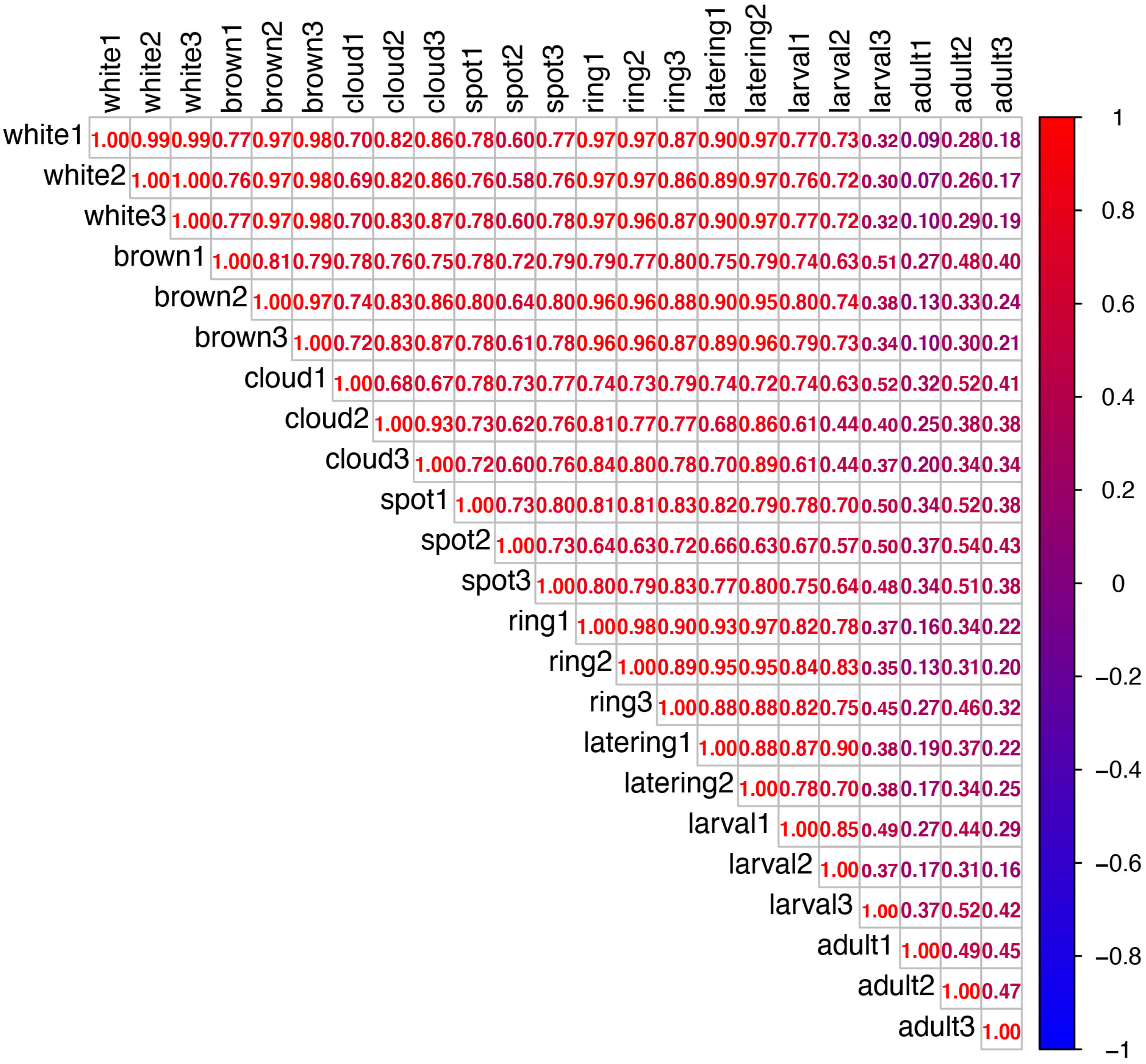
Supplementary material

Distal regulation, silencers and a shared combinatorial syntax are hallmarks of animal embryogenesis

Paola Cornejo-Páramo[1,2], Kathrein Roper[3], Sandie M Degnan[3], Bernard M Degnan[3], Emily S Wong[1,3,4]*

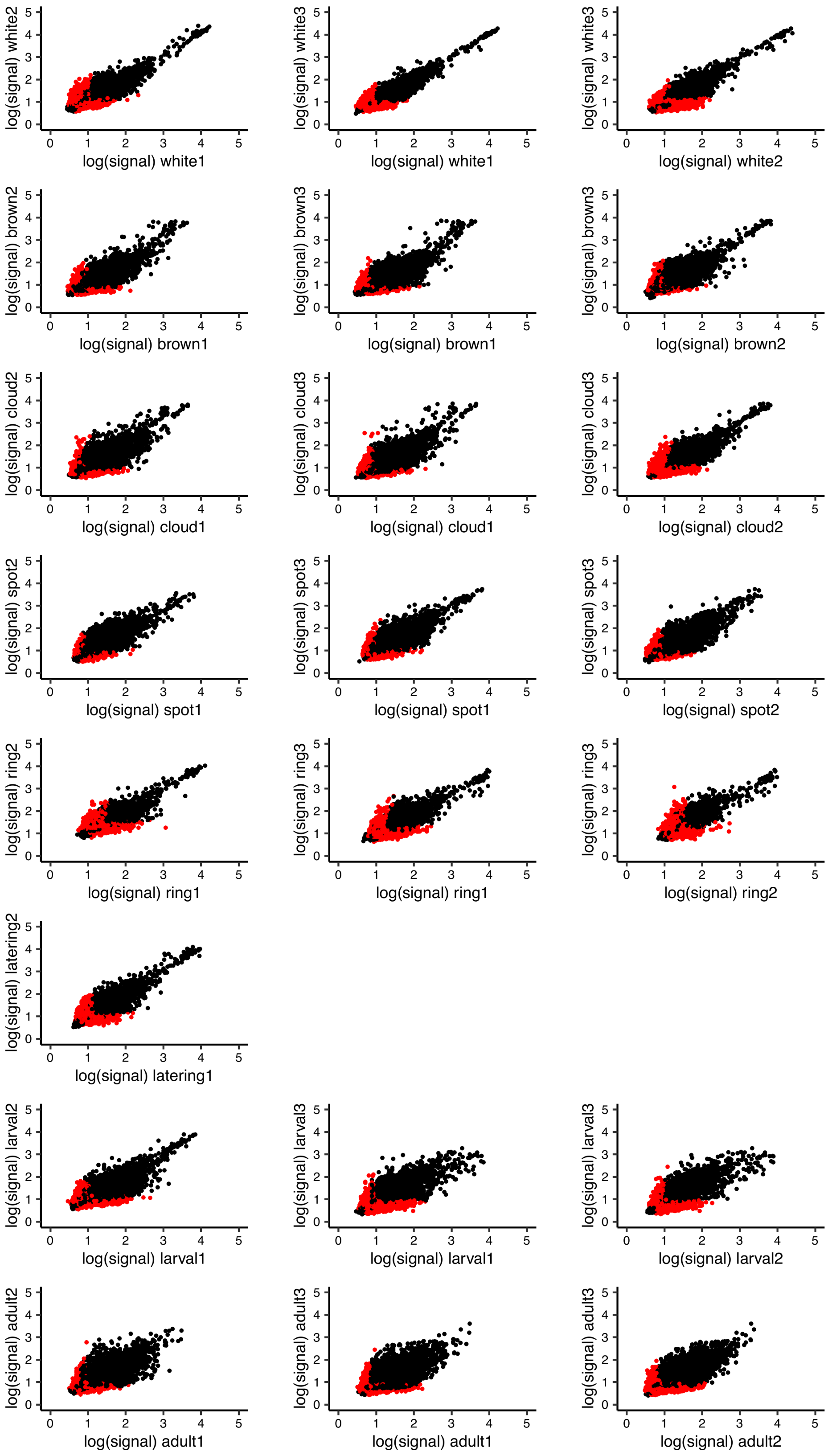
*correspondence to
e.wong@victorchang.edu.au

Table of Contents

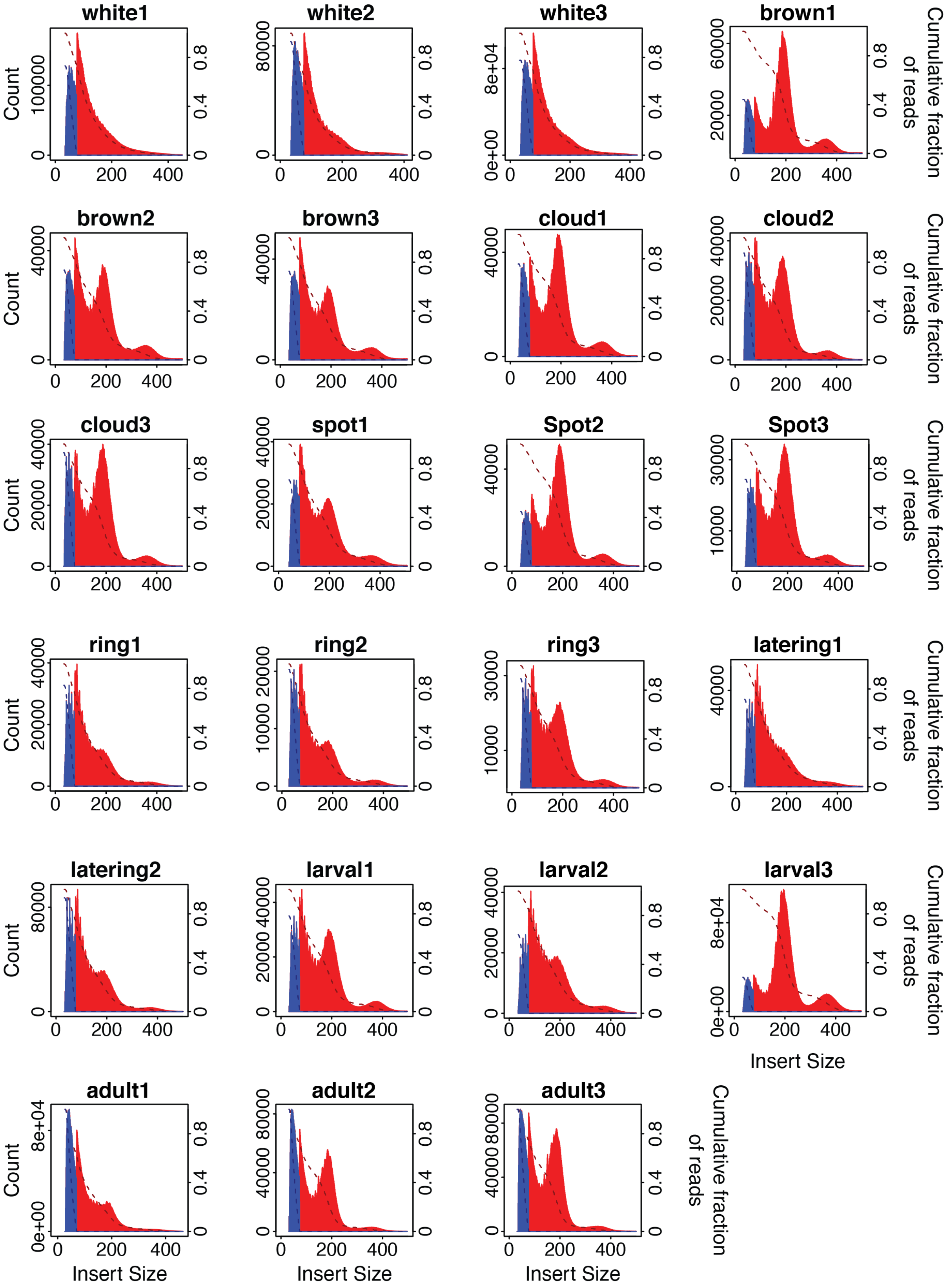


	white1	white2	white3	brown1	brown2	brown3	cloud1	cloud2	cloud3	spot1	spot2	spot3	ring1	ring2	ring3	latering1	latering2	larval1	larval2	larval3	adult1	adult2	adult3
white1	1.00	0.99	0.99	0.77	0.97	0.98	0.70	0.82	0.86	0.78	0.60	0.77	0.97	0.97	0.87	0.90	0.97	0.77	0.73	0.32	0.09	0.28	0.18
white2		1.00	1.00	0.76	0.97	0.98	0.69	0.82	0.86	0.76	0.58	0.76	0.97	0.97	0.86	0.89	0.97	0.76	0.72	0.30	0.07	0.26	0.17
white3			1.00	0.77	0.97	0.98	0.70	0.83	0.87	0.78	0.60	0.78	0.97	0.96	0.87	0.90	0.97	0.77	0.72	0.32	0.10	0.29	0.19
brown1				1.00	0.81	0.79	0.78	0.76	0.75	0.78	0.72	0.79	0.79	0.77	0.80	0.75	0.79	0.74	0.63	0.51	0.27	0.48	0.40
brown2					1.00	0.97	0.74	0.83	0.86	0.80	0.64	0.80	0.96	0.96	0.88	0.90	0.95	0.80	0.74	0.38	0.13	0.33	0.24
brown3						1.00	0.72	0.83	0.87	0.78	0.61	0.78	0.96	0.96	0.87	0.89	0.96	0.79	0.73	0.34	0.10	0.30	0.21
cloud1							1.00	0.68	0.67	0.78	0.73	0.77	0.74	0.73	0.79	0.74	0.72	0.74	0.63	0.52	0.32	0.52	0.41
cloud2								1.00	0.93	0.73	0.62	0.76	0.81	0.77	0.77	0.68	0.86	0.61	0.44	0.40	0.25	0.38	0.38
cloud3									1.00	0.72	0.60	0.76	0.84	0.80	0.78	0.70	0.89	0.61	0.44	0.37	0.20	0.34	0.34
spot1										1.00	0.73	0.80	0.81	0.81	0.83	0.82	0.79	0.78	0.70	0.50	0.34	0.52	0.38
spot2											1.00	0.73	0.64	0.63	0.72	0.66	0.63	0.67	0.57	0.50	0.37	0.54	0.43
spot3												1.00	0.80	0.79	0.83	0.77	0.80	0.75	0.64	0.48	0.34	0.51	0.38
ring1													1.00	0.98	0.90	0.93	0.97	0.82	0.78	0.37	0.16	0.34	0.22
ring2														1.00	0.89	0.95	0.95	0.84	0.83	0.35	0.13	0.31	0.20
ring3															1.00	0.88	0.88	0.82	0.75	0.45	0.27	0.46	0.32
latering1																1.00	0.88	0.87	0.90	0.38	0.19	0.37	0.22
latering2																	1.00	0.78	0.70	0.38	0.17	0.34	0.25
larval1																		1.00	0.85	0.49	0.27	0.44	0.29
larval2																			1.00	0.37	0.17	0.31	0.16
larval3																				1.00	0.37	0.52	0.42
adult1																					1.00	0.49	0.45
adult2																						1.00	0.47
adult3																							1.00

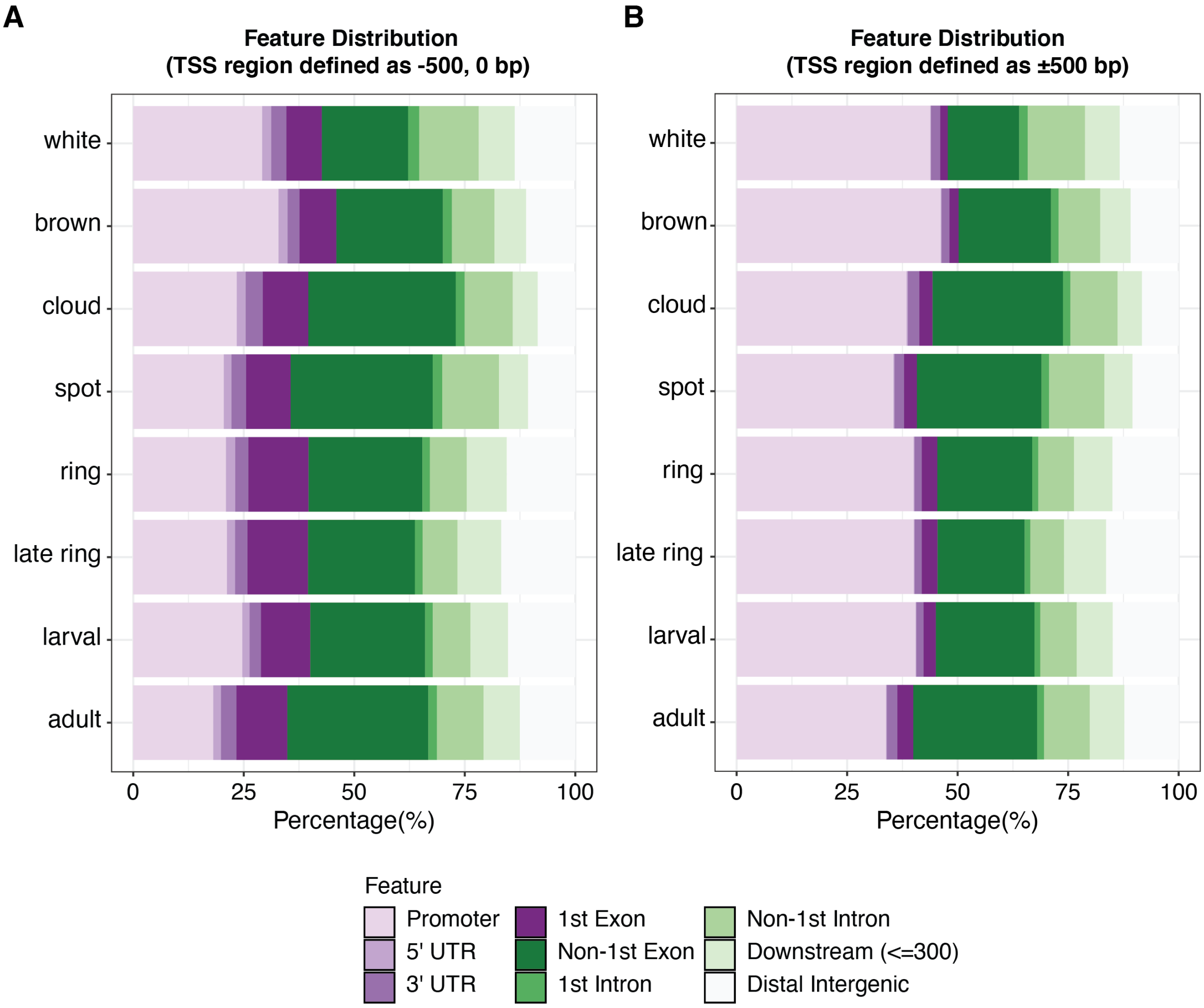
Supplemental Fig. S1. Pearson's correlation of ATAC-seq data across *Amphimedon* developmental stages Pearson's correlation of normalised ATAC-seq counts is shown for all *Amphimedon* ATAC-seq libraries (n = 23 libraries) across developmental stages (n = 3, n = 3, n = 3, n = 3, n = 3, n = 2, n = 3, and n = 3 libraries for white, brown, cloud, spot, ring, late ring, larval and adult stages respectively).



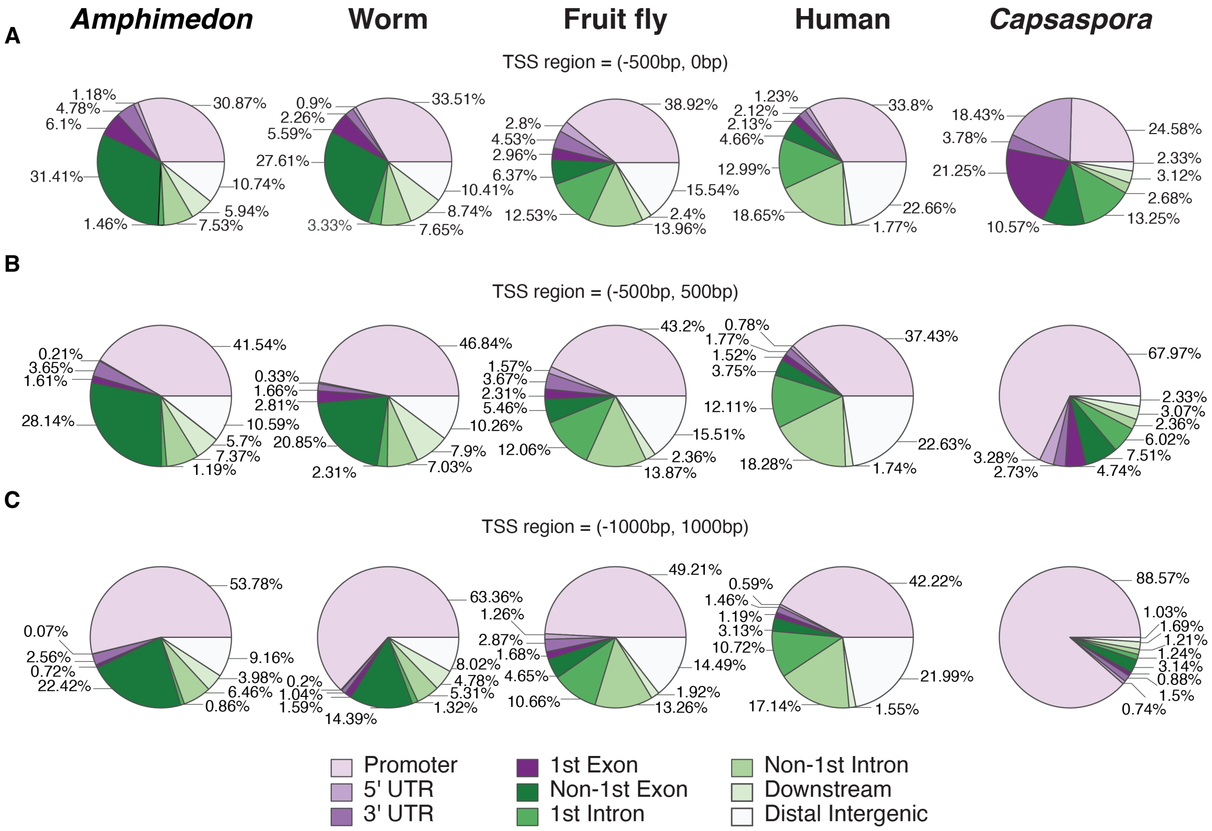
Supplemental Fig. S2. Irreproducible Discovery Rate (IDR) of *Amphimedon* ATAC-seq libraries Pairwise IDR of *Amphimedon* ATAC-seq libraries of every developmental stage is shown. Overlapping peaks with IDR <= 10% are illustrated in black.



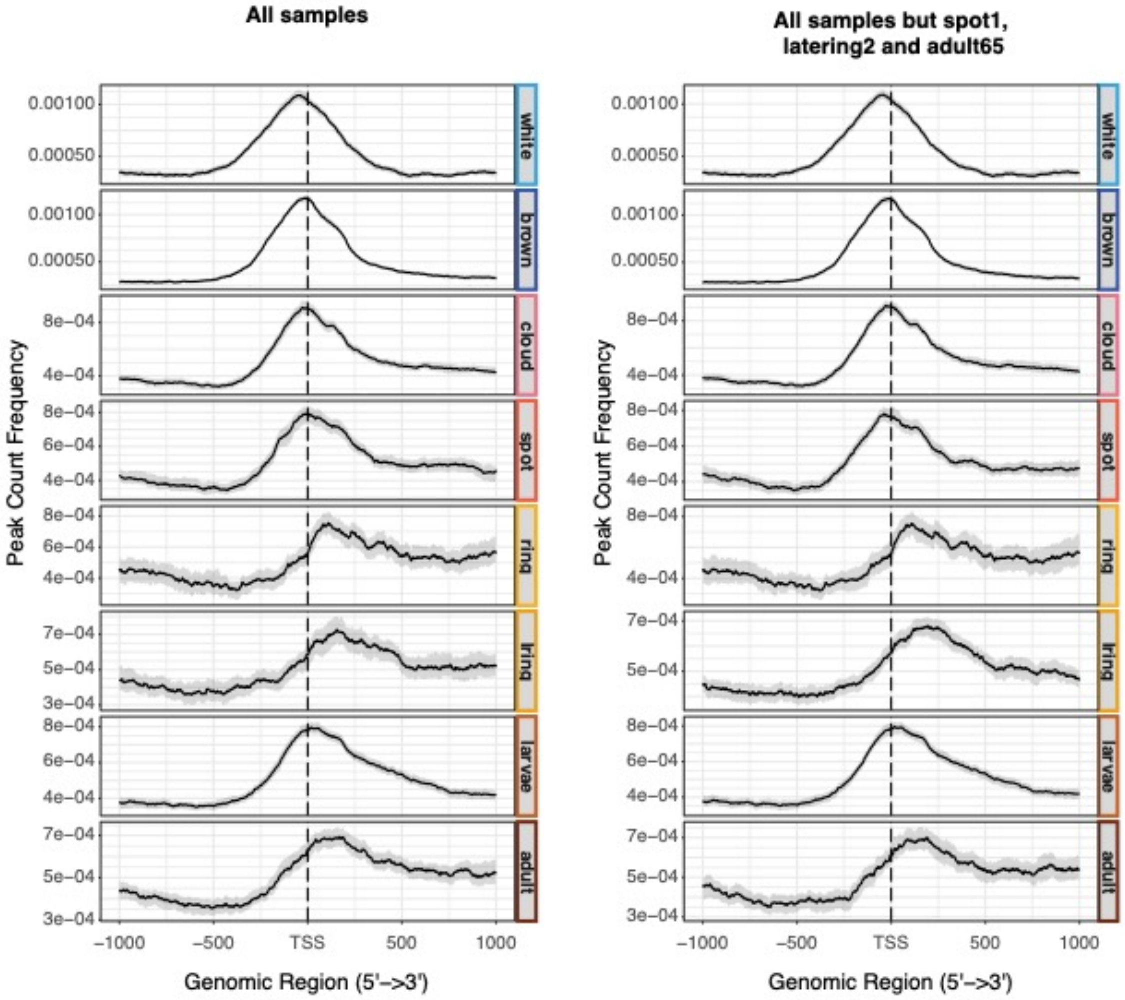
Supplemental Fig. S3. Density of insert fragment size in *Amphimedon* ATAC-seq libraries Forward to reverse (FR) fragments are shown in red and reverse to forward (RF) fragments are shown in blue.



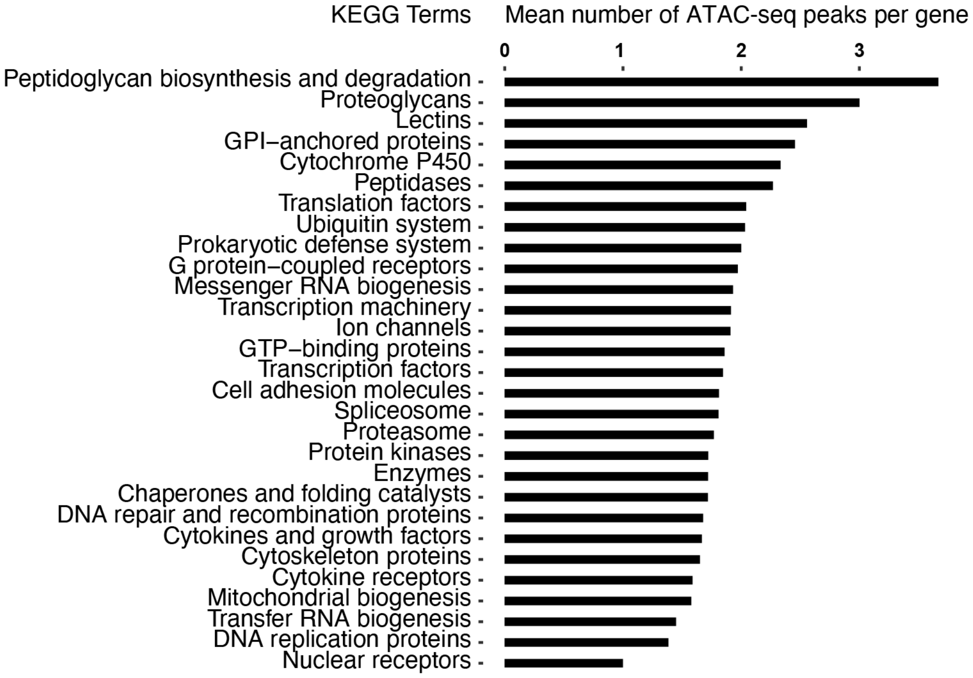
Supplemental Fig. S4. Distribution of *Amphimedon* ATAC-seq peaks in multiple genomic features by developmental stage *Amphimedon* developmental stages are indicated in chronological order (top to bottom), the colour code represents different genomic features. The TSS region was defined as -500 to 0 bp **(A)** and as ±500 bp **(B)**.



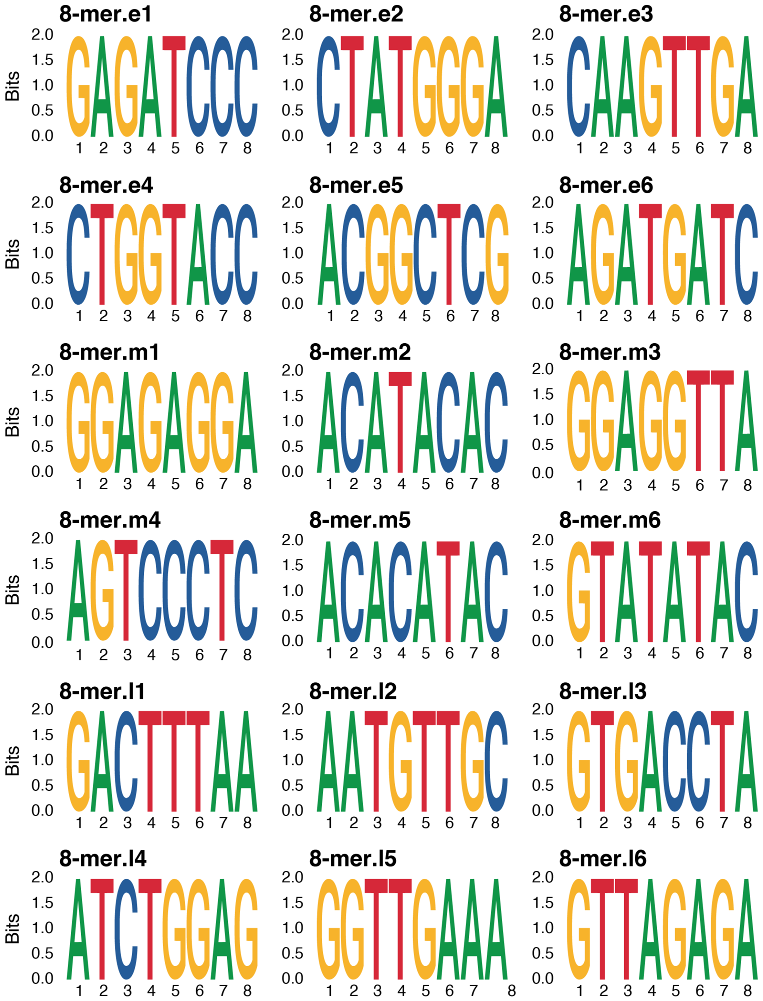
Supplemental Fig. S5. Distribution of *Amphimedon*, worm, fruit fly, human and *Capsaspora cis*-regulatory regions into multiple genomic features Transcription start site (TSS) region was defined from -500 to 0 bp from the TSS **(A)**, from -500 to 500 bp **(B)** and from -1000 to 1000 bp **(C)**.



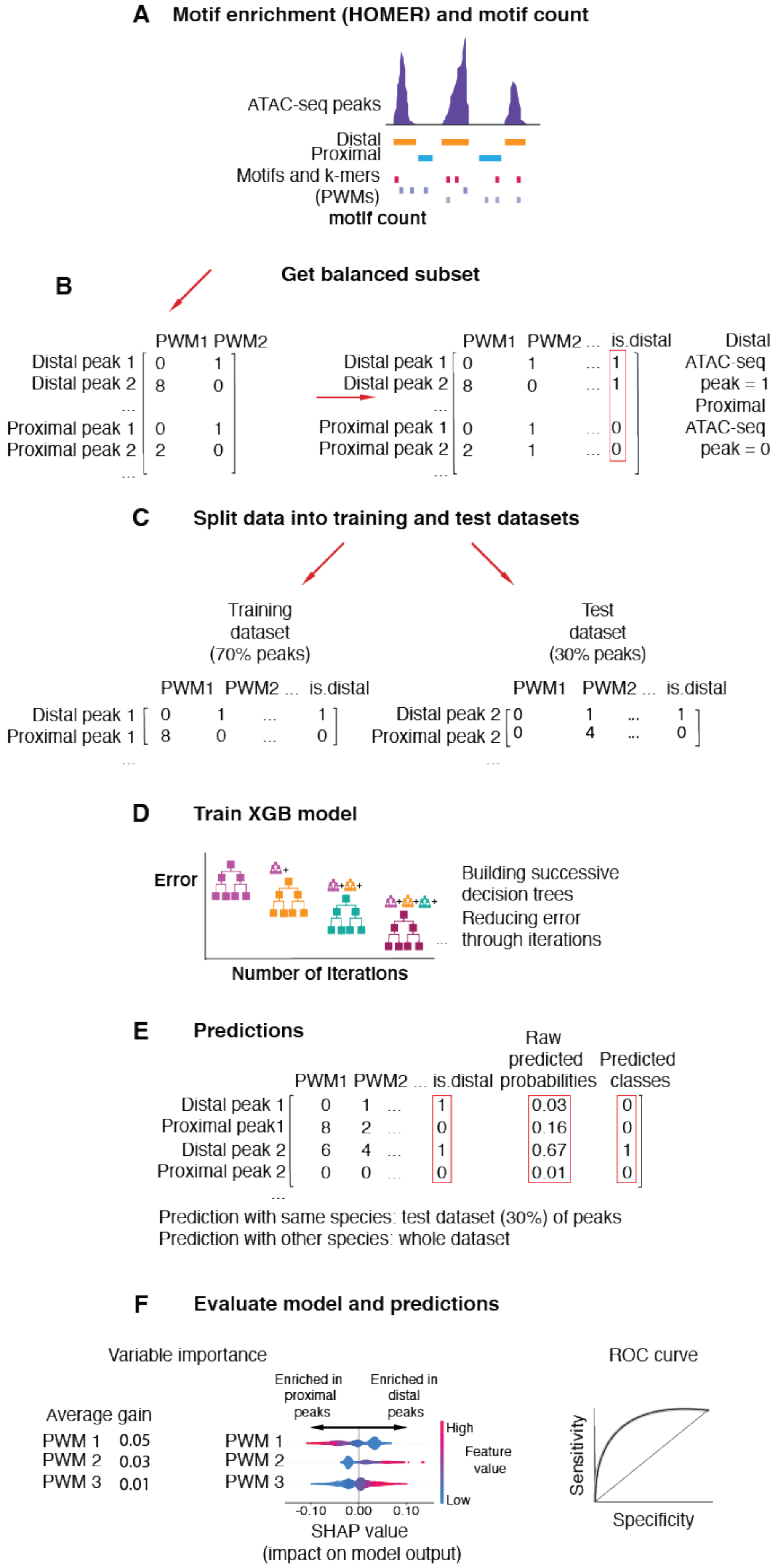
Supplemental Fig. S6. Little change to chromatin accessibility around the TSS with and without more variable samples Peaks were used only if at least 50% of bases overlapped across biological replicates.



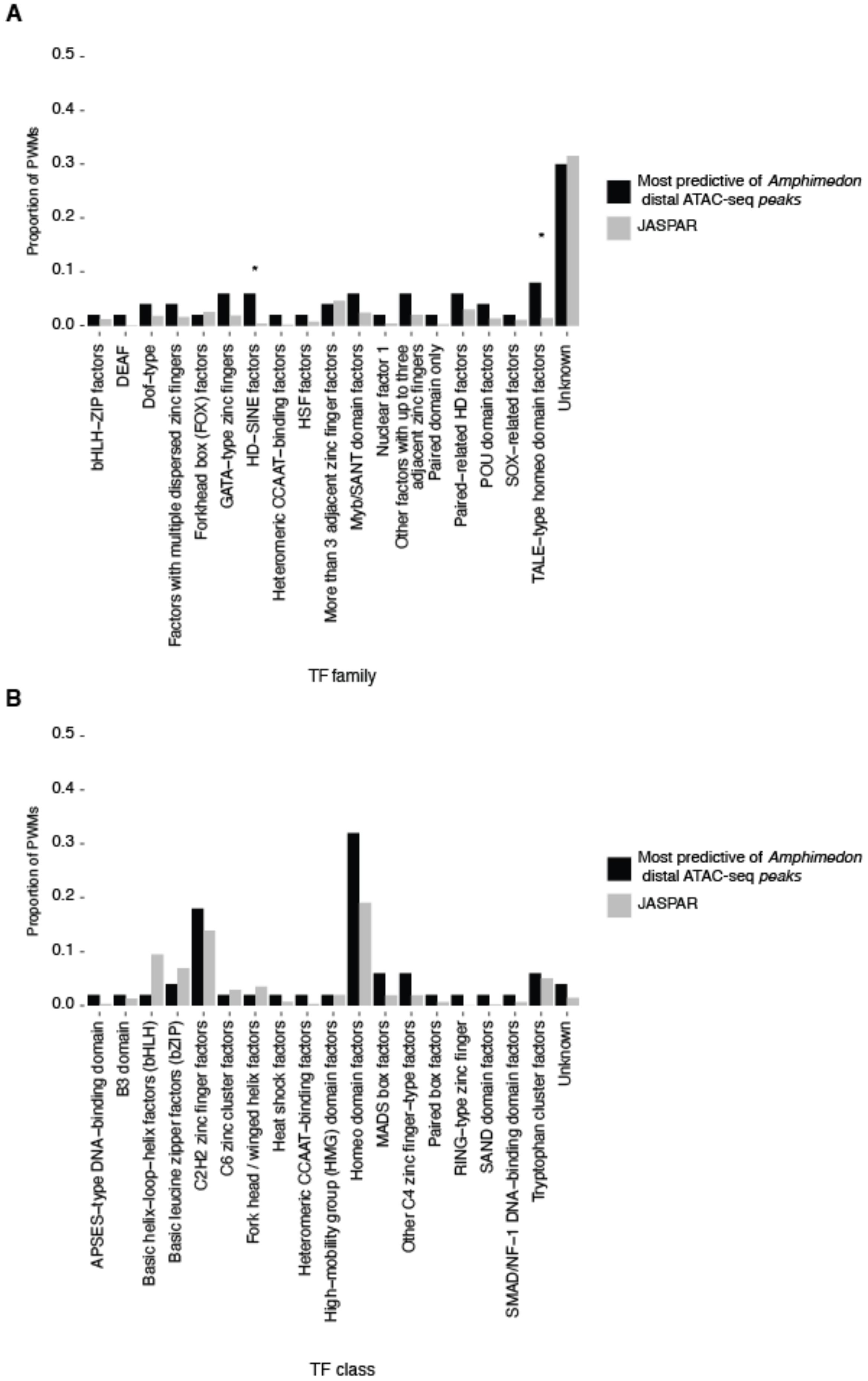
Supplemental Fig. S7. KEGG functional categories of genes proximal to ATAC-seq peaks Mean number of ATAC-seq peaks proximal to *Amphimedon* genes grouped by KEGG functional categories. Proximity defined as ±500 bp from TSS.



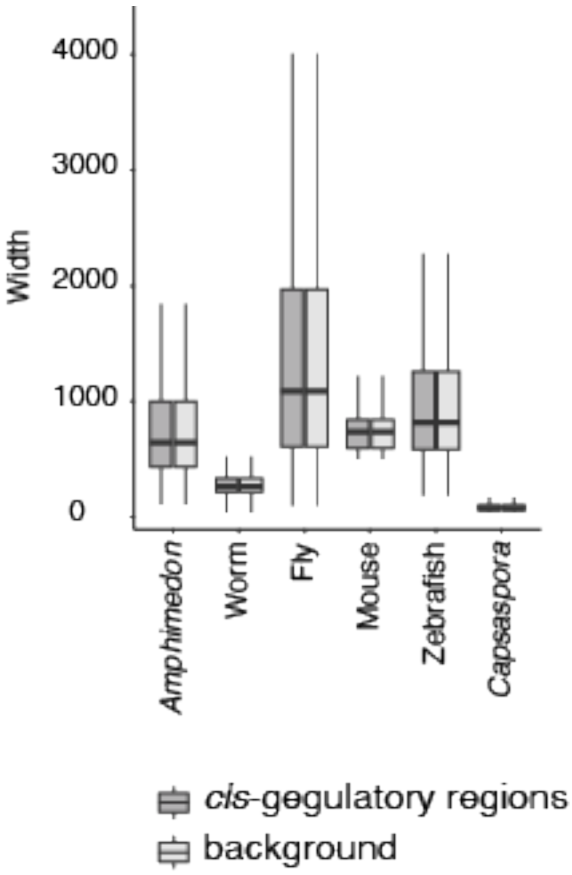
Supplemental Fig. S8. ***Amphimedon de novo k*-mers** *De novo* 8-mers that demarcate each broad developmental stage in *Amphimedon*, denoted as early: 8-mer.e, mid: 8-mer.m, late: 8-mer.l. List of 8-mers and their significance levels are shown in Table S6. Similarity of 8-mers and JASPAR motifs is shown in Table S7.


A
Motif enrichment (HOMER) and motif count
ATAC-seq peaks
Distal
Proximal
Motifs and k-mers
(PWMs)
motif count
B
Get balanced subset
PWM1 PWM2
Distal peak 1 | 0 1
Distal peak 2 | 8 0
...
Proximal peak 1 | 0 1
Proximal peak 2 | 2 0
...
PWM1 PWM2 ... is.distal
Distal peak 1 | 0 1 ... 1
Distal peak 2 | 8 0 ... 1
...
Proximal peak 1 | 0 1 ... 0
Proximal peak 2 | 2 1 ... 0
...
Distal ATAC-seq peak = 1
Proximal ATAC-seq peak = 0
C
Split data into training and test datasets
Training dataset (70% peaks)
Test dataset (30% peaks)
PWM1 PWM2 ... is.distal
Distal peak 1 | 0 1 ... 1
Proximal peak 1 | 8 0 ... 0
...
PWM1 PWM2 ... is.distal
Distal peak 2 | 0 1 ... 1
Proximal peak 2 | 0 4 ... 0
...
D
Train XGB model
Error
Number of Iterations
Building successive decision trees
Reducing error through iterations
E
Predictions
PWM1 PWM2 ... is.distal | Raw predicted probabilities | Predicted classes
Distal peak 1 | 0 1 ... 1 | 0.03 | 0
Proximal peak1 | 8 2 ... 0 | 0.16 | 0
Distal peak 2 | 6 4 ... 1 | 0.67 | 1
Proximal peak 2 | 0 0 ... 0 | 0.01 | 0
...
Prediction with same species: test dataset (30%) of peaks
Prediction with other species: whole dataset
F
Evaluate model and predictions
Variable importance
Average gain
PWM 1 0.05
PWM 2 0.03
PWM 3 0.01
Enriched in proximal peaks
Enriched in distal peaks
PWM 1
PWM 2
PWM 3
High
Feature value
Low
-0.10 0.00 0.10
SHAP value
(impact on model output)
ROC curve
Sensitivity
Specificity

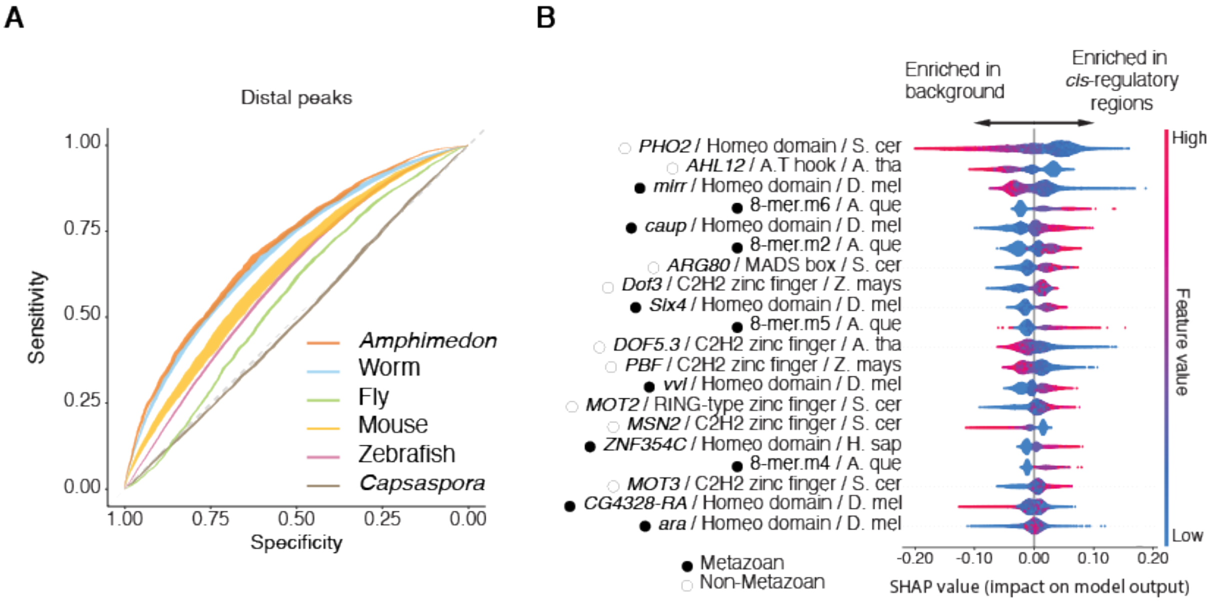
Supplemental Fig. S9. Extreme gradient boosting machine (XGB) pipeline for the prediction of distal *cis*-regulatory regions against proximal *cis*-regulatory regions using known and *de novo* PMWs in *Amphimedon* data (A) Motif enrichment of *Amphimedon* distal and proximal *cis*-regulatory regions. Peak state is codified in a binary variable (distal *cis*-regulatory regions = 1, proximal *cis*-regulatory regions = 0). **(B)** Selection of a balanced dataset of peaks (same number of distal and proximal *cis*-regulatory regions). **(C)** Splitting of data into 'training' and 'test' datasets. **(D)** training of XGB model. **(E)** Prediction of peaks states with *Amphimedon* test dataset and datasets of other species. **(F)** Evaluation of prediction performance by assessing the variable importance (average gain and SHAP values) of motifs and by analyzing ROC curves.


A
0.5
0.4
0.3
0.2
0.1
0.0
Proportion of PWMs
bHLH–ZIP factors
DEAF
Dof–type
Factors with multiple dispersed zinc fingers
Forkhead box (FOX) factors
GATA–type zinc fingers
HD–SINE factors
Heteromeric CCAAT–binding factors
HSF factors
More than 3 adjacent zinc finger factors
Myb/SANT domain factors
Nuclear factor 1
Other factors with up to three adjacent zinc fingers
Paired domain only
Paired–related HD factors
POU domain factors
SOX–related factors
TALE–type homeo domain factors
Unknown
*
*
TF family
Most predictive of *Amphimedon* distal ATAC-seq *peaks*
JASPAR
B
0.5
0.4
0.3
0.2
0.1
0.0
Proportion of PWMs
APSES–type DNA–binding domain
B3 domain
Basic helix–loop–helix factors (bHLH)
Basic leucine zipper factors (bZIP)
C2H2 zinc finger factors
C6 zinc cluster factors
Fork head / winged helix factors
Heat shock factors
Heteromeric CCAAT–binding factors
High–mobility group (HMG) domain factors
Homeo domain factors
MADS box factors
Other C4 zinc finger–type factors
Paired box factors
RING–type zinc finger
SAND domain factors
SMAD/NF–1 DNA–binding domain factors
Tryptophan cluster factors
Unknown
TF class
Most predictive of *Amphimedon* distal ATAC-seq *peaks*
JASPAR

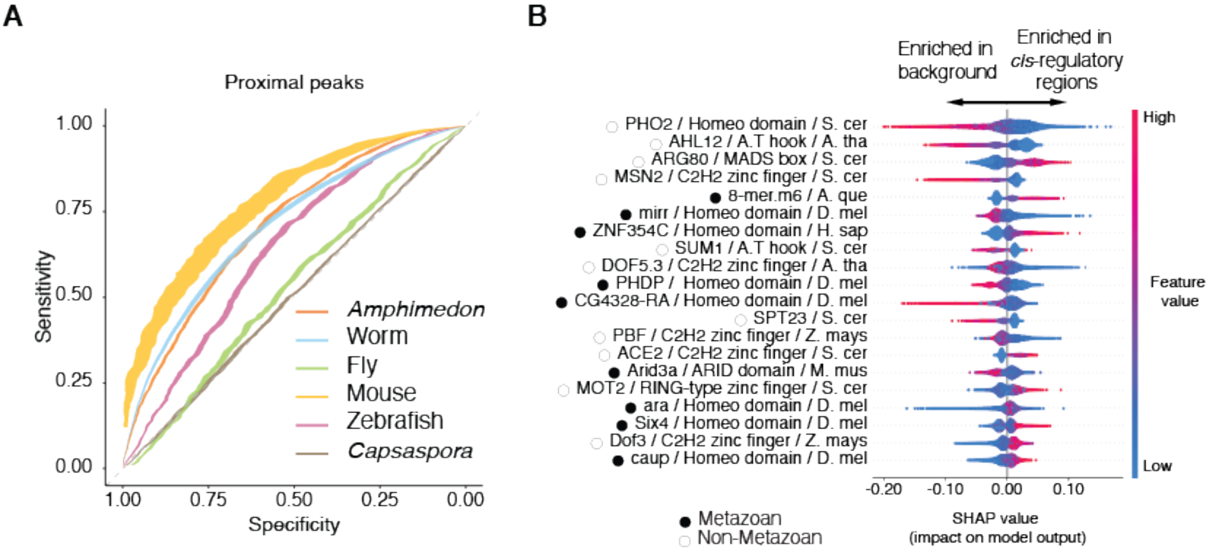
Supplemental Fig. S10. TF families and classes of the most predictive motifs of *Amphimedon* distal ATAC-seq peaks compared to the JASPAR reference set Proportion of unique PWMs belonging to different TF families **(A)** and classes **(B)** among the top 50 most predictive motifs of *Amphimedon* distal *cis*-regulatory regions (selected based on the greatest difference in TF motif numbers between ATAC-seq peaks and genome-wide background peak). The proportions of PWMs of the same classes and families in JASPAR database are shown (n = 1646 PWMs in JASPAR database). HD-SINE and TALE-type homeo domain factors were enriched among the most predictive motifs of distal *cis*-regulatory regions (FDR = 0.009 and FDR = 0.04, respectively).



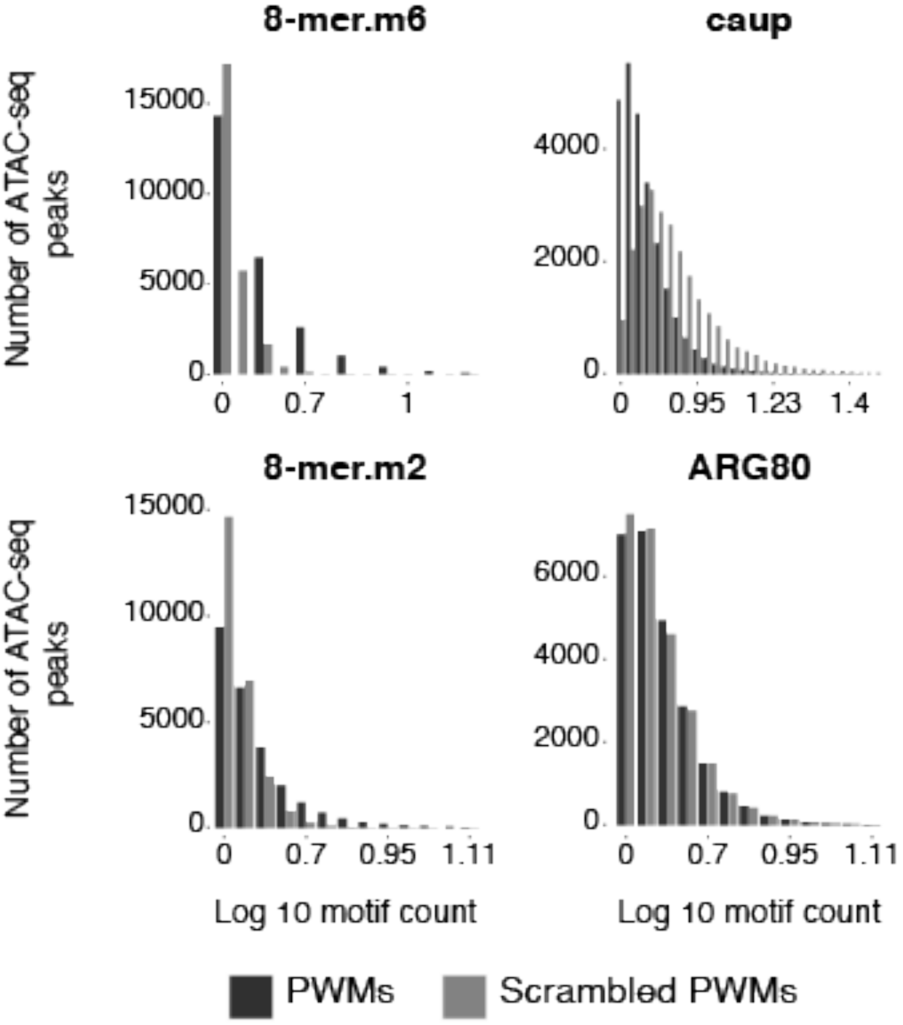
Supplemental Fig. S11. Width of *cis*-regulatory regions across species Width of *cis*-regulatory regions is shown for *Amphimedon*, worm, fly, mouse, zebrafish and *Capsapsora* (black) along with the corresponding background sequences (grey).



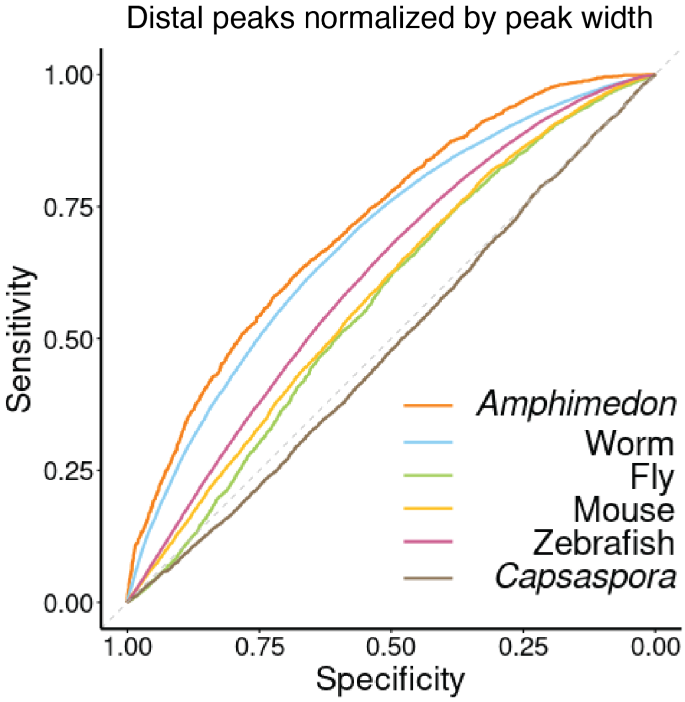
Supplemental Fig. S12. Receiver Operating Characteristic (ROC) curves and SHAP values of XGB models trained to distinguish distal *cis*-regulatory regions from background (A) ROC curves with 95% confidence intervals of the prediction of distal *cis*-regulatory regions from background sequences in *Amphimedon*, worm, fly, mouse, zebrafish and *Capsaspora* using JASPAR motifs and *Amphimedon* 8-mers counts (n = 10 XGB models). **(B)** SHAP values of most important motifs and 8-mers for the prediction of distal *cis*-regulatory regions (selected based on SHAP values, n = 1 XGB model). The plot shows motif importance and effect. Motifs are ordered according to their importance. Each dot reflects the motif at a peak. Colours reflect the count of the motif and the SHAP value show the impact on the prediction. S.cer = *S. cerevisiae*, A. tha = *A. thaliana*, D. mel = *D. melanogaster*, A. que = *A. queenslandica*, Z. mays = *Z. mays*, and H. sap = *H. sapiens*. Metazoan and non-metazoan PWMs are indicated with back filled and black outlined circles, respectively.



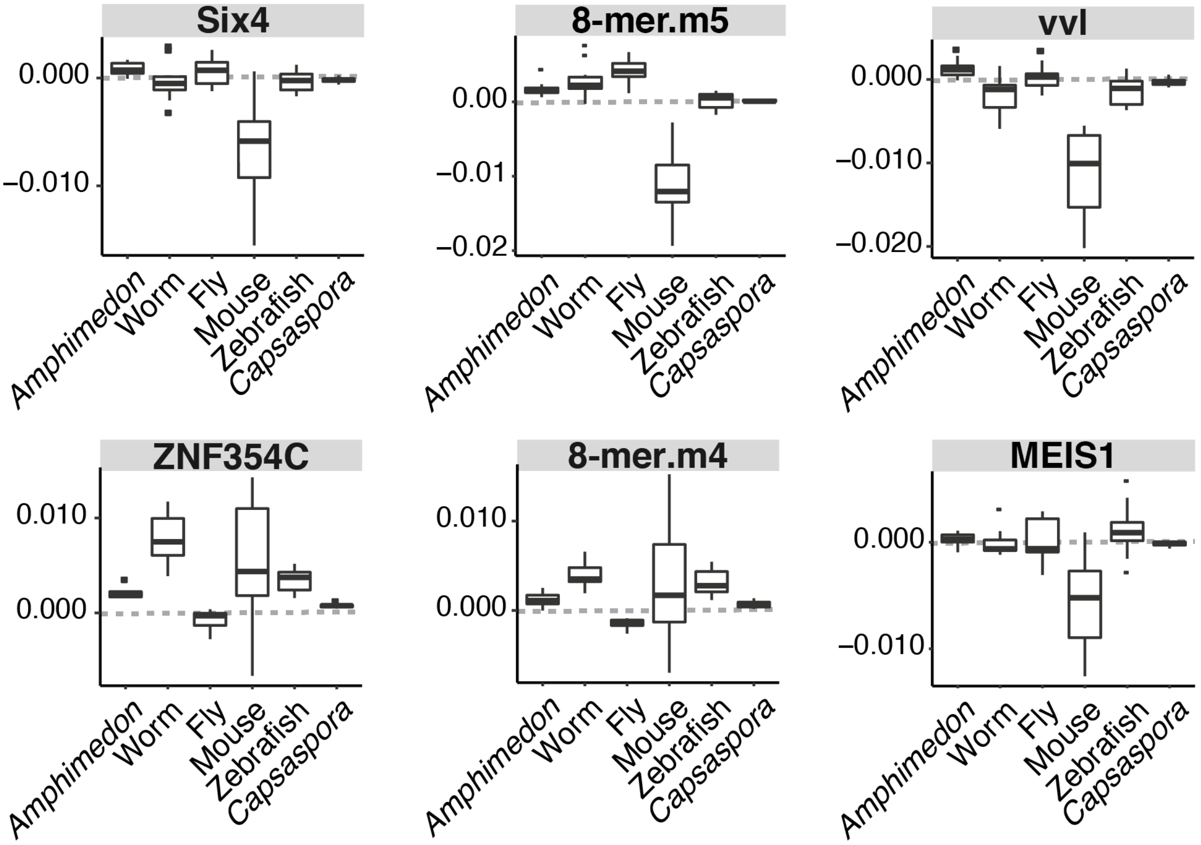
Supplemental Fig. S13. Receiver Operating Characteristic (ROC) curves and SHAP values of XGB models trained to distinguish proximal *cis*-regulatory regions from background (A) ROC curves with 95% confidence intervals of the prediction of proximal *cis*-regulatory regions from background sequences in *Amphimedon*, worm, fly, mouse, zebrafish and *Capsaspora* using JASPAR motifs and *Amphimedon* 8-mers counts (n = 10 XGB models). **(B)** SHAP values of most important motifs and 8-mers for the prediction of proximal *cis*-regulatory regions (selected based on SHAP values, n = 1 XGB model). The plot shows motif importance and effect. Motifs are ordered according to their importance. Each dot reflects the motif at a peak. Colours reflect the count of the motif and the SHAP value show the impact on the prediction. S.cer = *S. cerevisiae*, A. tha = *A. thaliana*, A. que = *A. queenslandica*, D. mel = *D. melanogaster*, H. sap = *H. sapiens*, Z. mays = *Z. mays* and M. mus = *M. musculus*. Metazoan and non-metazoan PWMs are indicated with back filled and black outlined circles, respectively.



Supplemental Fig. S14. Frequencies of motifs per peak for the top four most predictive motifs in *Amphimedon* distal regions JASPAR and 8-mers (black) and scrambled (grey) PWMs were used to identify motifs in *Amphimedon* distal ATAC-seq peaks and motif frequency per peak is shown. Counts were transformed by base 10 logarithm ($\log_{10}$(counts +1)). Bar plots were truncated to remove values with low frequency.



Supplemental Fig. S15. ROC curves of the prediction of distal *cis*-regulatory regions on PWMs counts normalised by peak width ROC curves of the prediction of distal *cis*-regulatory regions with an XGB model trained on the frequencies of JASPAR motifs plus *Amphimedon* 8-mers. These counts were adjusted for peak width by dividing the motif counts by peak width (bp) and multiplying by 10,000.



Supplemental Fig. S16. dAUC values of most predictive motifs of *Amphimedon* distal *cis*-regulatory regions dAUC values of highly predictive motifs and 8-mers of *Amphimedon* distal *cis*-regulatory regions (most predictive motifs shown in **Fig 5F**). Motifs were selected based on their SHAP values and their enrichment in *cis*-regulatory peak (Mann-Whitney *U*, FDR < 0.05).

Legends for Supplementary Tables

Supplemental Table S1. Alignment statistics of *Amphimedon* ATAC-seq libraries.

Supplemental Table S2. FRiP scores of *Amphimedon* ATAC-seq libraries.

Supplemental Table S3. Number of peaks, mean peak width and mean insert size in *Amphimedon* ATAC-seq libraries.

Supplemental Table S4. Gene ontology (GO) process of genes neighbouring distal *Amphimedon cis*-regulatory regions that align to the human genome.

Supplemental Table S5. Accessibility deviation scores of JASPAR motifs across *Amphimedon* developmental stages.

Supplemental Table S6. Top *de novo* motifs (8-mers) enriched in *Amphimedon* broad developmental stages (early, mid and late).

Supplemental Table S7. Similarity of *Amphimedon* 8-mers and known PMWs from JASPAR database.

Supplemental Table S8. Pairwise synergy of top 10 motifs enriched in every broad *Amphimedon* developmental stage.

Supplemental Table S9. Pairwise correlation coefficients of top motifs enriched in *Amphimedon* broad developmental stages at non-co-locating peaks.

Supplemental Table S10. Performance statistics of 10 XGB models trained on known motifs and *de novo Amphimedon k*-mers to distinguish distal from proximal *Amphimedon cis*-regulatory regions.

Supplemental Table S11. Variable importance (average gain) of known motifs and *de novo Amphimedon k*-mers across 10 XGB models trained to distinguish *Amphimedon* distal from proximal *cis*-regulatory regions.

Supplemental Table S12. Pearson's correlation of average gain values across XGB models trained on known motifs and *de novo Amphimedon* 8-mers to distinguish *Amphimedon* distal from proximal *cis*-regulatory regions.

Supplemental Table S13. Enrichment of motifs in distal *cis*-regulatory regions and background peaks of *Amphimedon*, worm, fly, mouse, zebrafish and *Capsaspora* (Mann-Whitney *U* test).

Supplemental Table S14. Variable importance (average gain) of known motifs and *de novo Amphimedon k*-mers across 10 XGB models trained to distinguish *Amphimedon* distal *cis*-regulatory regions.

Supplemental Table S15. Pearson's correlation of average gain values across XGB models trained on known motifs and *de novo Amphimedon* 8-mers to distinguish *Amphimedon* distal *cis*-regulatory regions.

Supplemental Table S16. Variable importance (average gain) of known motifs and *de novo Amphimedon k*-mers across 10 XGB models trained to distinguish *Amphimedon* proximal *cis*-regulatory regions.

Supplemental Table S17. Pearson's correlation of average gain values across XGB models trained on known motifs and *de novo Amphimedon* 8-mers to distinguish *Amphimedon* proximal *cis*-regulatory regions.

Supplemental Table S18. Performance statistics of 10 XGB models trained on known motifs and *de novo Amphimedon k*-mers on *Amphimedon* distal *cis*-regulatory regions and background sequences.

Supplemental Table S19. Performance statistics of 10 XGB models trained on known motifs and *de novo Amphimedon k*-mers on *Amphimedon* proximal *cis*-regulatory regions and background sequences.

Supplemental Table S20. Metazoan motifs not found in *Amphimedon cis*-regulatory regions using HOMER.