Species	Sample (MethylC-seq)	mCG	mCHG	mCHH	Lambda	Coverage
		CG	CHG	CHH	mC/C	X
Amphimedon queenslandica	Cleavage	83.277	1.191	1.685	0.441	25.77
	Precompetent larva	79.626	0.941	1.305	0.434	26.64
	Competent larva	80.534	1.065	1.546	0.446	18.51
	Juvenile	81.732	0.580	0.610	0.366	32.62
	Adult	81.363	1.074	1.636	0.689	20.84
Sycon ciliatum	Adult	57.172	0.467	0.411	0.321	10.45
Mnemiopsis leidyi	3 hpf	4.989	0.378	0.384	0.387	7.70
	4.5 hpf	5.883	0.340	0.322	0.307	12.54
	8.5 hpf	5.684	0.316	0.325	0.313	22.55
	14 hpf	5.767	0.394	0.420	0.382	13.07
	30 hpf	3.653	0.381	0.422	0.392	13.21
	Juvenile	7.222	0.334	0.337	0.404	6.40
Nematostella vectensis	Blastula	14.008	0.510	0.493	/	13.74
	Gastrula	14.573	0.435	0.438	0.415	15.76
	Planula	13.551	0.718	1.042	0.505	13.35

Supplementary Table 1. MethylC-seq global statistics. Global methylation levels for each of the whole genome bisulfite sequencing samples computed for different sequence contexts, where H = A,T or C. The methylation level for the unmethylated lambda phage genome spike-in represents the non-conversion rate after the bisulfite reaction.

Species	Sample (TAB-seq)	mCG	mCHG	mCHH	Lambda	Lambda	PUC19	Cov .(X)
		CG	CHG	CHH	mCG/CG	mCH/CH	hmC/C	
Amphimedon queenslandica	Precompetent larva	2.019	0.670	0.549	1.874	0.396	52.332	25.77
	Juvenile	1.455	0.354	0.350	1.532	0.372	50.619	23.13
Nematostella vectensis	Gastrula	0.511	0.372	0.372	1.411	0.393	50.271	8.88

Supplementary Table 2. TAB-seq global statistics. Uncorrected global hydroxy-methylation levels for each of the TAB-seq samples computed for different sequence contexts. The methylation level in the CG context for the *SspI* methylated lambda spike-in represents the non-oxidation rate after the *Tet* enzyme reaction, the methylation level in the CH context of the lambda phage genome spike-in represents the non-conversion rate after the bisulfite reaction and the methylation level in the PUC19 plasmid spike-in represents the protection rate of the hydroxymethyl groups.