Distinct core promoter codes drive transcription initiation at key developmental transitions in a marine chordate

Gemma B. Danks, Pavla Navratilova, Boris Lenhard, Eric Thompson

Supplemental figures and legends

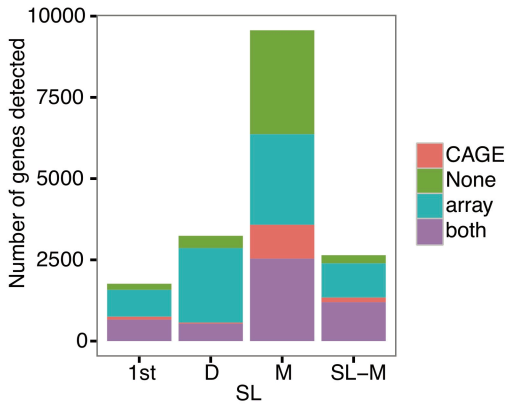

Figure S1.

The number of annotated genes detected by both CAGE and tiling arrays, CAGE only, tiling array only and those not detected by either method. CAGE libraries were depleted of *trans*-spliced transcripts and CAGE is unable to profile downstream operon genes unless they are transcribed from internal promoters. M = monocistronic (not *trans*-spliced); SL-M = *trans*-spliced monocistronic; 1^{st} = first gene in an operon; D = downstream operon gene.

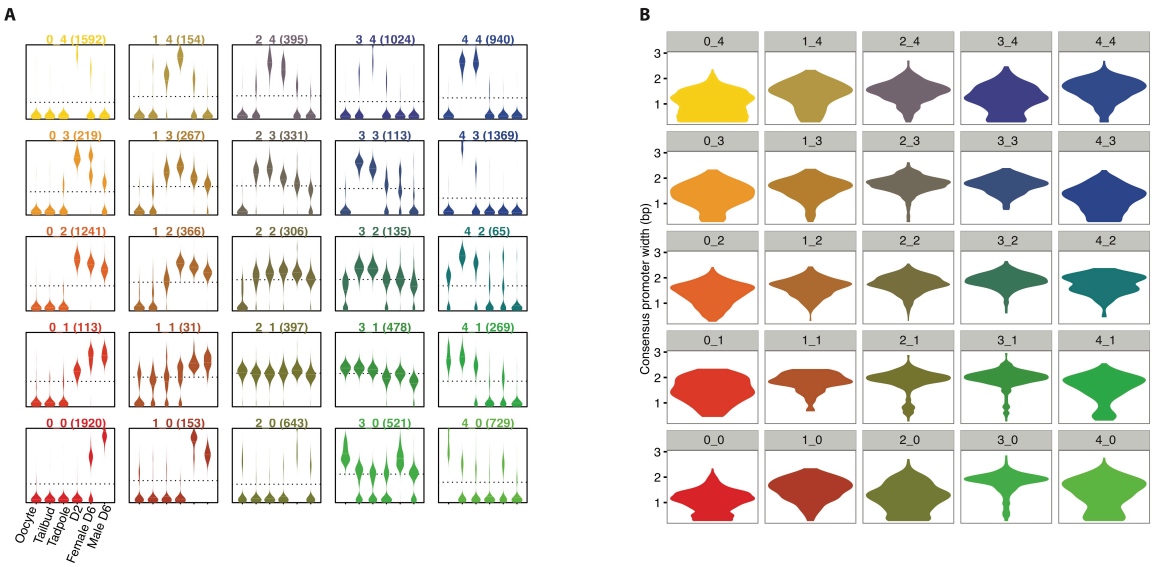

Figure S2.

Expression profiles obtained from self-organising map clustering of consensus promoter regions. **(A)** Each beanplot shows the distribution of relative expression of genes from promoters (number above each plot) within each cluster at each developmental stage (x-axis) indicated in the bottom left plot. **(B)** Beanplots show the distribution of promoter widths within each expression cluster in (**A**).

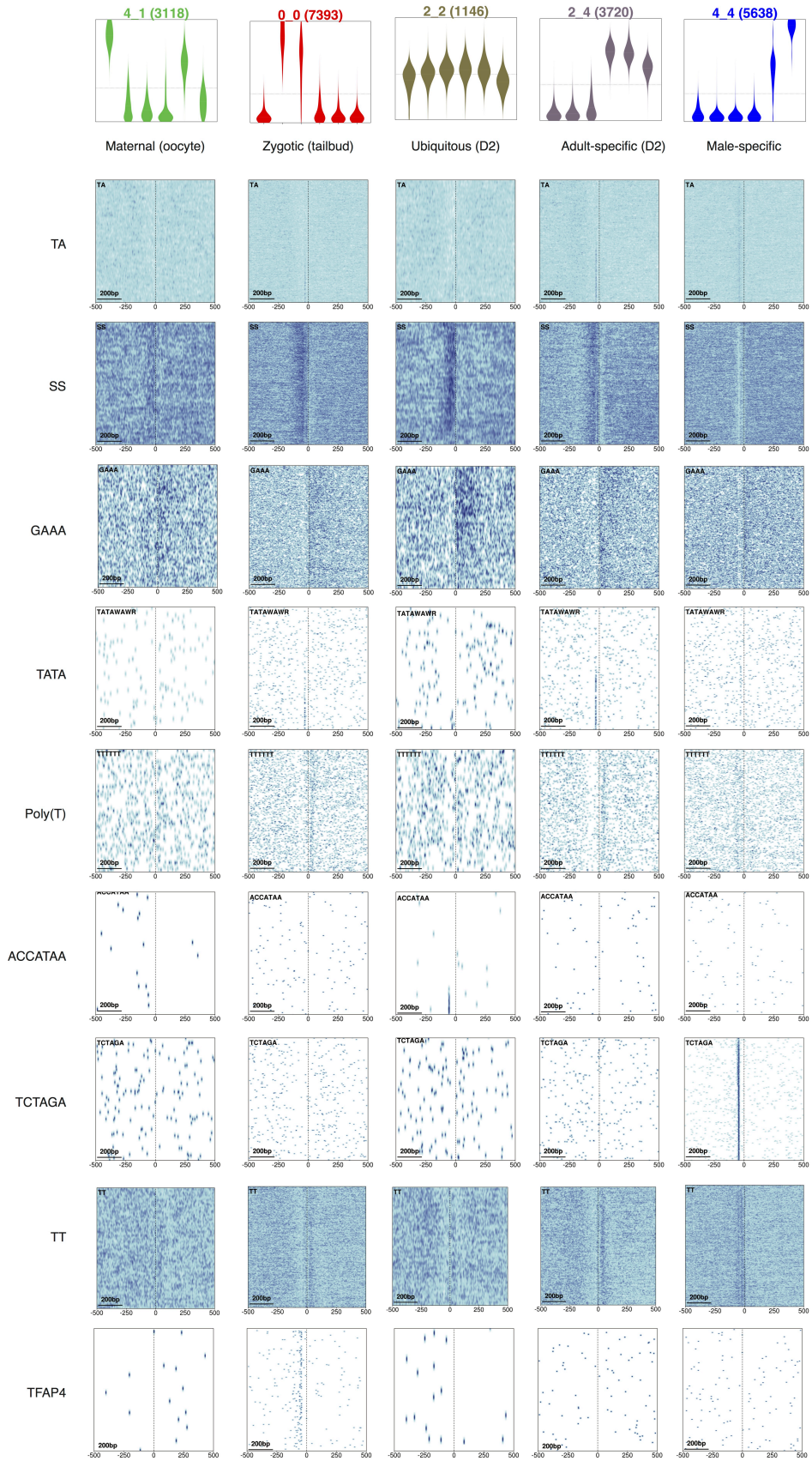
4_1 (3118)
0_0 (7393)
2_2 (1146)
2_4 (3720)
4_4 (5638)
Maternal (oocyte)
Zygotic (tailbud)
Ubiquitous (D2)
Adult-specific (D2)
Male-specific
TA
SS
GAAA
TATA
Poly(T)
ACCATAA
TCTAGA
TT
TFAP4
TATAWAWR
200bp
-500 -250 0 250 500

Figure S3.

Dinucleotide content and sequence motifs of *O. dioica* promoters. Each heatmap shows the density of the dinucleotide or motif, as indicated (left), at each position (x-axis) in the -500 to +500 bp region centred on the dominant TSS, for promoter sequences (rows) ordered by promoter width (top to bottom = broad to sharp). Darker blue indicates higher enrichment. Plots are shown for main expression clusters defined as indicated (top; expression profile for each also shown from in Figure 1C). Promoter sequences and TSS positions were taken from representative/dominant stages as indicated in parentheses (top).

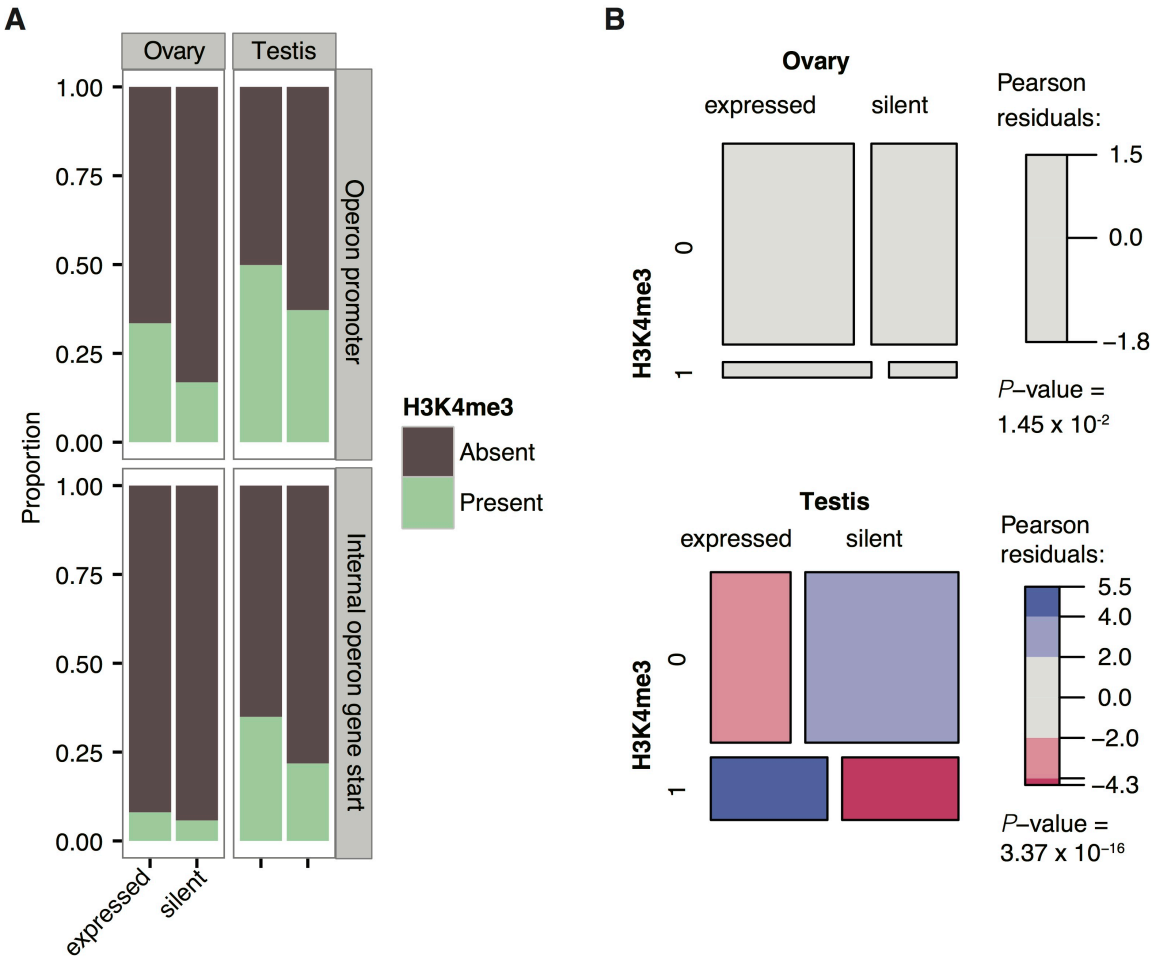

Figure S4.

Two modes of transcription for operon genes in the *O. dioica* ovary and testis. (A) Promoter regions of expressed operons are enriched for the active promoter mark H3K4me3 in both the ovary and the testis (upper panels). The start sites of downstream operon genes are also enriched for H3K4me3 in the testis only (lower panels), indicating cryptic male-specific internal promoters within operons. Each panel shows the proportion of expressed and silent genes/operons with a promoter region that overlaps an H3K4me3 ChIP-enriched region. **(B)** Mosaic plots visualizing the results of Pearson chi-squared tests for the association of H3K4me3 in internal operon gene promoter regions in the ovary (not significant) and testis (highly significant).

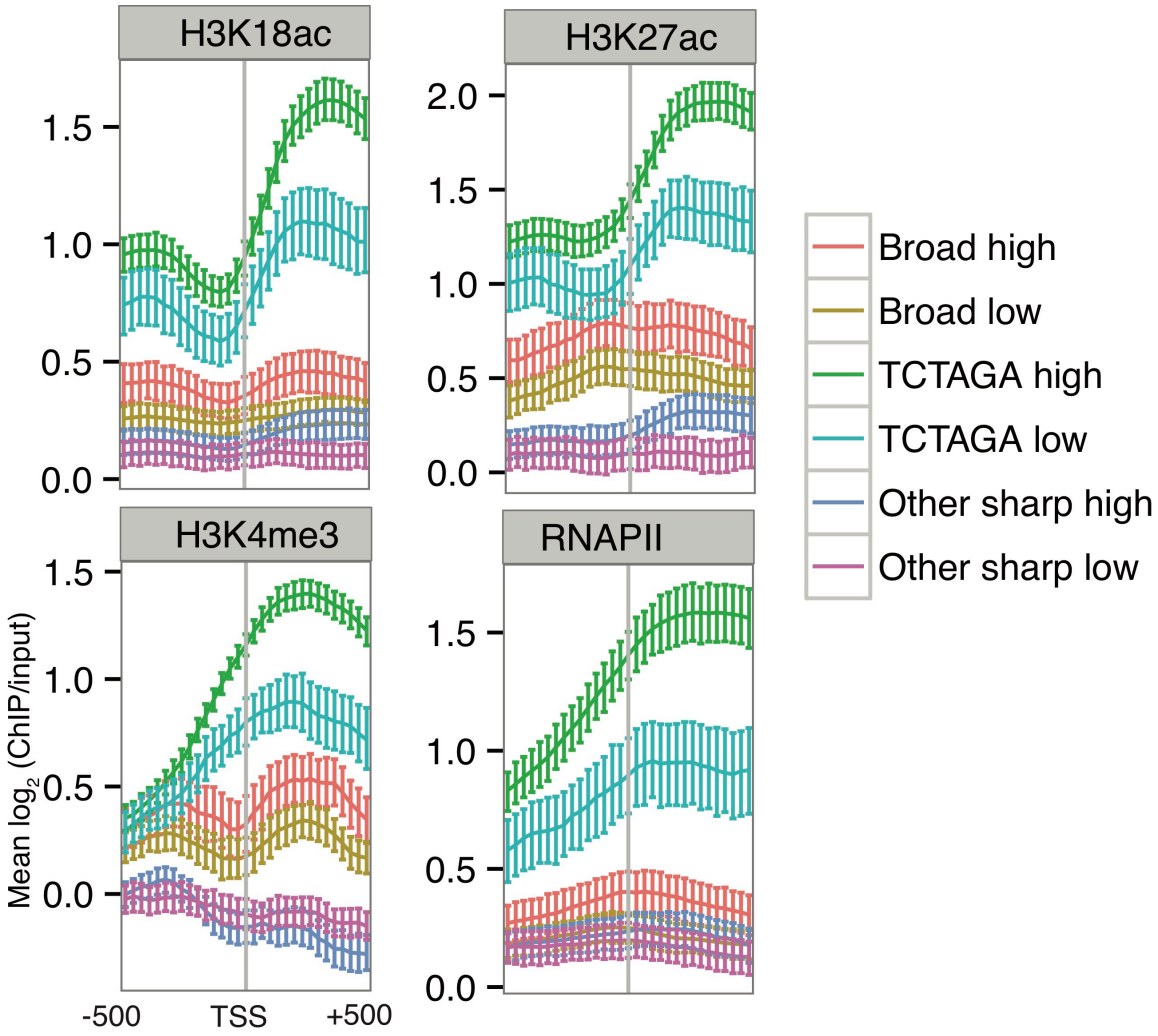

Figure S5.

Chromatin features of male-specific TCTAGA-promoters in the *O. dioica* testis. Data are shown for H3K18ac, H3K27ac, H3K4me3 and RNAPII ChIP-chip experiments. Each plot shows the mean $\log_2$ ratio of ChIP/input at each probe position in a 1000 bp window centered on the dominant CAGE TSS. Promoters were categorized according to promoter type (sharp=narrow region of TSSs; broad=dispersed region of TSSs) and activity (high indicates a tpm ≥ median; low indicates < median tpm). Sharp promoters were also subdivided according to the presence of a TCTAGA motif. Error bars show 95% confidence intervals for the mean obtained by bootstrapping. Regardless of expression level, TCTAGA-promoters have a higher enrichment for all features shown compared to promoters without this male-specific motif.

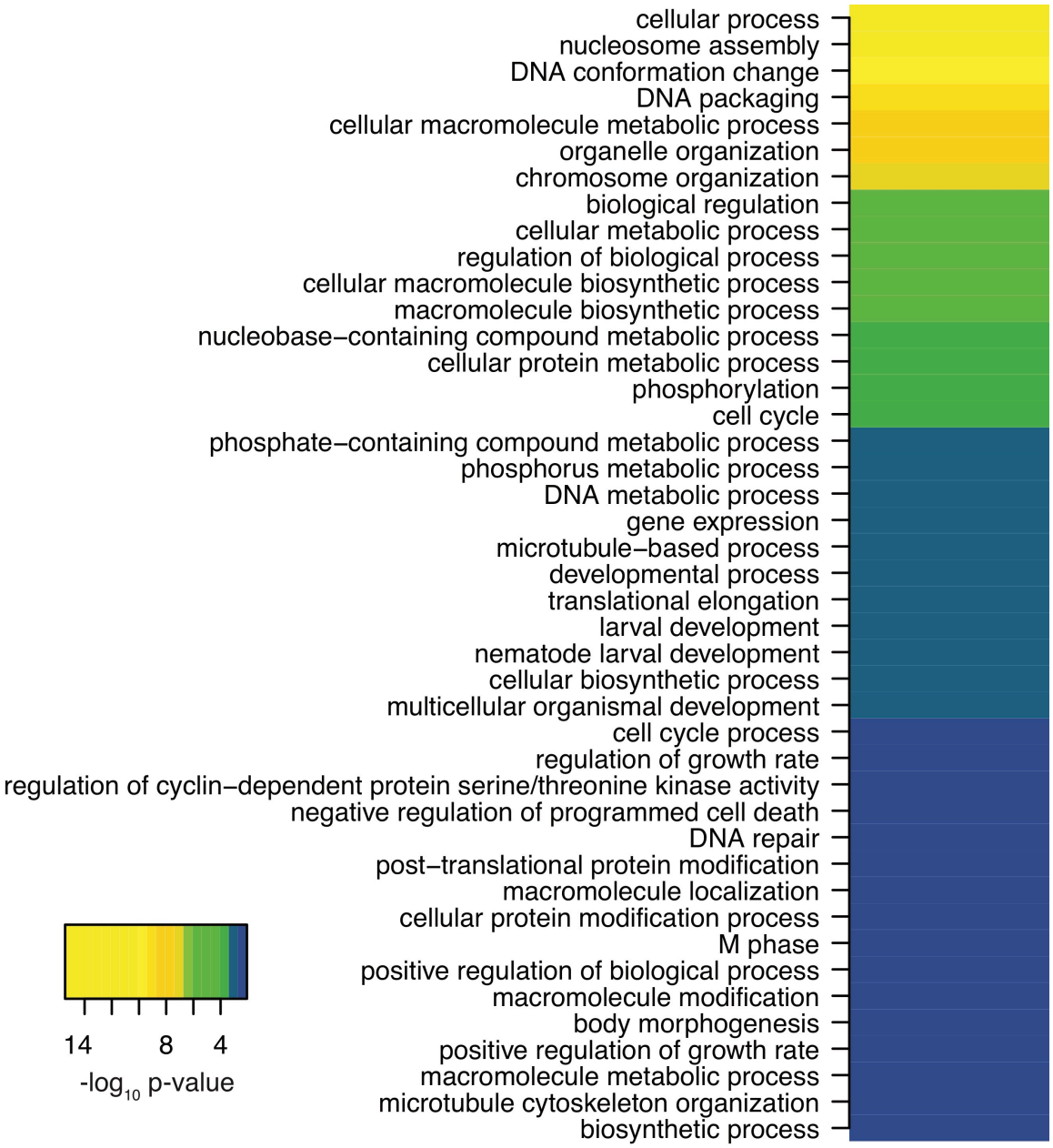

Figure S6.

Functions of E2F1-regulated genes. GO-terms enriched in genes associated with E2F1-bound broad promoters in the ovary, sorted and coloured according to respective *P*-values.

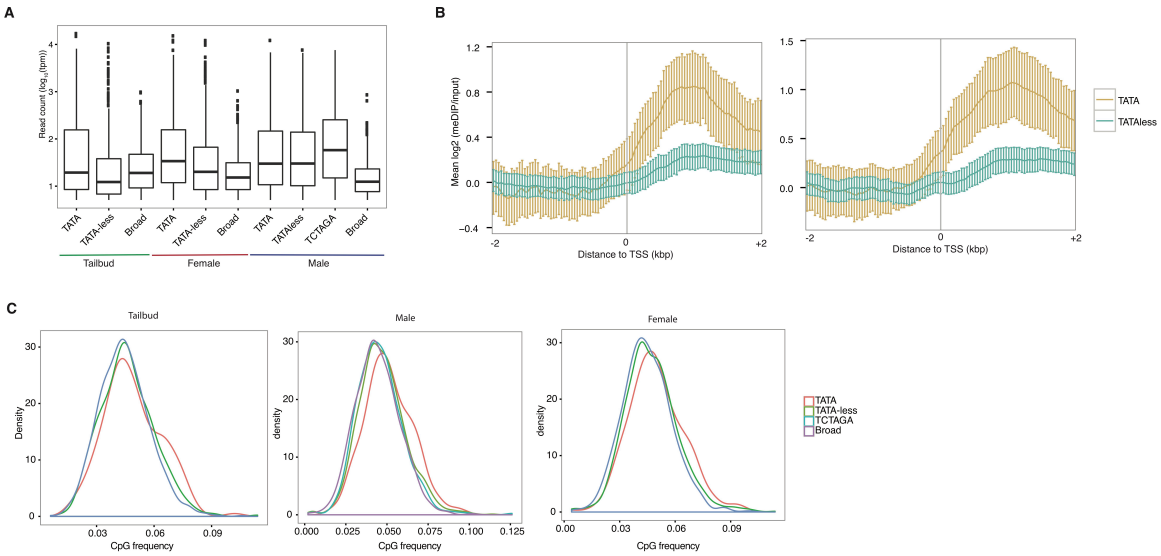

Figure S7.

DNA methylation, expression levels and downstream CpG content of TATA and TATA-less promoters. DNA-methylation enrichment downstream of TATA-dependent promoters is not explained by differing expression levels (**A**) and continues further downstream (**B**) but does coincide with a higher CpG content (**C**). Data in (**A**) and (**C**) correspond to the subset of promoters used in Figure 4. **(A)** Boxplot shows the distribution of expression levels ($\log_{10}$ tpm of dominant TSS) for broad promoters as well as sharp promoters with and without a TATA-element in the embryo (tailbud), male and female day 6 animals, and for promoters in the male with an upstream TCTAGA motif instead of a TATA-element. TATA-dependent and TATA-less promoters in the male did not have a significant difference in mean expression levels (Wilcoxon rank sum test: W = 142539.5, *P*-value = 0.9812). **(B)** Plots show the mean $\log_2$ ratio of methyl-DNA IP/input (y-axis) at each probe position (x-axis) in a 4 kb window centred on the dominant TSS, in two independent meDIP-chip biological replicates using *O. dioica* testes. Error bars show 95% confidence intervals for the mean obtained by bootstrapping. Each plot shows sharp promoters divided into those with a TATAA-element and those without. **(C)** Density plots show the distribution of CpG frequencies in the 500 bp region downstream of the dominant TSS for each subset of promoters in the embryo (tailbud), male and female animals.

Sample	Number of reads	% uniquely mapping (million reads)	% multiple mapping	% failed to align	Number of tag clusters	% assigned to annotated genes
Female D6	3,886,540	60.78 (2.4)	12.00	27.22	12945	76.27
Male D6	9,238,422	63.58 (5.9)	14.54	21.88	13562	77.74
Oocyte	4,276,690	58.87 (2.5)	10.68	30.46	6455	78.87
Tailbud	5,450,443	58.52 (4.5)	12.83	28.64	11433	70.78
Tadpole	8,244,476	63.76 (5.3)	11.33	24.91	13177	71.64
D2	7,625,608	53.95 (2.9)	14.56	31.48	12647	71.84

Table S1.
CAGE read mapping at different developmental stages.