


RESEARCH ARTICLE

Open Access



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

Gemma B. Danks^{1*†} , Pavla Navratilova^{1†}, Boris Lenhard^{1,2,3} and Eric M. Thompson^{1,4*}

Abstract

Background: Development is largely driven by transitions between transcriptional programs. The initiation of transcription at appropriate sites in the genome is a key component of this and yet few rules governing selection are known. Here, we used cap analysis of gene expression (CAGE) to generate bp-resolution maps of transcription start sites (TSSs) across the genome of *Oikopleura dioica*, a member of the closest living relatives to vertebrates.

Results: Our TSS maps revealed promoter features in common with vertebrates, as well as striking differences, and uncovered key roles for core promoter elements in the regulation of development. During spermatogenesis there is a genome-wide shift in mode of transcription initiation characterized by a novel core promoter element. This element was associated with > 70% of male-specific transcription, including the use of cryptic internal promoters within operons. In many cases this led to the exclusion of *trans*-splice sites, revealing a novel mechanism for regulating which mRNAs receive the spliced leader. Binding of the cell cycle regulator, E2F1, is enriched at the TSS of maternal genes in endocycling nurse nuclei. In addition, maternal promoters lack the TATA-like element found in zebrafish and have broad, rather than sharp, architectures with ordered nucleosomes. Promoters of ribosomal protein genes lack the highly conserved TCT initiator. We also report an association between DNA methylation on transcribed gene bodies and the TATA-box.

Conclusions: Our results reveal that distinct functional promoter classes and overlapping promoter codes are present in protochordates like in vertebrates, but show extraordinary lineage-specific innovations. Furthermore, we uncover a genome-wide, developmental stage-specific shift in the mode of TSS selection. Our results provide a rich resource for the study of promoter structure and evolution in Metazoa.

Keywords: Core promoter, DNA methylation, Histone modification, MZT, Oogenesis, Operons, Spermatogenesis, TATA-box, Transcription initiation

Background

Sites for the initiation of transcription are frequently marked in the genome by specific sequence elements, which are recognized and subsequently bound by basal transcription factors [1, 2]. The diversity of core promoter elements suggests that they play important roles in the differential regulation of subsets of genes. For example, the conserved TATA-box, which is bound by TATA-binding

protein, is responsible for transcription initiation at tissue-specific promoters in mouse [3], whereas a degenerate TATA-like element is associated with maternal transcription initiation in zebrafish [4]. Other core elements may be critical to development, but as yet none has been assigned a specific role(s).

In the promoters of vertebrates 5-methylcytosine (5mC) DNA methylation is associated with transcriptional repression. When DNA is methylated within gene bodies, however, it is instead associated with active transcription and splicing [5, 6], a feature that is conserved between animals and plants (although it may have been lost in certain lineages including *Caenorhabditis elegans* and

*Correspondence: gemma.danks@uib.no; eric.thompson@uib.no

†Equal contributors

¹Sars International Centre for Marine Molecular Biology, University of Bergen, N-5006 Bergen, Norway

⁴Department of Biology, University of Bergen, N-5006 Bergen, Norway
Full list of author information is available at the end of the article

Drosophila) [5], and functions in the repression of alternative intragenic promoters [7, 8]. The majority of DNA methylation in the genome of the urochordate, ascidian, *Ciona intestinalis* [9, 10] is gene body DNA methylation and is found only at a subset of genes, where it is positively correlated with gene expression level [9, 10]. How this subset is selected for methylation, and whether core promoter elements play a role in this, has so far remained unknown.

The identification of core promoter elements, and mapping of transcription start sites (TSSs), at single-nucleotide resolution, has been facilitated by Cap Analysis of Gene Expression (CAGE) [11]. This has led to the discovery of two main modes for specifying sites of transcription initiation [2, 12]. Sequence motifs bound by the pre-initiation complex result in transcription initiation within a narrow region and lead to “sharp” promoter architectures. Conversely, the positioning of nucleosomes defines a wider catchment area for the pre-initiation complex and leads to “broad” promoter architectures [4]. Promoter architectures can also show associations to downstream translational events. For example, promoters of ribosomal protein genes are usually sharp with a highly conserved TCT Initiator (Inr) sequence [2, 13, 14], which forms the beginning of a Terminal OligoPyrimidine (TOP) motif critical for nutrient-dependent translational control [15]. In mammals, these promoters, unusually, have both a TATA-box and CpG islands. In *C. intestinalis* they are sharp with a TCT initiator, but lack a TATA-box [13]. Recently, it has been shown that a genome-wide switch occurs in the mode of TSS selection during zebrafish embryogenesis [4]. Maternal promoters in zebrafish are sharp, or multiple sharp, with TATA-like, AT-rich (W-box) upstream elements guiding TSS selection. During the maternal to zygotic transition, nucleosomes with H3K4me3 are positioned at zygotic promoters that lack a W-box, leading to broad promoter architectures. The extent to which these, or similar, features are evolutionarily conserved is unknown.

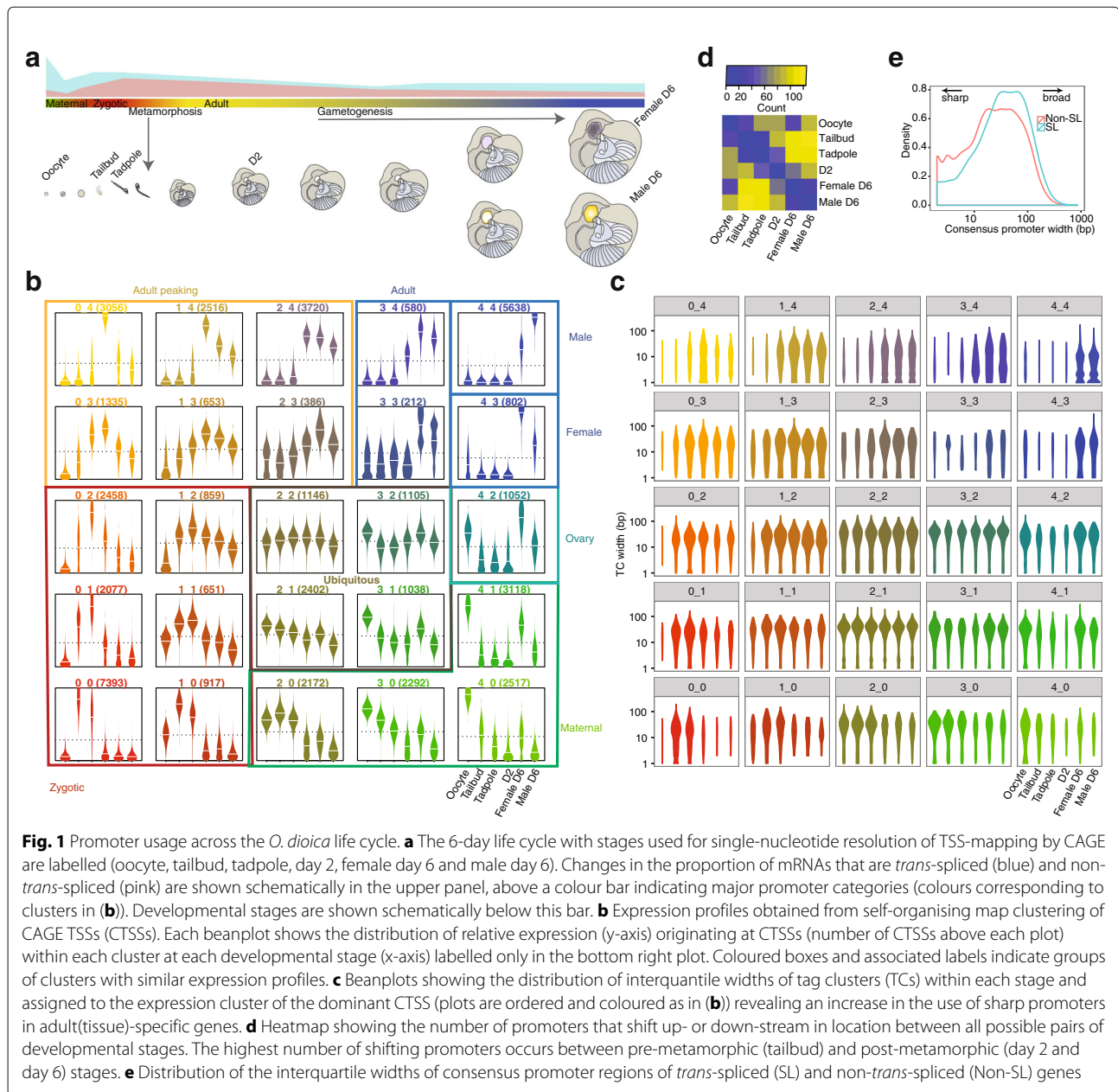
Oikopleura dioica is a marine, larvacean, chordate in the sister group to vertebrates and is well positioned to examine the evolution of TSS features and the dynamics of TSS selection. The *O. dioica* genome is the most compact of any animal genome sequenced so far and 27.8% of its genes are organised into operons [16]. Each operon contains two or more genes that are transcribed from a single promoter located upstream of the first gene. The resulting polycistronic mRNA is resolved via the *trans*-splicing [17, 18] of a spliced-leader (SL) sequence to unpaired acceptor sites at the 5' ends of each resulting monocistron. *Trans*-splicing in *O. dioica* [19] also occurs at monocistronic genes; 39% of all annotated genes give rise to mRNAs that are *trans*-spliced [20]. During *trans*-splicing a portion of the original 5' sequence upstream of the

trans-splicing acceptor site is removed. Here, we mapped TSSs at single-nucleotide resolution, using CAGE, in six key stages of *O. dioica* development, covering the entire 6-day life cycle. In order to maximise the mapping of original TSSs (rather than *trans*-splice sites) we sequenced only mRNAs without the SL sequence. We used our TSS maps, together with previously generated genome-wide maps of *trans*-splice sites [20], E2F1 binding sites, key histone modifications and DNA methylation [21], to derive TSS-selection criteria at major developmental transitions and identify novel modes of regulation. Our data show that *O. dioica* has some promoter features in common with vertebrates, including evidence of nucleosome positioning at broad promoters and tissue-specific expression of TATA-dependent promoters, but it differs markedly in its mode of maternal transcription initiation, which is characterized by the ordering of nucleosomes and the binding of the cell cycle regulator E2F1. *O. dioica* also employs a remarkable genome-wide shift in mode of TSS-selection during spermatogenesis, associated with a distinct, tissue-specific, TCTAGA core promoter motif, that has not been previously identified.

Results

Promoter usage across development

We extracted RNA for CAGE from *O. dioica* at six stages of development across the 6-day life cycle of the animal (Fig. 1a). Illumina sequencing generated > 39 M reads, of which 2.4–5.9 M (54–64%) for each stage mapped uniquely to the genome (Additional file 1: Table S1). Summing tags that mapped to unique positions gave the abundance of transcripts originating from each TSS. We normalized these counts to tags per million reads (tpm) and clustered neighbouring TSSs (allowing up to 20 bp between TSSs) to generate tag clusters (TCs), which revealed the set of promoter regions that are active within each stage. TCs (supported by at least 1 tpm in at least one stage) mapped to 6241 annotated genes, 4,937 of which were defined as expressed using previously generated tiling array data [22] across equivalent developmental stages (Additional file 1: Figure S1). Multiple genes within an operon are transcribed from a common TSS. In line with this we captured TCs for only 538 downstream operon genes, out of 2832 (19%) that were defined as expressed based on tiling array data (Additional file 1: Figure S1). TCs for these 538 genes include previously unidentified stage-specific use of cryptic internal promoters within operons. Previously, we generated a bp-resolution, genome-wide map of *trans*-splice sites in *O. dioica* [20] using pooled animals collected at the same developmental stages we used here. As previously, we define a gene as *trans*-spliced if it is associated with a mapped *trans*-splice site. Our newly generated CAGE dataset captures the original TSSs



of 51% (1341/2643) of all monocistronic (non-operon) *trans*-spliced genes allowing us to analyse promoter features of *trans*-spliced genes. Since *trans*-splicing is thought to occur co-transcriptionally, and some monocistronic genes can be both *trans*-spliced and non-*trans*-spliced depending on the developmental stage, we did not expect to capture a larger proportion of the promoters of all potentially *trans*-spliced monocistronic genes.

We defined 13,771 consensus promoter regions in the genome by clustering TCs, with > 5 tpm, across stages [4, 23]. Expression profiles of individual TSSs were clustered using a self-organizing map [4, 23] (SOM) in order to assess the dynamics of TSS selection across development

(Fig. 1b). Distinct ubiquitous, maternal and zygotic expression TSS clusters were present as well as a large cluster of male-specific TSSs. SOM clustering of consensus promoter region expression profiles revealed similar patterns (Additional file 1: Figure S2).

A genome-wide shift in mode of TSS selection during spermatogenesis

Maternal and ubiquitously expressed TSSs (identified by SOM clustering; Fig. 1b), and TSSs associated with *trans*-spliced genes, were found predominantly within broad TCs (Fig. 1c,e and Additional file 1: Figure S2) whereas TSSs used specifically in adult stages, particularly male-

specific TSSs, were predominantly found in sharp TCs (median width of male-specific promoters in day 6 male was 4 bp compared to 25 bp in maternal promoters in oocytes and 23 bp in ubiquitous promoters in day 2 animals); Fig. 1c, Additional file 1: Figure S2). The presence of sharp TCs suggests sequence motifs in these core promoters determine the selection of TSSs at a fixed distance downstream. We therefore examined all promoter sequences and identified a core promoter element (TCTAGA), embedded in a TT-rich sequence context, which was remarkably specific to male-specific TSSs (Fig. 2; see also Additional file 1: Figure S3 for frequencies of other motifs and dinucleotides, across the genome and around different classes of promoters). This element was present in 71.6% (1391/1943) of male-specific TCs in the male and was strictly positioned 40–50 nt upstream of the dominant TSS (with a strong preference for 45–48 bp; Fig. 2b). Given that the majority of the animal's mass at this stage is found in the gonad our data strongly suggests the use of a unique mode of gene regulation that is linked to spermatogenesis. Indeed, when we examined existing array data from dissected testes, ovaries and trunks of day 6 animals we found that 369/502 (73.5%) of genes that are specifically expressed in the testis (and represented in our

CAGE data set) are associated with a TCTAGA promoter element, compared to 100/906 (11.0%) that are specific to the ovary and 7/275 (2.5%) that are specific to the trunk (Additional file 1: Figure S4).

In order to determine whether or not this mode of regulation is found in mammals we re-analyzed existing CAGE data [23, 24] from a time course of 8 testis samples across mouse development from embryogenesis to adult tissues. We found no enrichment for a position-specific TCTAGA motif in promoter regions of any stage, nor of promoters with spermatogenesis-associated expression patterns (data not shown). In order to determine if this mode of regulation is present in other urochordate genomes we searched the promoter regions of 16,671 annotated genes in the *C. intestinalis* genome and found only 226 (1.3%) with a TCTAGA within 100 bp upstream of the annotated start site compared to 2088 (12.5%) that had a consensus TATA-box motif. This suggests a larvacean, lineage-specific evolution of this mode of TSS-selection for the activation of the spermatogenesis transcriptional program.

A single gene may have several alternative TSSs selected at different developmental stages. We identified all promoter regions with a shift in TSS usage [4, 23] between

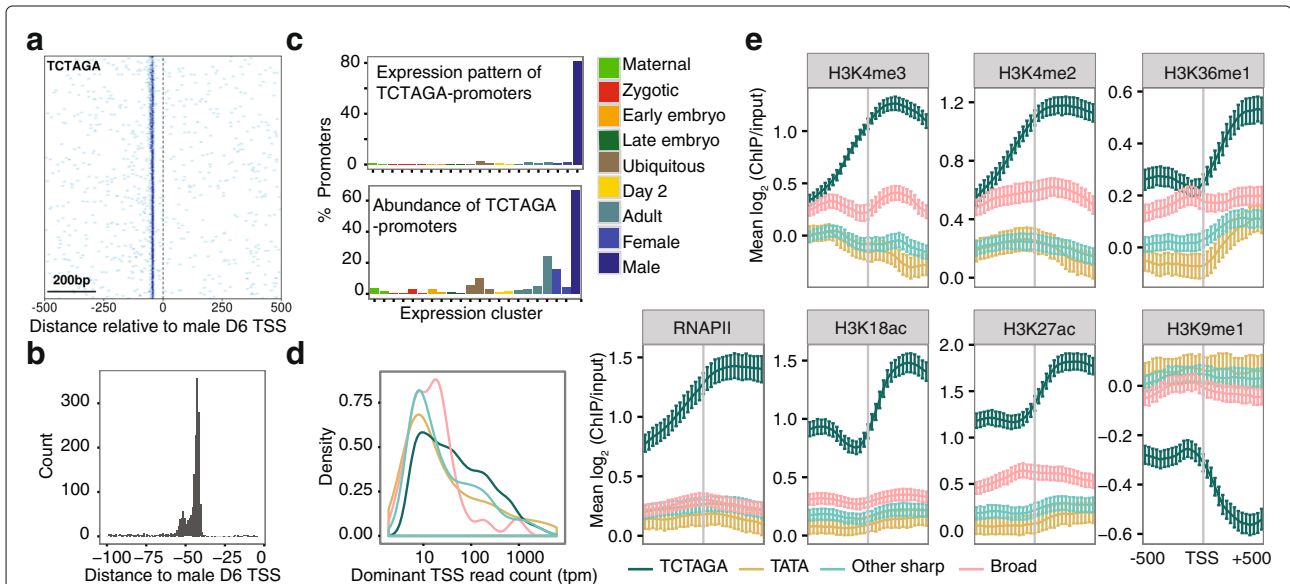


Fig. 2 Features of spermatogenic promoters in *O. dioica*. **a** Heatmap shows the density of TCTAGA at each position (x-axis) in a -500 to +500 bp region centred on the male dominant TSS for spermatogenic promoter sequences (rows) ordered by promoter width (top to bottom = broad to sharp). Darker blue indicates higher enrichment. **b** Distribution of distances (median = 44 bp) to the TCTAGA motif relative to the dominant TSS in male day 6 animals. **c** The TCTAGA motif is specific to transcription in the adult male (final bar in each plot = male-specific (4_4) cluster): of all promoters with a TCTAGA motif the majority have a male-specific expression profile (top). Moreover, the majority of spermatogenic transcription is associated with a TCTAGA motif: of all promoters with a male-specific expression profile the majority contain a TCTAGA, whereas very few promoters in other expression classes contain this motif (bottom). **d** Distribution of tag counts (tpm) for different promoter classes (see colour key code in **e**) in the male: TATA = TSS with upstream TATAW element; TCTAGA = TSS with upstream TCTAGA; Sharp = all other promoters with narrow region of TSSs; Broad = dispersed region of TSSs. **e** ChIP-chip data for day 6 testes are shown for RNAPII, H3K18ac, H3K27ac, H3K4me3, H3K4me2, H3K36me1 and H3K9me1. Each plot shows the mean log₂ ratio of ChIP/input at each probe position in a 1000 bp window centered on the dominant TSS. Promoters were categorized according to promoter type as in **d**. Error bars show 95% confidence intervals for the mean obtained by bootstrapping

any pair of developmental stages. We observed the highest number (124/519) of single promoter regions for which the TSS location changed (< 40% TC overlap) when comparing the embryo versus adult male stages (Fig. 1d; Additional file 1: Figure S5). In 43/124 of these cases there was a shift from promoters lacking a TCTAGA in the embryo to a TCTAGA-associated promoter in the male. This suggests that the TCTAGA promoter element may play a role in selecting alternative TSSs during spermatogenesis.

The *O. dioica* genome contains 1765 operons that comprise multiple genes that are transcribed from a single promoter into polycistronic mRNA. Cryptic internal promoters within operons, which drive tissue-specific expression of downstream genes, have been described in *C. elegans* [25] but the prevalence of these in the *O. dioica* genome is unknown. We identified 693 internal promoters within operons in *O. dioica*: male-specific promoters (208) were over-represented and the TCTAGA element was found more frequently (25.5%; 177 promoters) than expected ($\chi^2 = 142.98$, $df = 1$, $p < 2.2 \times 10^{-16}$). This suggests that during spermatogenesis the TCTAGA element plays a role in selecting TSSs at internal promoters within operons that are otherwise transcribed from a single upstream promoter in other stages. We therefore analyzed patterns of enrichment of the H3K4me3 promoter mark from ChIP-chip data [21] in the ovary and testis of animals at the same developmental stage as the day 6 male and female animals used to generate our CAGE dataset. In support of the presence of male-specific cryptic internal promoters within operons, we only found enrichment of H3K4me3 at the start sites of internal genes within operons in the testis, whereas the start sites of operons were enriched for this mark in both the ovary and testis (Additional file 1: Figure S6).

Sites for *trans*-splicing are determined by the presence of an unpaired AG acceptor site, which is usually followed by an adenine [20]. Remarkably, we found that 89 male-specific promoters in males (associated with 87 genes) had a TCTAGA motif with its AGA mapping to a *trans*-splice site (representing 16.4% of all TCTAGA male-specific genes that were annotated as *trans*-spliced). Transcription downstream of these TCTAGA elements during spermatogenesis therefore results in mRNAs that lack a *trans*-splice acceptor site and are therefore not *trans*-spliced with the SL sequence. Transcription driven by alternative upstream promoters during other stages of development leads to mRNAs with the *trans*-splice site intact and are therefore *trans*-spliced with the SL. This finding reveals a novel mechanism for the developmental regulation of *trans*-splicing.

Male-specific TCTAGA promoters had significantly higher expression levels compared to other promoter types in males (all $p < 0.05$; Fig. 2d). We analyzed the

profiles of a range of histone modifications as well as RNA pol II occupancy using ChIP-chip in the testis and ovaries of day 6 stage-matched animals [21]. We found that male-specific TCTAGA promoters were associated with higher RNA pol II occupancy and higher enrichment of histone modifications associated with active transcription (and depletion of repressive marks) in the testis, including specific marking by H3K18ac (Fig. 2e). Several of these marks were independent of expression level (Additional file 1: Figure S7). Together, our data revealed a unique transcription initiation code that was specific to male-specific core promoters. This code is associated with a chromatin state primed for high levels of transcription in the testis and directs both a genome-wide shift in promoter usage, and the developmental regulation of operon transcription and *trans*-splicing.

Maternal modes of TSS selection in endocycling nurse nuclei

Maternal promoters in zebrafish tend to be sharp, or multiple sharp, with a degenerate TATA-like motif (W-box) determining TSSs [4]. In contrast, we found that maternal promoters in *O. dioica* were broad (Fig. 1c) and lacked a W-box at the expected TATA-box position or any other enrichment of dinucleotides (Additional file 1: Figure S3). Broad promoters in zebrafish are associated with ordered nucleosomes, as shown by the precise positioning of histone H3K4me3 enrichment at the first nucleosome downstream from the dominant TSS [4]. Here, we used ChIP-chip data [21] from the ovaries of day 6 (stage-matched) *O. dioica* and analyzed the profiles of H3 and H3K4 histone modifications around dominant TSSs of maternal promoters. Distinct peaks of histone H3 enrichment flanked the dominant TSSs at broad promoters in the ovary (Fig. 3a) with a peak in H3K4me3 enrichment immediately downstream (Fig. 3a) as seen in vertebrate broad promoters. These data show that TSS-selection in *O. dioica* broad promoters has similar features to those in vertebrate broad promoters, indicating that this may be the main mode of TSS selection in (predominantly broad) maternal promoters in *O. dioica*.

We also found that the nucleosome-free region at the TSS of broad promoters in the day 6 female corresponds to an enrichment of the activating transcription factor E2F1 (Fig. 3a), a key regulator of the cell cycle [26]: 27.7% (1075/3882) of genes with strong CAGE support (≥ 5 tpm) had promoters bound by E2F1 in the ovary. These genes were enriched for, though not limited to, known E2F1-regulated functions (Additional File 1: Figure S8). These results suggest that E2F1 has a role in regulating maternal transcription in *O. dioica*.

Maternal promoters in *O. dioica* were located on the X-chromosome more frequently than expected, compared

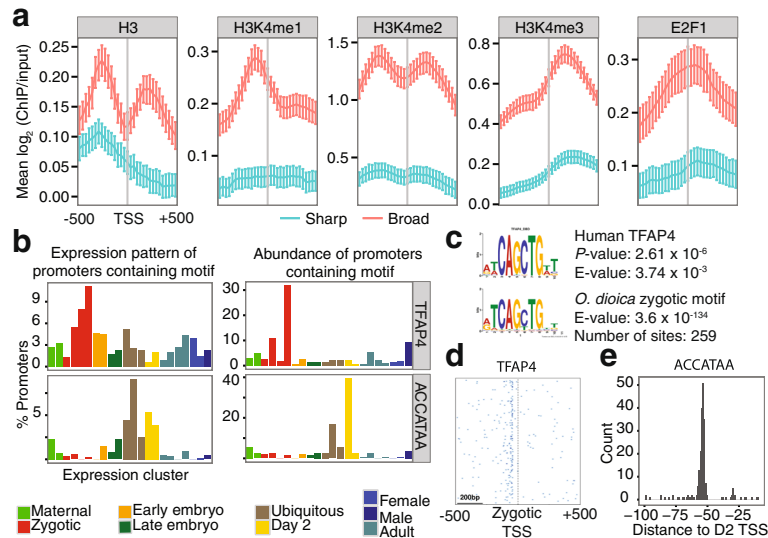


Fig. 3 Features of maternal, zygotic and ubiquitous promoters in *O. dioica*. **a** Ordered nucleosome positioning at broad promoters in the ovary. Data are shown for H3, H3K4me1, H3K4me2, H3K4me3 and E2F1 ChIP-chip experiments. Each plot shows the mean log₂ ratio of ChIP/input at each probe position in a 1000 bp window centred on the dominant TSS. Promoters were categorized into sharp (blue) and broad (pink) using lower and upper quartiles of widths across all stages. Error bars show 95% confidence intervals for the mean obtained by bootstrapping. **b** Percentage of all consensus promoters (left) within each expression cluster (x-axis) (profiles shown in Additional file 1: Figure S2A) containing sequence elements as labelled (*O. dioica* TFAP4-like motif, ACCATAA motif). Percentage of promoters that contain each motif falling within each expression cluster is also shown (right). Expression clusters are grouped and coloured as coded in Fig. 1b. **c** Sequence logo for an over-represented motif (E-value and number of sites as indicated) in zygotic promoters (tailbud sequences from TCs with dominant CTSS in cluster 0_0) in *O. dioica* (top), and its alignment to a significant match (E-value and *p* as indicated) for the binding site motif of human TFAP4. **d** Heatmap shows the density of the zygotic promoter motif matching TFAP4 at each position (x-axis) in a -500 to +500 bp region centred on the tailbud dominant TSS for zygotic promoter sequences (rows) ordered by promoter width (top to bottom = broad to sharp). Darker blue indicates higher enrichment. **e** Distance relative to the dominant TSS in day 2 animals for the ACCATAA motif found in sharp promoters that have ubiquitous and day 2-specific expression profiles

to zygotic promoters ($\chi^2 = 43.34$, $df = 1$, $p = 4.61 \times 10^{-11}$), revealing a female-bias of X-linked genes [27] in *O. dioica*.

Regulation of zygotic promoters

Zygotic promoters in *O. dioica* (TSS clusters with low maternal and high embryonic expression; Fig. 1b) contained an upstream GC-rich region, characteristic of broad promoters, and a downstream poly(T)-tract (Additional file 1: Figure S3). An E-box [28] motif with a significant match to the binding site of TFAP4 (activating enhancer binding protein 4), a regulator of cell proliferation, was over-represented in the region immediately upstream of 259 zygotic-specific TSSs in the embryonic tailbud stage (Fig. 3b-d; Additional file 1: Figure S9). Genes associated with these TSSs were enriched for GO terms related to muscle development (Additional file 1: Figure S10).

TSS selection in ubiquitous and ribosomal protein gene promoters

Most *O. dioica* promoters used to drive ubiquitous expression throughout the life cycle (Fig. 1b) had a broad architecture with a strong GC-rich band immediately

upstream of the TSS, as seen in zygotic promoters, and a clear GAAA signal at the expected +1 nucleosome position (Additional file 1: Figure S3). We also found a position-specific (median distance 56 bp upstream; Fig. 3e) ACCATAA sequence element associated with TSS-selection in sharp ubiquitous promoter regions (Fig. 3b and Additional file 1: Figure S3), as well as in sharp promoters specific to day 2 animals (juvenile animals; pre-gametogenesis). This motif was present in 215 consensus promoter regions.

Whereas a typical Initiator (Inr) CA dinucleotide was present in 53% of consensus promoter regions in *O. dioica*, the TCT initiator, which is highly conserved at ribosomal protein genes in other species, including *C. intestinalis*, was absent from all CAGE-detected ribosomal protein genes in *O. dioica* (29 detected out of 129 annotated; the majority being located within operons [20]). Unlike the sharp promoters of these genes in other species, TCs of these ribosomal protein genes in *O. dioica* were predominantly broad (only 6/51 were sharp; 2/6 contained a TATA-element), in line with other *trans*-spliced gene promoters in this animal (Fig. 1e), and we found a higher average CpG content than non-ribosomal protein genes (Welch Two Sample t-test: $t = 3.22$, $df = 35.164$,

$p = 1.379 \times 10^{-3}$). This indicates that these promoters have lost the specific transcriptional regulation conferred by the TCT initiator in other species and provides further evidence that the *trans*-spliced SL replaces the role of the TOP motif [20], which starts at this initiator sequence.

Conserved tissue-specific TATA-dependent TSS-selection is associated with higher levels of DNA methylation in gene bodies

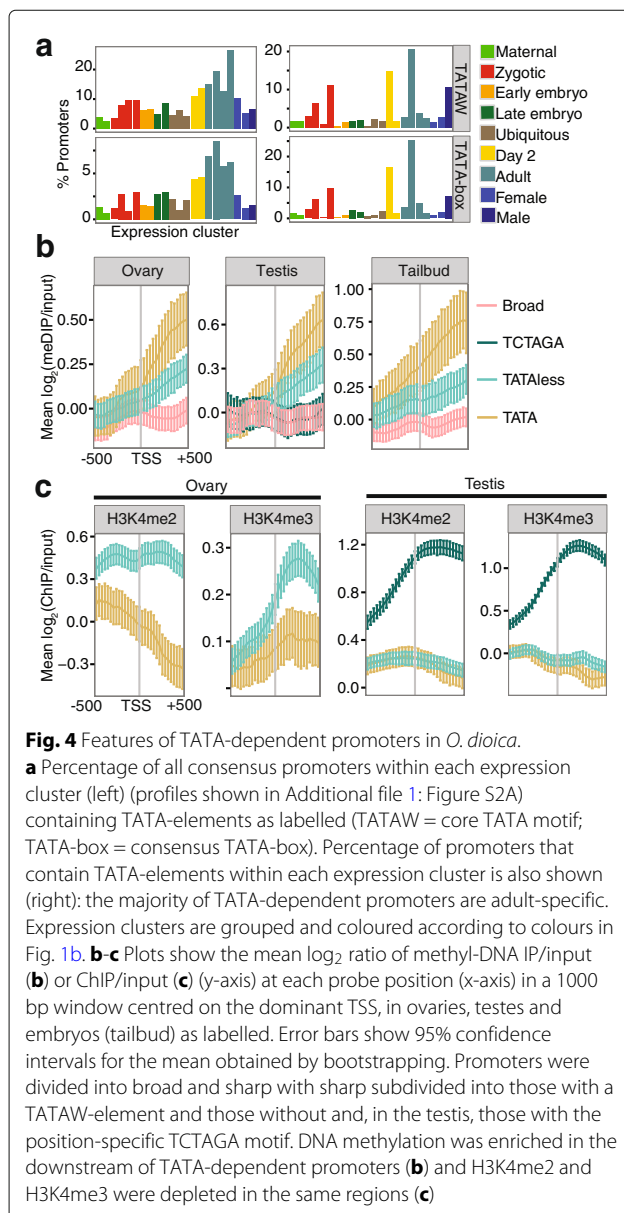
We next searched all our CAGE-defined promoters for the ancient TATA-box promoter element in order to assess the usage of this promoter motif in *O. dioica* development. A TATAW element was present in 10.7% of

consensus promoter regions, in line with the percentages of TATA-dependent promoters in mammals [29]. A lower percentage (3.8%) of promoters had a longer consensus TATA-box motif (TATAWAWR). The use of promoters with this consensus motif was specific to sharp promoters in adult stages (Fig. 4a and Additional file 1: Figure S3), indicating that this mode of TSS-selection at tissue-specific promoters is conserved between *O. dioica* and vertebrates. As in other species, the preferred location of the TATA box was 28–31 bp upstream.

We then analysed profiles of methylated DNA enrichment around promoters using methylated DNA immunoprecipitation followed by chip) data [21] from ovaries, testes and tailbud stage embryos, in order to uncover any associations with core promoter elements. Interestingly, we found that TATA-dependent sharp promoters had a higher average enrichment of DNA methylation in downstream gene bodies than TATA-less sharp promoters in embryos, ovaries and testes (Fig. 4b). This trend was not explained by expression level (Additional file 1: Figure S11A) or proximity to the promoter (Additional file 1: Figure S11B) but did correspond to a higher frequency of CpGs (Additional file 1: Figure S11C). A regression analysis showed that despite accounting for expression level ($B = 0.06$, $p = 2.04 \times 10^{-8}$), promoter width ($B = -0.04$, $p = 2.47 \times 10^{-4}$) and downstream CpG content ($B = 0.26$, $p < 2 \times 10^{-16}$) the presence of the most common core TATAA motif was a significant, independent, positive predictor ($B = 0.27$, $p = 1.32 \times 10^{-12}$) of downstream DNA-methylation levels in ovary, testis and tailbud, (the stage of development was not a significant predictor, overall fit of the model, $R^2 = 0.08$). H3K4me3, which inhibits the interaction of DNA methyltransferases with histone proteins [30], was depleted (as was H3K4me2) at the TSS and in the downstream regions of TATA-dependent promoters, compared to TATA-less sharp promoters, in both the ovary and testis (Fig. 4c). Together our findings reveal a specific association of gene body DNA methylation, and H3K4me3 depletion, with a TATA-dependent mode of TSS selection in *O. dioica*. We found a similar association in zebrafish, although the increase in DNA methylation compared to TATA-independent promoters was at the TSS rather than the gene body (Additional file 1: Results and Figure S12).

Discussion and conclusions

Here, we mapped sites of transcription initiation genome-wide at single nucleotide resolution across the life cycle of a marine chordate belonging to the sister group to vertebrates. Our data revealed a suite of TSS-selection criteria in *O. dioica* (Fig. 5) with features that are both shared with vertebrates and markedly different, particularly among maternal and spermatogenesis promoters (Additional file 1: Table S2).



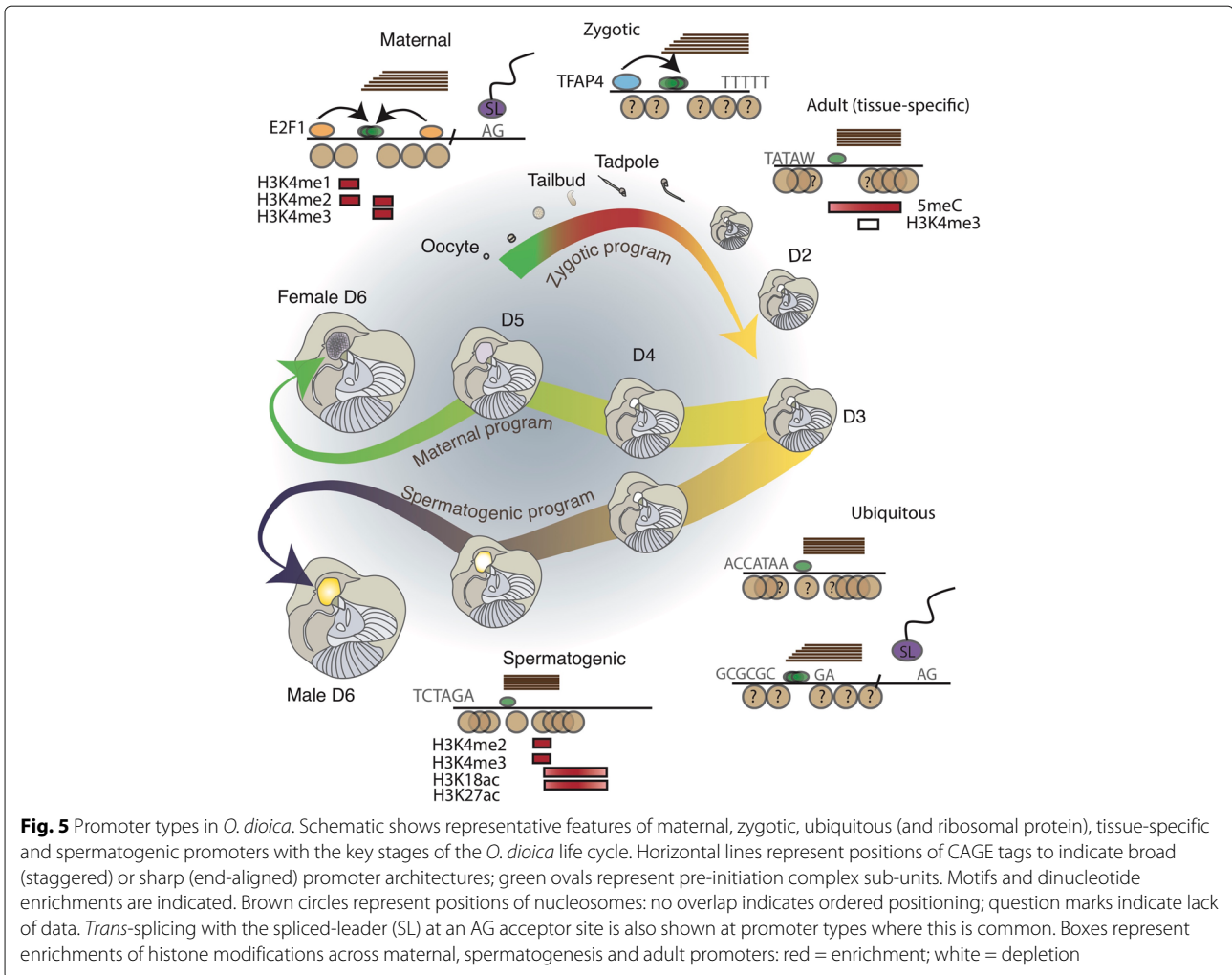


Fig. 5 Promoter types in *O. dioica*. Schematic shows representative features of maternal, zygotic, ubiquitous (and ribosomal protein), tissue-specific and spermatogenic promoters with the key stages of the *O. dioica* life cycle. Horizontal lines represent positions of CAGE tags to indicate broad (staggered) or sharp (end-aligned) promoter architectures; green ovals represent pre-initiation complex sub-units. Motifs and dinucleotide enrichments are indicated. Brown circles represent positions of nucleosomes: no overlap indicates ordered positioning; question marks indicate lack of data. *Trans*-splicing with the spliced-leader (SL) at an AG acceptor site is also shown at promoter types where this is common. Boxes represent enrichments of histone modifications across maternal, spermatogenesis and adult promoters: red = enrichment; white = depletion

A recent study in zebrafish found that maternal promoters are sharp (or multiple sharp) with a TATA-like upstream element whereas zygotic promoters are broad with transcription initiation guided by ordered nucleosome positioning [4]. In comparison, we found that in *O. dioica*, both maternal and zygotic promoters are broad. Moreover, we found evidence of nucleosome positioning as well as an enrichment for the binding of E2F1 at broad maternal promoters. These differences in maternal TSS-selection between zebrafish and *O. dioica* may be due to different modes of oogenesis and different sources of maternal transcripts. In zebrafish maternal transcripts originate from oocyte nuclei whereas in *O. dioica* the majority of maternal transcripts originate from terminally differentiated polyploid nurse nuclei within the single-cell coenocyst and are transported to oocytes through ring canals [31, 32].

Most maternal transcripts are *trans*-spliced in *O. dioica* [20] and this may have influenced the evolution of maternal promoter architectures. Since *trans*-splicing removes

the 5' end of a pre-mRNA (the "outtron") it follows that this sequence has little, if any, role in the post-transcriptional regulation of its mRNA. Indeed, one hypothesis for the function of *trans*-splicing in monocistrons is that it removes deleterious sequences at the 5' end of an mRNA (e.g. premature start codons). There is mounting evidence that the SL sequence itself plays an important role in translational control, particularly for TOP mRNAs, which are *trans*-spliced in *O. dioica* [20]. We have shown here that the conserved TCT initiator sequence, which constitutes the first two nucleotides of the TOP motif, is absent at *O. dioica* ribosomal protein TSSs. We hypothesize that there is no requirement for a strict site of transcription initiation for *trans*-spliced genes since the spliced leader provides any necessary 5' post-transcriptional regulatory motifs. Promoters of *trans*-spliced genes are then permitted to adopt a broad architecture governed by chromatin state rather than sequence motifs.

Our data also revealed a remarkable genome-wide shift in mode of TSS-selection during spermatogenesis to

one associated with a position-specific core promoter motif (TCTAGA). This shift is likely regulated by a basal transcription factor that is specifically expressed in the male. With a distance (44 bp) from the TSS similar to that of the TATA-box (36 bp) it is tempting to speculate that a TFIID complex containing a male-specific variant of the TATA-binding protein (TBP) is the factor binding TCTAGA and driving expression of male-specific genes. Indeed, variants of the basal transcription factor (TF) machinery are known to play roles in development and gametogenesis across metazoans [33]. One consequence of this shift is the developmental regulation of *trans*-splicing during spermatogenesis: many mRNAs that are *trans*-spliced in other stages (often in operons) are transcribed during spermatogenesis from an alternative TSS, driven by a TCTAGA promoter located downstream of the *trans*-splice acceptor site. This *trans*-splice acceptor site is thereby excluded from resulting mRNAs, which are no longer *trans*-spliced with the SL sequence. This may lead to a switch in the translational control of these transcripts to one that is independent of nutrient levels [20]. We hypothesise that this translational control is not required during the non-vitellogenic process of spermatogenesis. Nutrient-dependent control over initiation of meiosis has, however, been described in both sexes of *O. dioica* [34]. The TCTAGA promoter motif may play a role in this regulation in males if its binding by a transcription factor is nutrient-dependent.

A recent study found that genes with transcription-associated gene body methylation encode more highly conserved proteins with typical “housekeeping” functions [9]. We discovered a strong association of gene body DNA methylation with TATA-dependent promoters in *O. dioica*. This relationship is present during early development as well as in both the male and female germ lines, despite these differing substantially in their chromatin landscapes [21]. Promoters with the male-specific TCTAGA motif did not exhibit this downstream DNA methylation enrichment, despite this motif being position-specific and located in the expected TATA-box position. This indicates that gene body methylation in a subset of *O. dioica* genes is driven by core promoter features, specifically the TATA-box. A study in *C. intestinalis* found that gene bodies in near identical sets of genes are methylated in different cellular contexts [35], which is similar to our observations in *O. dioica*. This study also showed, however, that features within two ubiquitously expressed promoters are not the primary determinant of gene body DNA methylation. Analysis of additional *C. intestinalis* promoters may nevertheless reveal a relationship with the TATA-box similar to what we observe in *O. dioica*. Further exploration of sequence context in both species may also reveal a role for additional factors.

Given that DNA methylation in gene bodies suppresses transcription from alternative downstream promoters [7, 8] it is tempting to speculate that TATA-dependent sharp promoters employ DNA methylation as additional insurance for the strict positioning of transcription initiation. We also observed a depletion of H3K4me3 at, and downstream of, TATA-dependent promoters, in line with the inhibitory effect of H3K4me3 on DNA methyltransferases. Since TFIID is anchored at H3K4me3 on the +1 nucleosome [36] this indicates that TATA-dependent promoters are bound by TBP as part of an alternative complex. In yeast, TATA-dependent promoters are depleted of both TFIID and a nucleosome positioned downstream of the TSS and TBP is instead directed to the TATA-box by the SAGA complex [37]. Further investigation is required to establish whether or not a similar situation exists in metazoans.

Our results support previous findings of overlapping promoter codes [4], while revealing additional diversity and differential usage during complex developmental transitions. We provide the first links between acquisition of *trans*-splicing and the reorganization of promoter architectures for a conserved set of core metabolic genes, probably arising at least in part, because of regulatory sequences encoded in the SL. We also show shifts in TSS selection associated with a previously unidentified core promoter motif during the spermatogenic program. Further work on a range of additional models would provide a better framework in understanding the evolution of core promoter architectures, particularly with respect to innovations within major lineages.

Methods

Modified cap analysis of gene expression (CAGE)

Total RNA from each stage of development was isolated using RNAqueous Micro (Ambion) and treated by TerminatorTM 5'-Phosphate-Dependent Exonuclease to deplete excess small RNAs. A modified CAGEScan protocol [11] was carried out at DNAFORM, Yokohama City, Japan. The standard CAGEScan protocol was modified in order to separate *trans*-spliced from non-*trans*-spliced transcripts by first using a custom designed 5' linker, specific to the 5' spliced leader sequence, before using standard linkers for non-*trans*-spliced mRNAs. Sequenced libraries for each stage therefore included only non-*trans*-spliced transcripts.

Mapping reads

Illumina sequencing generated a total of 39,124,333 reads, 37 nt in length. We mapped these to the *O. dioica* reference genome [16] using Bowtie [38] with default parameters (allowing 2 mismatches per read). The 5' coordinates of all uniquely mapping read (CAGE tag) alignments were extracted from the Bowtie

output to give positions of CAGE transcription start sites (CTSSes), and the number of tags at each position was computed to give a tag count for each CTSS. We normalized tag counts to tags per million reads (tpm).

Promoter types

We used the R package “CAGEr” [23] to cluster CTSSes into CAGE tag clusters, excluding those with < 1 tpm and singletons < 5 tpm, using a maximum distance of 20 bp between CTSSes within a cluster. We calculated the interquartile range ($q_{0.1} - q_{0.9}$) of promoter widths (a measure of how broad/dispersed or peaked/focused a promoter’s TSS usage is that is more robust to expression level than using the full promoter width). We used this to group promoters into four classes using the mean and upper and lower quartiles as thresholds. We defined the upper and lower quartiles as “broad” and “sharp” respectively. We categorized promoters by CpG frequency in a 200 bp window centered on the dominant CTSS. Promoters with a CpG frequency in the upper quartile of CpG frequencies were classed as high CpG (HCG) and promoters with a CpG frequency in the lower quartile classed as low CpG (LCG). Using CAGEr we grouped all tag clusters with > 5 tpm across all stages into consensus promoter regions, using the interquartile range ($q_{0.1} - q_{0.9}$) of tag cluster widths and a distance of 100 bp to merge clusters into one region. We used SOM clustering both at the level of individual CTSSes and consensus promoter regions to generate 25 expression profiles in each case.

Shifting promoters

We calculated a shifting score and p -value of Kolmogorov-Smirnov test for all consensus promoters for all pairwise comparisons. We used a score > 0.6 and FDR < 0.01 to define a promoter shift – identifying promoters that have at least 60% of transcription initiation in the sample with lower expression occurring either upstream or downstream of transcription initiation in the compared sample.

Assigning CTSSes to gene models and operons

We used Genoscope gene model predictions and annotations of polycistrons (www.genoscope.fr) to classify genes into operons and non-operons. A CTSS was associated with a gene model if it overlapped a gene body or its 500 bp upstream region. Using previously published CAGE data for *trans*-spliced transcripts [20], we classed a gene as SL *trans*-spliced if there was a SL CTSS within the gene body or within a 500 bp upstream region, if it was supported by > 1 tag count and if it had an ‘AG’ acceptor site motif immediately upstream.

GO analysis

We used *O. dioica* GO annotations [22]. We used the Bioconductor GOstats package in R to compute hypergeometric p -values for over-representation of GO terms in different sets of genes.

Motif analyses

Over-represented motifs in core promoter regions were identified using MEME with default parameters on sequences in a 200 bp window, centred on the dominant CTSS within each tag cluster, for groups of CTSSes of interest. We also identified position-specific motifs (including initiator trinucleotides) by scanning core promoter regions for the occurrence of all possible k -mers (for $k=1-6$). We used TOMTOM to match position-weight matrices of motifs identified by MEME to known transcription factor binding sites [39]. We plotted the dinucleotide content of promoters using the R package “seq-Pattern”. We searched for TATA elements in the region 37–22 bp upstream of the dominant CTSS in each TC. We searched for TCTAGA motifs in the 22–52 bp and 52–101 bp upstream regions. We searched for ACCATAA motifs in the 32–72 bp upstream region. We used zygotic (CTSS SOM cluster 0_0) promoter sequences (200 bp centred on the dominant CTSS) from the tailbud stage to identify over-represented zygotic motifs. We used the “Biostrings” R package to scan (using a minimum score of 85%) the 101 bp upstream region with the position weight matrix discovered by MEME that matched the binding site for human TFAP4.

ChIP-chip analysis

We analysed previously published meDIP-chip data and ChIP-chip data for E2F1, H3 and histone modifications H3K4me1, H3K4me2, H3K4me3 and H3K27me3 from mature *O. dioica* ovaries and testes [21]. We used the Bioconductor R package Ringo [40] for pre-processing all ChIP-chip data. Briefly, we normalized raw probe intensities from each sample (Cy5 channel) to corresponding input DNA probe intensities (Cy3 channel) by computing $\log_2(\text{Cy5}/\text{Cy3})$. We used the NimbleGen normalization method, which adjusts for systematic dye and labeling biases by subtracting from individual \log_2 ratios the Tukey’s biweight mean, computed across each sample’s \log_2 ratios. To reduce noise in the data we smoothed the normalized \log_2 ratios using a running median across a 150 bp window (the approximate size of a single nucleosome) with a minimum threshold of 3 non-zero probes. For each group of promoters we plotted the mean \log_2 ratio at each probe position for all probes in a 1000 bp window centred on the dominant CTSSes of promoter regions of interest. We excluded promoters with flanking regions that overlap. We defined regions of ChIP-enrichment genome-wide as previously described [21].

Operon transcription analysis

We used tiling array data generated from *O. dioica* testes and ovaries [22] to categorize genes within operons as expressed or silent. We then defined an operon as expressed if any of its genes are classed as expressed. We intersected H3K4me3 ChIP-enriched regions with operon promoters, as well as potential promoters of internal operon genes, using the region 500 bp upstream and 100 bp downstream of annotated start sites. Any overlap was defined as a presence of H3K4me3 in a candidate promoter region.

5' RACE

RACE was performed using SMARTER RACE kit from Clontech according to manual.

Re-analysis of mouse testis CAGE data

Analysis of TSS data followed that found in [23] using data downloaded from <http://promshift.genereg.net/CAGEr/InputData/> consisting of TSSs from 8 stages of mouse testis development. Briefly, we used the CAGEr [23] package to normalize tag counts and cluster TSSs into TCs for each stage. We plotted the frequency of TCTAGA motif around TSSs from each stage, sorted by the width of TCs and saw no enrichment. We then used a self organizing map (SOM) to cluster the expression profiles of each TSS and identified a cluster with expression specific to later development which was previously annotated as being enriched for TSSs associated with spermatogenesis genes [23]. We plotted the TCTAGA frequency around the TSSs of this cluster and also saw no enrichment.

Search for TCTAGA and TATA-box motifs in *C. intestinalis* promoters

We searched the 100 bp region upstream of all annotated Ensembl 87 KH *C. intestinalis* genes for “TCTAGA” motif and the consensus TATA-box motif “TATAWAR” using the “Biostrings” R package.

Additional file

Additional file 1: Supplementary Material. Supplemental Results, Supplemental Methods, **Table S1**, **Table S2**, **Figures S1 - S12** and Supplemental References. (PDF 13 464 kb)

Abbreviations

5mC: 5-methylcytosine; CAGE: Cap analysis of gene expression; ChIP: chromatin immunoprecipitation; CTSS: CAGE transcription start site; GO: Gene ontology; Inr: Initiator; meDIP: Methylated DNA immunoprecipitation; RACE: Rapid amplification of cDNA ends; SL: Spliced leader; SOM: Self-organizing map; TC: Tag cluster; TF: Transcription factor; TOP: Terminal OligoPyrimidine; tpm: Tags per million; TSS: Transcription start site

Acknowledgements

We thank Matthias Harbers and his staff at DNAFORM (Yokohama, Japan) for their assistance in the development of the modified CAGE protocol. We thank Jean-Marie Bouquet, Magnus Reeve and Anne Aasjord for supplying animals as part of the animal culture facility.

Funding

This work was supported by grants 183690/S10 and 133335/V40 from the Norwegian Research Council.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the NCBI Gene Expression Omnibus under accession number GSE78794 and GSE78915.

Authors' contributions

Conceptualization, GBD, EMT, and BL; Methodology, GBD, PN, EMT, and BL; Investigation, GBD, PN, BL, and EMT; Formal Analysis, GBD; Visualization, GBD; Validation, PN; Writing – Original Draft, GBD; Writing – Review & Editing, GBD, PN, BL, and EMT; Funding Acquisition, EMT; Supervision, EMT and BL.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Sars International Centre for Marine Molecular Biology, University of Bergen, N-5006 Bergen, Norway. ²Computational Regulatory Genomics, MRC London Institute of Medical Sciences, W12 0NN London, United Kingdom. ³Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, W12 0NN London, United Kingdom. ⁴Department of Biology, University of Bergen, N-5006 Bergen, Norway.

Received: 28 September 2017 Accepted: 28 January 2018

Published online: 26 February 2018

References

- Juven-Gershon T, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol*. 2010;339(2):225–9.
- Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*. 2012;13(4):233–45.
- Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol*. 2006;7(8):78.
- Haberle V, Li N, Hadzhiiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, van IJcken WFJ, Armant O, Ferg M, Strähle U, Carninci P, Muller F, Lenhard B. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*. 2014;507(7492):381–5.
- Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science*. 2010;328(5980):916–9.
- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev Genet*. 2012;13(7):484–92.
- Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJM, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*. 2010;466(7303):253–7.
- Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, Maldotti M, Anselmi F, Oliviero S. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*. 2017;543(7643):72–7.
- Sarda S, Zeng J, Hunt BG, Yi SV. The evolution of invertebrate gene body methylation. *Mole Biol Evol*. 2012;29(8):1907–1916.
- Suzuki MM, Kerr ARW, De Sousa D, Bird A. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res*. 2007;17(5):625–31.
- Kodzios R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P. CAGE: cap analysis of gene expression. *Nature methods*. 2006;3(3):211–22.

12. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesni A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* 2006;38(6):626–35.
13. Yokomori R, Shimai K, Nishitsuji K, Suzuki Y, Kusakabe TG, Nakai K. Genome-wide identification and characterization of transcription start sites and promoters in the tunicate *Ciona intestinalis*. *Genome Res.* 2016;26(1):140–50.
14. Parry TJ, Theisen JWM, Hsu JY, Wang YL, Corcoran DL, Eustice M, Ohler U, Kadonaga JT. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes & Dev.* 2010;24(18):2013–018.
15. Meyuhas O. Synthesis of the translational apparatus is regulated at the translational level. *European J Biochem.* 2001;267(21):6321–330.
16. Denoeud F, Henriot S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Cañestro C, Bouquet JM, Danks G, Poulain J, Campsteijn C, Adamski M, Cross I, Yadetie F, Muffato M, Louis A, Butcher S, Tsagkogeorga G, Konrad A, Singh S, Jensen MF, Cong EH, Eikeseth-Otteraa H, Noel B, Anthouard V, Porcel BM, Kachouri-Lafond R, Nishino A, Ugolini M, Chourrout P, Nishida H, Aasland R, Huzurbazar S, Westhof E, Delsuc F, Lehrach H, Reinhardt R, Weissenbach J, Roy SW, Artiguenave F, Postlethwait JH, Manak JR, Thompson EM, Jaillon O, Du Pasquier L, Boudinot P, Liberles DA, Volff JN, Philippe H, Lenhard B, Roest Crolihus H, Wincker P, Chourrout D. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science.* 2010;330(6009):1381–1385.
17. Blumenthal T. Operons in eukaryotes. *Brief Funct Genom & Proteomics.* 2004;3(3):199–211.
18. Hastings KEM. SL trans-splicing: easy come or easy go?. *Trends Genet.* 2005;21(4):240–7.
19. Ganot P, Kallesøe T, Reinhardt R, Chourrout D, Thompson EM. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mole Cell Biol.* 2004;24(17):7795–805.
20. Danks GB, Raasholm M, Campsteijn C, Long AM, Manak JR, Lenhard B, Thompson EM. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mole Biol Evol.* 2015;32(3):585–99.
21. Navratilova P, Danks GB, Long A, Butcher S, Manak JR, Thompson EM. Sex-specific chromatin landscapes in an ultra-compact chordate genome. *Epigenetics & Chromatin.* 2017;10:3.
22. Danks G, Campsteijn C, Parida M, Butcher S, Doddapaneni H, Fu B, Petrin R, Metpally R, Lenhard B, Wincker P, Chourrout D, Thompson EM, Manak JR. OikoBase: a genomics and developmental transcriptomics resource for the urochordate *Oikopleura dioica*. *Nucleic Acids Res.* 2013;41(Database issue):845–53.
23. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* 2015;43(8):51–1.
24. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, De Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmid C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescato M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JAC, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drabløs F, Edge ASB, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furino M, Furusawa J-i, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiacki L, Iida K, Ikawa T, Jankovic BR, Jia H, Joshi A, Jurman G, Kaczkowski B, Kai C, Kaida K, Kaiho A, Kajiyama K, Kanamori-Katayama M, Kasianov AS, Kasukawa T, Katayama S, Kato S, Kawaguchi S, Kawamoto H, Kawamura YI, Kawashima T, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klincken SP, Knox AJ, Kojima M, Kojima S, Kondo N, Koseki H, Koyasu S, Krampitz S, Kubosaki A, Kwon AT, Laros JFJ, Lee W, Lennartsson A, Li K, Lilje B, Lipovich L, Mackay-Sim A, Manabe R-i, Mar JC, Marchand B, Mathelier A, Mejhert N, Meynert A, Mizuno Y, de Lima Morais DA, Morikawa H, Morimoto M, Moro K, Motakis E, Motohashi H, Mummery CL, Murata M, Nagao-Sato S, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nakazato K, Van Nimwegen E, Ninomiya N, Nishiyori H, Noma S, Nozaki T, Ogishima S, Ohkura N, Ohimiya H, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Pain A, Passier R, Patrikakis M, Persson H, Piazza S, Prendergast JGD, Rackham OJL, Ramilowski JA, Rashid M, Ravasi T, Rizzu P, Roncador M, Roy S, Rye MB, Saijyo E, Sajantila A, Saka A, Sakaguchi S, Sakai M, Sato H, Savvi S, Saxena A, Schneider C, Schultes EA, Schulze-Tanzil GG, Schwegmann A, Sengstag T, Sheng G, Shimoji H, Shimoni Y, Shin JW, Simon C, Sugiyama D, Sugiyama T, Suzuki M, Suzuki N, Swoboda RK, 't Hoen PAC, Tagami M, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tatum Z, Thompson M, Toyodo H, Toyoda T, Valen E, van de Wetering M, van den Berg L. A promoter-level mammalian expression atlas. *Nature.* 2014;507(7493):462–70.
25. Huang P, Pleasance ED, Maydan JS, Hunt-Newbury R, O'Neil NJ, Mah A, Baillie DL, Marra MA, Moerman DG, Jones SJM. Identification and analysis of internal promoters in *Caenorhabditis elegans* operons. *Genome Res.* 2007;17(10):1478–1485.
26. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes & Dev.* 2002;16(2):245–56.
27. Meisel RP, Malone JH, Clark AG. Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome research.* 2012;22(7):1255–1265.
28. Massari ME, Murre C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Molecular and cellular biology.* 2000;20(2):429–40.
29. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev Genet.* 2007;8(6):424–36.
30. Smallwood SA, Kelsey G. De novo DNA methylation: a germ cell perspective. *Trends Genet.* 2012;28(1):33–42.
31. Ganot P, Bouquet JM, Kallesøe T, Thompson EM. The *Oikopleura* coenocyst, a unique chordate germ cell permitting rapid, extensive modulation of oocyte production. *Dev Biol.* 2007;302(2):591–600.
32. Ganot P, Kallesøe T, Thompson EM. The cytoskeleton organizes germ nuclei with divergent fates and asynchronous cycles in a common cytoplasm during oogenesis in the chordate *Oikopleura*. *Dev Biol.* 2007;302(2):577–90.
33. Akhtar W, Veenstra GJC. TBP-related factors: a paradigm of diversity in transcription initiation. *Cell & Biosci.* 2011;1(1):23.
34. Subramaniam G, Campsteijn C, Thompson EM. Lifespan extension in a semelparous chordate occurs via developmental growth arrest just prior to meiotic entry. *PloS ONE.* 2014;9(4):93787.
35. Suzuki MM, Yoshinari A, Obara M, Takuno S, Shigenobu S, Sasakura Y, Kerr AR, Webb S, Bird A, Nakayama A. Identical sets of methylated and nonmethylated genes in *Ciona intestinalis* sperm and muscle cells. *Epigenetics & chromatin.* 2013;6(1):38.
36. Vermeulen M, Mulder KW, Denisov S, Pijnappel WWMP, van Schaik FMA, Varier RA, Baltissen MPA, Stunnenberg HG, Mann M, Timmers HTM. Selective anchoring of TFIIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell.* 2007;131(1):58–69.
37. Rhee HS, Pugh BF. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature.* 2012;483(7389):295–301.
38. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):25.
39. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37(Web Server):202–8.
40. Toedling J, Sklyar O, Huber W. Ringo – an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC bioinformatics.* 2007;8(1):221.