Promoter architecture and sex-specific gene expression in the microcrustacean *Daphnia pulex* revealed by large-scale profiling of 5′-mRNA ends

R. Taylor Raborn*[1], Ken Spitze[1], Volker P. Brendel[1,2,3] and Michael Lynch[1,3]

[1]Department of Biology, Indiana University

[2]School of Informatics and Computing, Indiana University

[3]These authors jointly supervised this work.

April 20, 2016

Keywords: CAGE, Daphnia, gene regulation, meiosis, promoter architecture, transcription start sites

*Correspondence: 212 South Hawthorne Drive, Simon Hall 205B. Bloomington, IN 47405. Email: rtraborn@indiana.edu

Abstract

Large-scale identification of transcription start sites (TSSs) using Cap Analysis of Gene Expression (CAGE) has yielded insight into promoter location, architecture, and regulation for a small set of taxa representing major model organisms, including human. However, comparative and evolutionary genomics studies of *cis*-regulatory control of transcription initiation for a wider spectrum of Metazoa are still outstanding. To broaden our understanding of core promoter structure in species from metazoan clades with currently scant genome data, we sought to characterize the landscape of *cis*-regulatory elements in the microcrustacean water flea *Daphnia pulex*, an important model organism for studies in ecology, toxicology, and genetics. We performed CAGE from total RNA derived from three states: sexual females, asexual females, and males, reflecting distinct sexes and modes of reproduction. We mapped over 120 million CAGE reads to the *D. pulex* genome and generated a *Daphnia* promoter atlas containing 12,662 unique promoters. Characterization of the transcription initiation sites showed the expected enrichment of the CA-dinucleotide at TSSs [-1,+1] (associated with Initiator-motif containing promoters) but also significant over-representation of GN-dinucleotides. Overall, these data suggest that *D. pulex* initiation sites are among the most GC-rich yet observed in metazoans. Computational *de novo* motif discovery around CAGE-identified TSSs revealed eight putative core promoter elements, including the canonical TATA (TATAWAA) and Initiator (CAGWY) motifs, as well as statistically significant motifs with no obvious orthologs in other metazoans. Analysis of the differentially-expressed genes suggests that a considerable number of cell cycle genes (each with net negative regulatory effects on meiosis) are upregulated in asexual females, providing a glimpse of the molecular events that underpin the cyclical parthenogenesis in *D. pulex*. Taken together, this work provides the first picture of transcription initiation and promoter architecture within Crustacea. The *Daphnia* promoter atlas we present here provides a basis for future study among *Daphnia* spp. as well as for comparative genomic analyses of metazoan transcriptional control.

Introduction

All biological processes, including development, differentiation, and maintenance of homeostasis, rely upon precise, coordinate regulation of gene expression. A key early step in gene expression is transcription initiation at the core promoter, a short genomic region containing the transcription start site (TSS) (Kadonaga 2012). During initiation, sequences within core promoters recruit general transcription factors (GTFs), which is followed by binding of RNA polymerase II (RNAPII) and formation of the pre-initiation complex (PIC) (Cosma 2002). Identifying the locations and composition of promoters is fundamental for understanding the basis for gene expression regulation. Recent work (Frith et al. 2008, Hoskins et al. 2011) demonstrates that core promoters are structurally more diverse than previously appreciated. This diversity is thought to reflect large numbers of developmental programs and regulatory strategies (Lenhard et al. 2012), but the precise rules and mechanisms underlying promoter function remain unclear.

Genome-scale TSS profiling has identified promoters in a number of metazoans (FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014, Lenhard et al. 2012). CAGE (Cap Analysis of Gene Expression) (Kodzius et al. 2006, Kurosawa et al. 2011), the most prominent TSS profiling method, identifies core promoter positions at high resolution. This approach revealed that most genes do not possess a single TSS, but instead exhibit sets of closely spaced TSSs that will be referred to as **T**ranscription **S**tart **R**egions (TSRs) in the following. While the largest number of TSS profiling studies have been performed in mammalian (human and mouse) systems (Djebali et al. 2008, FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014), CAGE has also been performed in non-mammalian metazoans, including fruit fly (Hoskins et al. 2011), nematode (Nepal et al. 2013), and zebrafish (Haberle et al. 2014). Overall, these studies indicate that the majority of core promoters in metazoan genomes lack TATA elements (Lenhard et al. 2012), an unanticipated finding given previously established models for transcription initiation. At least two major promoter classes are evident. In human and mouse, the largest class is known as CpG island promoters (CPI) (Saxonov et al. 2006, Lenhard et al. 2012). These promoters are located near CpG islands and are generally of high GC-content and depleted for TATA elements. Sequences in the other major promoter class, called "low-CpG", exhibit low GC-content and are enriched for TATA boxes. This latter class of promoter is consistent with conventional models of promoter structure, such as TATA-dependent transcription initiation (Kadonaga 2012).

Characterization of CAGE-defined promoters in a wider taxonomic context uncovered two distinct patterns of TSS distributions within a given promoter (Carninci et al. 2006, Hoskins et al. 2011, Lenhard et al. 2012). "Peaked" promoters exhibit CAGE signal from a narrow genomic region surrounding a single prominent TSS, whereas "broad" promoters instead feature multiple TSSs distributed across a wide (30 bp and longer) genomic region (Kadonaga 2012, Lenhard et al. 2012). These TSS distribution patterns appear to coincide with the aforementioned (mammalian) classes of promoter architecture: peaked promoters are highly associated with the low-CpG promoter class, whereas broad TSS distributions tend to be found at high-CpG promoters. Peaked and broad promoters also regulate separate functional gene classes: genes with peaked promoters tend to be developmentally regulated or tissue-specific, while genes with broad promoters tend to be housekeeping genes exhibiting constitutive expression (Lenhard et al. 2012). Recent work using CAGE from a variety of mammalian cell types unexpectedly detected widespread enrichment of TSSs at enhancers (Andersson et al. 2014). The new class of RNA defined by this work, *enhancer RNAs* (eRNAs), are short, transient, RNAPII-derived transcripts generated at active enhancer regions. While enhancers appeared to be distinguishable from promoters on the basis of transcript stability and bidirectionality (Andersson et al. 2014), subsequent reports suggest that enhancers and promoters possess common properties, including motif composition and activity (Arner et al. 2015).

Despite recent progress, considerable gaps remain in the understanding of promoter architecture across metazoan diversity. To date, high-resolution TSS profiling has been reported in just two arthropod species, both closely-related drosophilids: *D. melanogaster* (Hoskins et al. 2011) and *D. pseudoobscura* (Chen et al. 2014). Promoter profiling in a broader set of taxa is necessary to establish robust comparative genomic analyses of *cis*-regulatory regions in metazoa. To address this need, we performed TSS profiling using CAGE in the water flea *Daphnia pulex.* A freshwater microcrustacean with a cosmopolitan distribution, *D. pulex* is notable for its ability to reproduce both sexually and asexually, high levels of heterozygosity, and relatively large effective population sizes (N_e) compared to other broadly dispersed arthropods (Tucker et al. 2013, Haag et al. 2009). *D. pulex* serves as a key model system throughout the biological sciences, from ecosystem ecology to molecular genetics. By mapping TSSs for *D. pulex* from active promoters within the three developmental states of sexual females, asexual females, and adult (sexual) males, we sought to characterize the architecture of core promoters in *D. pulex* and also explore meiosis- and sex-specific gene regulatory programs. We successfully identified TSSs at high resolution across the entire genome, defining promoters for all genes expressed under the experimental conditions. We then performed computational *de novo* motif discovery using this set of set of mapped TSSs, obtaining consensus sequences of canonical core promoter elements, including TATA and Initiator (Inr). The quantitative tag counts from the CAGE datasets allowed us to identify differentially-expressed genes within each of the three states surveyed, including those regulated in a sex-specific manner. The resultant *D. pulex* promoter atlas extends our knowledge of metazoan *cis*-regulation into Crustacea, a taxonomic expansion that will also serve as a public resource for functional and comparative genomics.

Results

Profiling 5′ mRNA ends characterizes the global landscape of transcription initiation

Interrogation of capped 5′-ends of mRNAs identifies the locations and patterns of transcription initiation within a genome. Through biochemical capture of these 5′ transcript ends (see Methods), CAGE ultimately generates short, strand-specific sequences (CAGE tags), the 5′-ends of which correspond to the first base of the associated mRNA. Sequenced CAGE tags (47bp in length) were aligned to the genome (Figure 1A, panel i). The coordinate corresponding to the 5′ aligned base of each aligned read is defined as a CAGE-detected TSS (CTSS; Figure 1A, panel ii). Multiple CAGE tags mapping to identical CTSS coordinates provide a quantitative measure of the abundance of mRNA ends that originated from that position. Individual CTSSs supported by sufficient numbers of CAGE tags (significant CTSSs; abbreviated sCTSSs; see Methods) occurring in close proximity in the genome were clustered to yield transcription start regions (TSRs) that correspond to genomic intervals that coincide with transcriptionally active promoters (Figure 1A, panel iii). Finally, when CAGE data from multiple conditions or tissues were compared, we define TSRs that agree (*i.e.* overlap) in all cases as “consensus promoters” (Figure 1A, panel iv).

A promoter atlas in *Daphnia pulex*

D. pulex can reproduce asexually, through ameiotically-produced eggs that develop directly, and sexually, through diapausing eggs. We generated CAGE datasets from three distinct adult states of *D. pulex* (Figure 1B; leftmost panel): males, parthenogenetic females (hereafter asexual females), and pre-ephippial females (hereafter sexual females). These states were chosen to potentially identify distinct genes and gene networks associated with meiosis, parthenogenesis, and sex-specificity. We sequenced eight libraries, generating 1.82×10^8 CAGE reads overall (Table 1), of which 1.22×10^8 (67.0%) mapped successfully to the current version (JGIv1.1) of the *D. pulex* assembly (Colbourne et al. 2011). After normalization (see Methods), replicates for each state were highly correlated (Pearson coefficient >0.97; Figure S1 and S2). We then applied a computational analysis pipeline to identify CTSSs, TSRs and consensus promoters (Figure 1A) from CAGE reads across each of the three states (See Methods).

We evaluated our CAGE definitions in their entirety by considering their locations within the *D. pulex* genome. Among CTSSs (n=2,332,582) pooled across all states, we observe that a sizable fraction (67.5%) were located within 1 kb of a CDS, while 9.88% were present in the first 1 kb downstream of a stop codon (Figure 1C), an observation also reported in *D. melanogaster* (Hoskins et al. 2011). When CAGE tags are considered individually (rather than unique CTSSs alone), we report a substantially larger percentage (82.3%) located within the first 1 kb upstream of the translation start site of coding genes, while only a small fraction (1.95%) were located downstream of annotated CDSs (Figure 1C). From this we conclude that CTSSs supported by many CAGE reads are more likely to be positioned upstream of coding genes than those supported by fewer reads.

Similar numbers of TSRs (between 11,289 and 11,558) are identified within the three individual states, totaling 12,662 unique TSRs overall (Table 2). The majority of identified promoters (83.1%) were positioned within the first 1 kb upstream of coding genes, indicating general but incomplete agreement with the current *D. pulex* gene annotation (Figure 1C). This work represents a comprehensive, sex-specific promoter atlas in adult *D. pulex*, the first of its kind in crustaceans.

Promoter shape, base composition, and expression class

The property of the distribution of TSSs is known to be key descriptor of the structure and composition of the underlying promoter in metazoans (Rach et al. 2009, Hoskins et al. 2011). We evaluated CAGE tag distributions at consensus promoters (n=10,580) using two criteria. The first is *width*, which is defined as the length of the genomic segment occupied by all CTSSs within a TSR or consensus promoter. We observe an ample range of widths (2–163 bp), including a small number (1104; 10.4%) of TSRs with widths >30 bp (Figure 2A). Overall, We observe a median width of 5 bp, and a mean width of 12 bp for all consensus promoters. We applied a second metric, promoter shape, which measures the stability of the CAGE tag distribution at a TSR. For example, a TSR with a sharp distribution of CAGE tags surrounding a single major CTSS would be considered *peaked*, whereas a TSR with numerous distinct CTSSs supported by roughly equivalent numbers of CAGE tags would be *broad*. We applied the Hoskins Shape Index (SI) (Hoskins et al. 2011) to measure shape across all consensus promoters. We also observe a wide range of consensus promoter shapes (Figure 1A, inset); the observed median and mean SI values were -0.42 and -0.54, respectively.

Two distinct promoter classes have been proposed in mouse, human and *Drosophila*, defined according to the shape of empirical (generally CAGE-based) 5′-end distributions (Carninci et al. 2006, Hoskins et al. 2011, Kadonaga 2012). We reasoned that if two distinct classes of promoter exist in *D. pulex*, then the shapes we observe should be bimodally-distributed. We fit the distribution of consensus promoter shapes using an expectation-maximization (EM) algorithm (see Methods), and see strong support for a two-component mixture model (Figure 2A, inset), consistent with broad and peaked consensus promoter shapes. This result provides evidence for the existence two classes of promoter in *D. pulex* and is consistent with previous findings. We classified consensus promoters into categories according to SI, peaked (n=738), broad (n=1318) or unclassified (see Methods). An example of a peaked and broad consensus promoters found within our CAGE dataset is shown in Figure 2C. We then asked if promoter expression (the abundance of CAGE tags associated with a consensus promoter) varied by promoter shape class (see Methods). We find that broad TSRs have significantly higher expression [p <0.0003710] than peaked and unclassified TSRs (Figures 2D and 2E). However, we do not observe a similar relationship between expression and promoter width (data not shown). This suggests that, in *D. pulex*, shape is more reflective of promoter properties than width.

Dinucleotide preferences of *D. pulex* TSSs

Global studies of transcription initiation across metazoan diversity identified distinct dinucleotide compositions at the TSS (Frith et al. 2008, Nepal et al. 2013). We investigated dinucleotide preferences in *D. pulex*, measuring the dinucleotide frequencies present within the [-1,+1] interval relative to CTSSs. We observe a strong preference for CA, GA, GC, GG, and GT relative to background (p <0.01; see

Methods) and considerable depletion for AT-rich dinucleotides AA, AT and TT (p <0.02, 0.01 and 0.01, respectively; Figure 2B).

De novo discovery of consensus promoter elements in *D. pulex*

Core promoter elements and their motif consensus sequences have been identified in *D. melanogaster* (Ohler et al. 2002, Down et al. 2007, Kadonaga 2012), mammals (*i.e.* human and mouse) (Carninci et al. 2006, FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014) and other metazoan model organisms: worm, *C. elegans* (Saito et al. 2013), and zebrafish, *D. rerio* (Nepal et al. 2013, Haberle et al. 2014).

Cis-regulatory motifs of any kind in *D. pulex* are unknown, so we sought to identify core promoter elements using the CAGE data generated in this study. To accomplish this, we performed *de novo* motif discovery using CAGE evidence (see Methods), applying sequence windows corresponding to core promoters ([-50,+50]). This procedure revealed a set of eight core promoter elements in *D. pulex* (Figure 3). To evaluate their similarity to known core promoter elements, we performed sequence alignment of each position weight matrix (PWM) against two motif sets: the complete JASPAR database (Portales-Casamar et al. 2009) and a curated list of 14 non-redundant core promoter motifs in *D. melanogaster*. We find two motifs within our set with strong sequence identity to the most well-characterized metazoan core promoter elements. The motif *Dpm2*, which has the consensus TATAWAA, has significant identity to the TBP-binding motif consensus in JASPAR (MA0108.1_TBP, e-value = 6.19×10^{-9}) in addition to the TATA element of *D. melanogaster* (E-value = 7.49×10^{-10}). The TATA-like *Dpm2* was observed in 9.48% of promoters. The motif *Dpm3*, with the consensus NCAGTY, has significant sequence similarity to the Initiator (Inr) element (consensus TCAKTY) (E-value = 6.097×10^{-6}) of *D. melanogaster* and is found at 12.04% of promoters.

In addition to TATA and Inr, we report a variety of motifs within our set of *D. pulex* core promoter elements (Figure 3). *Dpm5* (consensus TGGCAACNYYG), exhibits significant similarity to (E-value = 5.76×10^{-8}) to the "Ohler8" motif in *D. melanogaster* (Ohler et al. 2002). All of the remaining motifs match significantly with at least one motif in the JASPAR database. Among these, three motifs exhibit similarity to well-characterized transcription factor binding sites (TFBSs): *Dpm4* (consensus ARATGGC) matches the CTCF motif in JASPAR (MA0139.1_CTCF) (E-value = 5.51×10^{-5}), *Dpm6*, (CGCTAGA) matches the ABF transcription factor binding site consensus (MA0266.1_ABF2) (E-value = 5.51×10^{-6}) (Portales-Casamar et al. 2009), and the motif *Dpm5* (consensus CARCGTTGCC) exhibits a significant match to the TFBS consensus of RFX1 (MA0365.1) (E-value = 2.12×10^{6}).

Motif co-occurrence at promoters

After completing *de novo* discovery of core promoter elements in *D. pulex* (Figure 3), we sought to characterize the overall motif composition of promoters within the *Daphnia* promoter atlas. We used the consensus sequences of each of the eight motifs in the *Daphnia* promoter set and searched within a sequence window of [-200,+50] surrounding the midpoint of all annotated promoters. Using this information, we constructed a co-occurrence matrix for all identified promoter motifs, asking as to the overall coincidence of motifs within promoter regions. Several patterns of motif co-occurrence are observed among the *Dpm* motifs (Figure 4A). We find that TATA (*Dpm2*)-containing promoters are not enriched for other identified *Daphnia* motifs and are depleted for *Dpm4* and *Dpm5*. Inr (*Dpm3*) promoters are enriched for *Dpm6* and have fewer *Dpm4* and *Dpm5* motifs than expected. *Dpm4* while strongly enriched for *Dpm1* also exhibits significant enrichment for *Dpm5* and *Dpm6*. We observe strong co-occurrence between *Dpm6* and *Dpm7*. Three motifs, *Dpm1*, *Dpm6* and *Dpm7* have greater than expected frequencies of co-occurrence. Of note, none of the other core promoter elements are co-enriched with (*Dpm2*), and two (*Dpm4*, *Dpm5*) are depleted. This line of evidence suggests that TATA-containing promoters do not frequently act in combination with the other identified elements.

Positional enrichment of identified *D. pulex* core promoter elements

Many characterized core promoter elements are known to occur at specific locations relative to the TSS (+1). To determine the spatial characteristics of each of the *D. pulex* motifs, we evaluated their positional distributions relative to CTSSs and found that four of the eight *Dpm* motifs exhibit positional enrichment. We observe strong positional enrichment of *Dpm2* (TATA-like) and *Dpm3* (Inr-like) relative to *D. pulex* promoters (Figure 4B), with peaks at -30 and +1, respectively, consistent with the positions of TATA and Inr within other metazoans (Kadonaga 2012). *Dpm1* exhibits a modest peak at approximately +50, while *Dpm5* is enriched between -50 and -40 (Figure 4B and Figure 4C). *Dpm4* shows an irregular distribution within promoters, with two distinct peaks near -50 and +10 (Figure 4D). We do not observe a positional enrichment for motifs *Dpm6, Dpm7* and *Dpm8* (Figure 4D and data not shown). Taken together, this positional information allows us to construct an initial working model of the known core promoter elements in *D. pulex* (Figure 4E), and draw a comparison between canonical core promoter elements in *D. pulex* and *D. melanogaster* (Figure 4F).

Patterns of transcription initiation are known to relate to underlying promoter architecture in *Drosophila* (Rach et al. 2009, Hoskins et al. 2011) and mammals (Kadonaga 2012), so we asked whether possession of the two major core promoter elements Inr and TATA (*Dpm2* and *Dpm3*, respectively) is associated with TSR shape in *D. pulex.* Using the Shape Index (as previously described) to measure the focus and dispersion of CTSSs within a promoter we find that both Inr- and TATA-containing consensus promoters are significantly more peaked overall than TATA-less promoters <0.001 (Figure 4G), consistent with our expectations and the evidence in other metaozan model organisms including *D. melanogaster* (Rach et al. 2009, Hoskins et al. 2011).

Differential expression of *D. pulex* promoters

The abundance of CAGE tags that map to a putative promoter region provides quantitative measurement of the extent of transcription initiation at that site; this is capable of estimating expression of the associated genes (Murata et al. 2014, Balwierz et al. 2009), so we sought to identify differentially-expressed genes across the three states surveyed by our CAGE experiment. We used our defined set of consensus promoters (Table 2; n=10,665) and compared the normalized quantities of CAGE reads within a given state. Consensus promoter expression (*i.e.* the abundance of CAGE tags present at a consensus promoter in a given state) was measured using the number of mapped CAGE tags within the promoter and were represented in units of tags per million (tpm). An illustration of tag abundance within consensus promoters across the three states surveyed in this study is presented in Figure 5A. We carried out differential expression analysis across all libraries using limma (Ritchie et al. 2015), applying the mean-variance relationship of log-tpm (see Methods). During our analysis we compared promoter expression between each state separately (*e.g.,* sexual females vs. asexual females, *etc.*) in addition to the following comparisons: males vs. both females, sexual vs. asexual females, comprising five comparisons in total. We observe that an average of 1359 consensus promoters were differentially-expressed within each comparison: an average of 690 promoters exhibited significantly increased activity and 669 promoters had significantly decreased activity (Figure 5B). We observe the greatest number of differentially-expressed promoters (n=1206; upregulated, n=1052; downregulated) in the comparison between males and asexual females. Differentially expressed consensus promoters exhibit a complex topology of enrichment patterns across all three states; representative comparisons for asexual females are shown in Figure 5C and Figure 5D. Heatmaps of differentially-expressed promoters from other comparisons are presented in Figure S3.

Differentially-expressed promoters are enriched for endocrine and environmental response functions

We investigated the set of differentially-expressed promoters between each state, asking if the members of each respective gene set were enriched for common functions. We carried this out using the Gene Ontology (GO), using GO terms associated with the gene adjacent to each differentially-expressed consensus promoter. We observe significantly enriched GO categories for every comparison (data not shown). Results for the differentially-expressed genes between asexual and sexual females are summarized in Figure S5. Among asexual females, enriched categories among upregulated genes include *nitrogen compound metabolic process* (GO:0006807; $p < 1.2 \times 10^{-7}$). In sexual females (Figure S5), we observe enrichment of several GO categories, including (*hormone activity* (GO:0003735; $p < 0.014$) and *organic cyclic compound metabolic process* (GO:1901360; $p < 2.9 \times 10^{-6}$).

Differential upregulation of promoters of meiosis genes in asexual (parthenogenetic) females

We then asked whether there was evidence of enrichment of specific pathways within differentially-expressed promoters. Among genes upregulated in asexual females (vs. sexual females) (Figure 5B), we detect enrichment for pathways associated with cell cycle progression and oocyte meiosis (Figure S6), including *cell cycle* (04110 ;$p < 1.57 \times 10^{-5}$), *p53 signaling pathways* (04115;$p < 3.80 \times 10^{-3}$) and *oocyte meiosis* (04114;6.88×10^{-3}). Upon inspection of the genes associated with these terms, we observe substantial overlaps with annotated meiotic genes in *D. pulex*. From the differentially expressed genes associated with the *cell cycle* KEGG pathway (Figure S6), 5 out of 9 (Cdc20, CycA, CycB, CycE and Cdk2; 55.6%) are functionally designated as "meiotic" by at least one study (Schurko et al. 2009). Additionally, 3 of 7 upregulated genes within the *Oocyte meiosis* category (Cdc20, Cdk2 and CycE) are annotated in meiosis within *D. pulex* (Schurko et al. 2009), with two others (Plk1 (Pahlavan et al. 2000) and AurA (Crane et al. 2004)) being directly implicated in meiosis in other model systems. Given their positions within gene networks, upregulation of these genes would be expected to have a negative regulatory impact on meiotic progression overall. Relative expression of the detected promoters of meiosis genes among two comparisons: males vs. females and asexual females vs sexuals (*i.e.* males and sexual females), respectively, are shown (Figure 5E).

We investigated the set of upregulated genes in the (facultatively) asexual females within our study, asking about the extent of the concordance between the differentially-upregulated genes and scaffolds known to be physically linked to obligate asexuality (Tucker et al. 2013). Considering the genomic locations of differentially-upregulated genes, we unexpectedly find that a fraction (4/15 genes) are located on scaffolds linked to "asexual" chromosomes. This list includes Cdk2 (scaffold_77/ChrVIII), Tim-C (scaffold_76/ChrVIII), Plk1-C (scaffold_9/ChrIX) and HDAC (scaffold_13/ChrIX). We also note that two of the 15 genes, CycE (scaffold_163) and β-TrCP (scaffold_169) are located on short scaffolds that were not previously tested (Tucker et al. 2013).

Dramatic, sex-specific differential expression of a hemoglobin gene

In evaluating the differentially-expressed consensus promoter data (Figure 5A), we note several genes that are dramatically upregulated in a single condition. Among these is the *2-domain hemoglobin protein subunit* (ID:315053) gene on scaffold 13. We observe approximately 400-fold more CAGE tags at the promoter of this gene within sexual females than the other two states (males and asexual females) (Figure 6A and 6B), indicating considerable apparent state-specific upregulation of hemoglobin. The striking abundance of CAGE tags at the consensus promoter in sexual females (20,791 tpm) represents just over 2% of all sequenced CAGE tags within that state. An illustration of the core and proximal promoter region of the gene is shown in Figure 6C, including the consensus promoter region and major CTSS identified by this study. The core promoter contains a TATA box (5'-TATATA-3') at -27. We looked in the proximal promoter region for the juvenoid response element (JRE; 5'-CTGGTTA-3') identical to the one reported in *D. magna* (Gorr et al. 2006), but did not find one. We anticipate that future investigation will identify the precise cognate *cis*-regulatory elements within this region. An additional example of sex-specific expression is shown in Figure S4, where upregulation of the consensus promoter for the gene encoding the egg protein vitellogennin among asexual females is presented.

Discussion

In this study, we performed CAGE (Kodzius et al. 2006, Takahashi et al. 2012b) to map 5′-mRNA ends and identify active promoters within the ubiquitous aquatic microcrustacean *Daphnia pulex*, providing a taxonomic extension to the picture of metazoan promoter architecture. We report an average of 11,448 TSRs across the three conditions, 12,662 unique TSRs, and 10,580 consensus promoters. This *D. pulex* promoter atlas provides the first comprehensive collection of *cis*-regulatory elements within Crustacea.

We measured the occurrence of our CAGE-derived annotations with sites within the *D. pulex* genome, finding that they are generally located in positions consistent with promoter regions. The observation of CTSSs downstream of coding regions is consistent with the findings in *D. melanogaster*, where 17% of CAGE peaks were detected within annotated 3′UTR regions (Hoskins et al. 2011). The possible functions of CTSSs observed in CDSs and downstream of coding genes are challenging to interpret: they could represent the biochemical background of CAGE (Hoskins et al. 2011) or could alternatively represent *bona fide* RNA Pol II-derived transcripts. The latter case would suggest conflict with existing gene annotations, which can be resolved as more transcriptome analysis is performed in *D. pulex*. Approximately 82% of total aligned CAGE tags map upstream of annotated protein-coding genes (Figure 1C), a similar figure to that reported in *Drosophila* embryos (86%) (Hoskins et al. 2011). The overall incidence of TSRs upstream of coding genes (83%) mirrors that of CAGE tags (82.3%), suggesting that most TSRs in our dataset are positioned in locations consistent with the promoters of coding genes. The collection of TSRs (17%) located elsewhere is likely to contain a number of *bona fide* promoters.

The total number of unique TSRs defined here, 12,662, is close to the total of 12,454 promoters reported in *D. melanogaster* (Hoskins et al. 2011). This result may indicate a greater similarity in the number of protein coding genes between *D. pulex* and *D. melanogaster* than is presently predicted by the present genome annotation. The existing gene count for *D. pulex* (30,907) (Colbourne et al. 2011) is considerably larger than the approximately 17,000 currently annotated in *D. melanogaster*. The high depth of sampling and variety of stages measured in this study would be expected to reveal a similar ratio of active TSRs to annotated genes to what was observed in *D. melanogaster* (Hoskins et al. 2011). However, given the limited functional genomic evidence in *D. pulex* currently available, we cannot unequivocally conclude how many of the TSRs we report are, in fact, “true” promoters beyond evaluating their relationship to the current gene annotation. As it currently stands, this reality may lend greater weight to those TSRs that are found upstream of annotated coding genes. Further functional genomic (*e.g.* RNA-seq) analysis will be helpful to reconcile these existing discrepancies. We propose that the promoter atlas presented here be utilized to form an important component of a new, improved gene annotation in *D. pulex.*

We explored the properties of the consensus promoters within our *D. pulex* promoter atlas. The distribution of consensus promoter widths observed are consistent with what is seen in *D. melanogaster* (Figure 2A) (Hoskins et al. 2011, Chen et al. 2014). A proportion of the consensus promoter widths are long, including 1104 (10.4%) with widths longer than 30 bp (Figure 2A). This value is also similar to the amount observed (10.8%) in *D. melanogaster* (Hoskins et al. 2011). Promoters with long widths have also been observed in human, mouse (Carninci et al. 2006), and more recently, *C. elegans* (Saito et al. 2013). The distribution of consensus promoter shapes (Figure 2A, inset) indicates that both broad and peaked transcription initiation patterns are observed at *D. pulex promoters.* The observation that shape distribution is bimodal (Figure 2A, inset) agrees with previous models of promoter classes and provides rationale for the classification of promoters according to shape. We found that broad promoters exhibited higher promoter expression than did peaked promoters (Figure 2E), but we did not observe the same relationship between width and expression (data not shown). This suggests that shape is a more faithful representation of CTSS distribution and TSR properties than breadth alone. Our finding that broad promoters have higher promoter expression agrees with the available evidence in other species. In *D. melanogaster*, promoter width was positively associated with CAGE tag count (the equivalent to "expression" as defined here) (Hoskins et al. 2011). In *D. melanogaster* and elsewhere, *broad* promoters are associated with higher expression and genes with constitutive expression (Lenhard et al. 2012). While we did not directly address the relationship between promoter class and gene function in this study, such a comparison will be possible using these data, particularly as the functional annotation (*i.e* the Gene Ontology) of *D. pulex* genes improves.

We observe a strong preference for specific dinucleotides (CA, GA, GC, GG and GT) at CTSSs (Figure 2B). These results are partly in line with what is known elsewhere; the CA dinucleotide is located at [-1,+1] in Initiator (Inr)-containing promoters (Kadonaga 2012), and purines (A and G) are enriched at the TSS in metazoans, where studied (Nepal et al. 2013, Sandelin et al. 2007, Fitzgerald et al. 2006). However, three of the four over-represented dinucleotides (GA, GC, and GG) have guanines at

-1, which is observed less commonly in metazoans. *D. melanogaster*, the most closely related species for which CAGE data are available (Hoskins et al. 2011, Chen et al. 2014), is enriched for YR at [-1,+1]; no enrichment of dinucleotides with G at -1 is reported. In human, where core promoters tend to be GC-rich (Fitzgerald et al. 2004; 2006), YR, but no GN dinucleotides, are enriched at initiation sites (Frith et al. 2008, Sandelin et al. 2007).

Our data suggest the overall nucleotide preferences of *D. pulex* are unusual compared of other metazoans that have been similarly surveyed. We observe the CA dinucleotide at approximately 12% of CTSSs, which is identical to canonical YR code at initiation sites and agrees with the sequence of Initiator (Inr) at the [-1,+1] position (Butler and Kadonaga 2002). By contrast, the other four enriched dinucleotides reported here are observed less frequently at the initiation sites in other metazoans. This may suggest the presence of one or more alternative initiators in *D. pulex*. We exclude the trivial explanation, 5′ guanine addition bias sometimes observed in CAGE studies (Carninci et al. 2006), for the observed GN enrichment because these were corrected for by our analysis pipeline (see Methods).

Our *de novo* discovery revealed eight distinct enriched motifs that we call the *D. pulex* core promoter set (*Dpm1*-*Dpm7*; Figure 3). Of the eight *D. pulex* core promoter elements, three have significant sequence identity with a core promoter element in *D. melanogaster*. We find correspondence to major metazoan core promoter elements: *Dpm2*, with the consensus TATAWAA, displays similarity to the TATA element in *Drosophila* (TATAAA), and the consensus of the putative Inr motif *Dpm3* (NCAGT) has significant identity to the Initiator motif (Inr) of fruit fly, which is NCAKTY (Ohler et al. 2002) (Figure 4F). The putative TATA *Dpm2* and Inr *Dpm3* are enriched between -30 and +1 (Figure 4B), respectively, consistent with their positions elsewhere within metazoans (Juven-Gershon and Kadonaga 2010). This strongly suggests that we have identified the TATA and Initiator motifs in *D. pulex*. The motif *Dpm5* (TGGCAAC), observed at 15.3% of promoters, bears significant identity to the Ohler8 motif (–YGGCARC–) in *D. melanogaster* (Ohler et al. 2002). *Dpm5* is enriched at approximately +50 (Figure 4D); the *D. melanogaster* Ohler8 motif has an equivalent, but more modest, peak at the same position (Down et al. 2007). The *cis*-regulatory role of Ohler8 is unknown, but it has been validated separately on several occasions since its initial discovery (Fitzgerald et al. 2006, Hoskins et al. 2011). In our study, the Ohler8-like *Dpm5* motif was observed in a smaller fraction of promoters than observed in *D. melanogaster* (15.3% vs. 23.2%) (Ohler et al. 2002).

The remainder of the *Daphnia* promoter motif set is less well-characterized. The five other motifs within our *D. pulex* core promoter set, *Dpm1*, *Dpm6*, *Dpm7* and *Dpm8* (Figure 3), lack similarity to any member of the core promoter list in *D. melanogaster*. Two of these exhibit a degree of positional enrichment relative to the TSS. *Dpm1* is enriched broadly between approximately -40 and -75. *Dpm4* exhibits a sharp positional enrichment at -10, and a second, wider distribution surrounding -50. No positional enrichment was observed among *Dpm6*, *Dpm7*, and *Dpm8* (Figure 4D and data not shown), suggesting that they lack location preferences within core promoter regions.

The core promoter motif discovery described in this study is the first comprehensive glimpse into the *cis*-regulatory repertoire of *D. pulex*, and indeed for any crustacean. We observe strong cognates to core promoter elements in more well-studied metozoan genomes, including *D. melanogaster*. Collectively, these data support a model for the composition of the *D. pulex* core promoter (Figure 4E). Comparisons between our *D. pulex* core promoter model and the established model in *D. melanogaster* highlight the similarity of the reported TATA and Inr elements between the two species, but also underscores the absence of two canonical fly core promoter elements (BRE and DPE) (Butler and Kadonaga 2002) in our set of core promoters (Figure 4F). A finely-tuned motif discovery approach that selects only specific promoter classes (*e.g.* only Inr-containing promoters) is necessary as it would be more suited for discovery of BRE and DPE, which are less abundant than TATA and Inr.

In total, 3 of 8 *Dpm* motifs identified by our study lack obvious homologs in *Drosophila*. While we cannot propose precise functions for these putative core promoter elements, the overall positional enrichment and motif co-occurrence data (Figure 4A–4D) suggests that core promoters in *D. pulex* may group into TATA and TATA-less categories. In *D. melanogaster*, promoters that contain TATA, Inr, and a small number of other elements (including Pause Button, which we do not find in our set) are very likely to exhibit a peaked shape (Hoskins et al. 2011). By contrast, broad promoters are depleted for TATA and Inr (Rach et al. 2009, Hoskins et al. 2011); in mammals, they are associated with CpG Islands (Lenhard et al. 2012). Our finding that TATA and Inr-containing promoters have a more peaked shape than TATA-less promoters (Figure 4G) is consistent with this model. A complete characterization of the relationship between core promoter (*i.e. Dpm*) motif composition (especially TATA and Inr) and TSR shape and expression will require further analysis of the evidence generated in this study.

D. pulex is an important model in which to study the maintenance of sexual and asexual reproduction (Hebert 1981, Tucker et al. 2013); we analyzed the genes associated with differentially-expressed promoters observed between asexual females and sexual females (Figure 5C and 5D) and both sexuals (sexual females and adult males; Figure S3). Our observation of strong enrichment cell-cycle pathways (KEGG IDs: 04110 and 04115) among genes upregulated in asexual females (Figure S6) was unexpected. Upon closer inspection, we find strong overlap between genes in these categories and those belonging to two enriched meiosis-related pathways (*Progesterone-mediated oocyte maturation* (04914) and *Oocyte meiosis* (04114); a number have been annotated as meiotic in *D. pulex* (Schurko et al. 2009). The observation of upregulated meiosis genes in asexual females (Figure 5E) was surprising, but is consistent with what is known about the functions of some of the genes in question. The most compelling of these examples is Cdc20 (ID:326123; NCBI_GNO_7600067), which is more than two-fold upregulated (169.4tpm to 76.2tpm) in asexual females. In mammals, Cdc20 acts with the APC to trigger progression through prophase during Meiosis I (Homer et al. 2009). Increased expression of Cdc20 would be expected to hasten the exit from Meiosis I-like cell-division. Cdc20 misexpression is known to disrupt Meiosis I; mice hypomorphic for Cdc20 were shown to be infertile (or nearly so) due to chromosomal lagging and mis-alignment during Meiosis I (Jin et al. 2010).

Although we lack comparable sources of expression data in *Daphnia*, the apparent increase in Cdc20 expression we observe here parthenogeneisis is consistent with current model of parthenogenic oogenesis in *D. pulex*, which is known to consist of abortive Meiosis I followed by a normal, Meiosis II-like division (Hiruta et al. 2010). We posit that the apparent differential regulation of meiosis and cell-cycle genes observed here is evidence for the transcriptional changes to meiosis that accompany parthenogenesis in *D. pulex*. However, it must be emphasized that additional molecular and cytological work will be required to appropriately address this question.

Finally, the identity and genomic position of several genes upregulated in asexual females on scaffolds associated with the evolution of asexuality (Figure 5E) is worth noting. Among these are Cdc20 (scaffold 76/ChrVIII) and HDAC (scaffold 13/ChrIX), two genes that were recently shown to be strongly upregulated in cyclic parthenogenesis (relative to obligate parthenogenesis) in Bdelloid rotifers (Hanson et al. 2013).

Taken together, our large-scale analysis of transcription initiation in the microcrustacean *D. pulex* provides the first glimpse of *cis*-regulation and core promoter architecture in Crustacea. We find that *D.*

pulex exhibits similar features of promoter architectures relative to fly and mammals, including peaked promoters associated with TATA and Inr and constitutively-expressed broad promoters. We also detect major constituents of its core promoter that lack an obvious ortholog in fly, suggesting some degree of novelty within the core promoter of *D. pulex*. It is intended that the data presented here, including the *D. pulex* promoter atlas described here, serve as a resource for future investigations within *D. pulex*, and comparative genomic analysis across metazoan diversity. We anticipate that, using this resource, comparisons between *D. pulex* and the fruit fly and fellow arthropod *D. melanogaster*, which are ˜600My diverged (Hedges et al. 2006), will be of particular utility.

Methods

Focal genotype and maintenance of individuals

The *Daphnia pulex* genotype used in this work was isolated from Portland Arch Pond (Warren County, Indiana, USA; geographic coordinates: 40.2096°, -87.3294°) and is identified as PA13-42 (hereafter PA42). The PA42 clone originates from a well-characterized natural population (Lynch et al. 1989). *D. pulex* individuals from the PA42 clone are cyclical parthenogens, meaning that they are capable of reproducing both asexually through eggs that develop directly or sexually through diapausing eggs. All individuals used in this study were the result of asexual reproduction. Females were maintained in 3L containers containing COMBO media (Kilham et al. 1998) (diluted 1:1 with water) at 20°C and fed *Scenedesmus* at approximately 100,000 cells/mL. New offspring were removed and placed in new containers daily. Asexual females, pre-ephipial (sexual) females, and males were isolated from culture on separate occasions using strainers of differential sizes and visual identification under a dissecting microscope. Males can be visually distinguished from females based on the criteria of enlarged atennules and flattened ventral carapace margin. The current reproductive mode of females can be determined by phenotyptic differences in yolk-filled ovaries: females currently reproducing asexually have more "bulbous" ovaraies that tend to be more green in color, while females currently reproducing sexually have blackish yolks of reduced size and a smoother external margin.

RNA isolation and quantification

Whole *D. pulex* individuals (approximately 50–75) were collected from fresh cultures from each of the three aforementioned states. Collections were homogenized manually using a small pestle in microcentrifuge tubes containing lysis buffer. Isolation of total RNA was performed using solid phase extraction (Bioline, Inc). Samples were snap-frozen in liquid nitrogen and stored at -80°C. RNA samples were quantified and evaluated for quality and using the Bioanalyzer 2100 (Agilent Technologies).

CAGE library preparation and sequencing

A multiplexed CAGE library was constructed as described (Takahashi et al. 2012a) from 5μg total RNA sample using the nAnT-iCAGE protocol (Murata et al. 2014) (K. K. DNAForm, Yokohama, Japan). Briefly, total RNA was reverse transcribed using a random "N6 plus base 3" primer (TCTNNNNNN), using SuperScript III reverse transcriptase (Thermo Fisher). Following oxidation (with sodium peroxide) and biotinylation of the ^{m7}G cap structures, 1st-strand-complete mRNA:cDNA hybrids were bound with streptavadin beads, pulled down with a magnet, and released. This was followed by ligation of the 5′ linker, which includes the 3nt barcode (*e.g.* iCAGE_01 N6 5′-CGACGCTCTTCCGATCTACCNNNNNN-3′) followed by 3′ linker ligation. Finally, 2nd-strand synthesis was performed using the nAnT-iCAGE 2nd primer (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT-3′), creating the final dsDNA product. For a more detailed protocol, please see the following: (Murata et al. 2014).

qRT-PCR evaluation of CAGE libraries

Prior to sequencing, relative mRNA:rRNA ratios were measured for each CAGE library using quantitative reverse-transcriptase PCR (qRT-PCR) with SYBR Green I (Life Technologies). The control gene GAPDH (glyceraldehyde-3-phosphate dehydrogenase) was selected, (Forward Primer: 5′-ACCACTGTCCATGCCATCACT-3′, Reverse Primer: 5′-CACGCCACAACTTTCCAGAA-3′) and was measured against 18S ribosomal mRNA (Forward Primer: 5′-CCGGCGACGTATCTTTCAA-3′, Reverse Primer: 5′-CACGCCACAACTTTCCAGAA-3′). Biological replicates of each of the three states were reflected in the final CAGE library (n=3 for both female groups, n=2 for males). Finally, the completed CAGE library was sequenced using Illumina HiSeq2000 (single-end, 50bp reads) at the University of

California, Berkeley Genome Sequencing Laboratory (Berkeley, CA, USA).

CAGE processing, alignment, and rRNA filtering

All CAGE-adapted sequence reads (1.82×10^8) were demultiplexed (`http://hannonlab.cshl.edu/fastx_toolkit/index.html`), creating eight separate fastq files corresponding to the original CAGE libraries. All CAGE-adapted sequences (47bp) from each library were aligned separately using bwa (Li and Durbin 2009) to the *D. pulex* assembly v1.1 (JGI) (Colbourne et al. 2011). Prior to downstream analysis, CAGE alignments (in .bam format) were subjected to a filtering step (rRNAdust; `http://fantom.gsc.riken.jp/5/sstar/Protocols:rRNAdust`) to remove rRNA sequences (28S, 18S, and 5S). The SAM flags of identified rRNA reads in the alignment were changed to “unmapped”. Overall, 1.22×10^8 CAGE reads (67.0% of the total) mapped successfully (Table 1), and these were utilized in subsequent analyses. Evaluations of CAGE alignments and pooling of multiple libraries was performed using Samtools (Li et al. 2009). The distribution of CAGE tags within the *D. pulex* genome was determined using BEDtools (Quinlan 2014). Non-overlapping genomic intervals were created using BEDtools from the Joint Genome Institute’s (JGI) *Frozen Gene Catalog* annotation (“FrozenGeneCatalog20110204.gff3”) located at `http://genome.jgi.doe.gov/Dappu1/Dappu1.download.html`.

Analysis of mapped CAGE tags

TSRs were defined from mapped CAGE tags using the CAGEr package (Haberle et al. 2015) in R Bioconductor (Huber et al. 2015). Aligned reads from each library were normalized by fitting to a power law distribution as described (FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). The 5′ coordinate (CAGE adapter-adjacent) of each aligned read was designated as a CTSS, and the CAGE tag abundance at each genomic position was quantified in tags per million (tpm). CTSSs with CAGE tag support above 2 tpm (significant CTSSs; sCTSSs) were clustered into TSRs using the *distclu* algorithm in CAGEr, which merges sCTSSs below a maximum distance of 20bp apart. Correlation of sCTSS abundance across biological replicates showed extremely high within-sample concordance (R^2 >0.97). TSR *width* was defined as the length of the genomic segment occupied by sCTSSs within a TSR. Where specified, we calculated TSR width using the interquantile range using the “quantilePositions” function in CAGEr (Haberle et al. 2015). We selected the interquantile range between the 10th and 90th percentile of all CAGE signal within a TSR, where the n^{th} percentile refers to the genomic position where $n\%$ of the CAGE signal is $5'$ of the entirety of the CAGE signal within a TSR (Haberle et al. 2015). The CAGE mapping pipeline is available in Supplementary Scripts (CAGE_Promoter_mapping_TCO.R).

Promoter definitions

TSRs were reported for a given condition if the evidence from all replicates were in agreement (n=3 for sexual females and asexual females, n=2 for males). Consensus promoters are the genomic coordinates of promoters found in all CAGE datasets and were calculated using interquantile widths (10^{th} - 90^{th}) using CAGEr (Haberle et al. 2015). CAGE definitions are illustrated in Figure 1.

Classification of promoter shape

We measured promoter shape by calculating the diversity of CTSSs within a given TSR or consensus promoter. To do this, we applied the Shape Index (SI) as described (Hoskins et al. 2011), which is itself based on the Shannon entropy (Shannon 1948). The Shape Index is calculated as follows using TSSs within a given promoter:

$$SI = 2 + \sum_{i}^{L} p_i log_2 p_i,$$

where p is the probability of CTSS position i being observed among all L CTSS positions within the TSR (or consensus promoter). TSRs that contain a single unique CTSS position will have a Shape Index equal to 2, while the Shape Index value becomes more negative as the number of distinct CTSSs within the TSR increases.

TSRs and consensus promoters were labeled as either *broad* (SI <-2), *peaked* (SI >1.5), or *unclassified* (all others) according to their associated SI values.

Test for bimodality of TSR shapes

We tested the calculated shape value (in units of SI, as described above) of all consensus promoters for bimodality. The distribution of shape values was evaluated using the Expectation-Maximization (EM) algorithm implemented in the Mixtools package (Benaglia et al. 2009) in R. The results support a 2-component mixture within the distribution. Fitted Gaussian densities of the two components (shaded in coral and blue, respectively) were plotted against the overall distribution of calculated consensus promoter shapes (Figure 1A, inset).

Dinucleotide preference at initiation sites

Dinucleotide frequencies were calculated using bedtools nuc (Quinlan 2014) from 2bp intervals (position: [-1,+1]) created from i) CTSSs and ii) randomly sampled background intervals derived from the *D. pulex* genome. A statistical test of the observed dinucleotide preferences was performed by repeating this procedure iteratively for all consecutive dinucleotides within the the [-1,-100] window (the control) relative to +1, and evaluating the the resulting dinucleotide frequencies observed for each. Dinucleotide frequencies within the window [-1,+1] relative to CTSSs were considered significant if they fell in the top or bottom 5 (0.05) of all control observations. We did not test dinucleotide frequencies downstream of +1 in our test to avoid the potential confounding effects of codon bias.

De novo motif discovery

Daphnia core promoter motifs were discovered using hypergeometric enrichment in Homer (Heinz et al. 2010). This procedure was performed as follows: first, CAGE peaks (using the peak-finding algorithm of Homer) from pooled (*i.e.* in all three states) alignments were detected using annotatePeaks.pl to create a peak interval file. Next, we retrieved motifs that were enriched within 150bp sequences ([-100,+50]) surrounding the CAGE peaks relative to background (findMotifsGenome.pl). We searched for motifs of 6, 8, 10 and 12bp, reflecting the typical size range of *cis*-regulatory motifs.

Statistical validation of predicted *de novo* motifs

Promoter motifs were determined using 10-fold cross-validation. The CAGE peak position file was divided into ten folds (subsamples) of equal size. For each round of validation, one of the folds was labeled as the test set, and the other nine were identified as the training set. This process was iterated ten times, such that each fold served as the test set exactly once. *De novo* motif prediction was performed on each of the ten training sets using Homer as described above.

We evaluated motifs within all ten training sets by measuring the consistency with which a motif is found within a training set. For example, if a given motif is found only in a handful of the ten training sets, it is unlikely to be a *bona fide* core promoter motif. Predicted motifs from each of the ten training sets were grouped and clustered according to their pairwise distance (Pearson correlation coefficient) using the Tomtom module (Gupta et al. 2007) of the MEME Suite package (Bailey et al. 2015). To group identical motifs within the training set, we generated a graph with the python module "NetworkX" (Schult and Swart 2008) from the significant hits between motifs from the Tomtom output, with each pairwise match between motifs becoming an undirected edge. We identified connected components containing 8 or more nodes, and selected all motifs associated with these. Eight groups met this criteria; these were used to build corresponding 8 motif sets. Finally, PWMs from each motif set were aligned (MotifSetReduce.pl; see Supplementary Scripts)) to create a single consensus PWM, generating 8 motifs overall. These consensus PWMs were designated **D**aphnia (core) **p**romoter **m**otif (Dpm). Motif logos were generated for each Dpm PWM using the motif2Logo.pl function in Homer. The similarity of the each member of the Dpm motif set to core promoter elements in *D. melanogaster* was determined by sequence alignment STAMP (Mahony and Benos 2007) against the JASPAR database (Portales-Casamar et al. 2009). The E-value of the best alignment was recorded for every Dpm motif. The enrichment score of a representative PWM from the motif set was selected to reflect each Dpm motif in Table 1.

Differential expression analysis

Differential expression of promoters was performed using defined consensus promoters (n=10,665) along with their normalized expression values (in tpm) observed in each condition. We utilized the most recent version of the *limma* package in R (Ritchie et al. 2015) to determine the differentially-expressed promoters across all three conditions. *Limma*, which implements a linear modeling algorithm, also incorporates *voom* (variance modeling at the observational level), a method that estimates the mean-variance relationship in a counts-based fashion (Law et al. 2014).

Analysis of mean-variance and linear model

Genomic coordinates and expression values (in tpm) for all consensus promoters within a library were used to construct an ExpressionSet object (Lawrence and Morgan 2014) in R. Biological replicates from a given stage were labeled and used to construct a "contrasts matrix" to establish comparisons between stages (*i.e.* males - sexual females). Analysis of mean variance (voom) was performed for every consensus promoter containing more than 25 tags (TSSs) on aggregate across all CAGE libraries. The log-ratios from the previous step were fit to a linear model (lmFit; (Ritchie et al. 2015)), followed by a "contrasts fit" using the aforementioned contrasts matrix, which calculates the standard error for each *contrast*, or between-stage comparison. An empirical Bayes method (*ebayes*) was applied to the model fits from the previous step, generating moderated t- and F-statistics, respectively, and a log-odds differential expression value for each consensus promoter. A *decide test* was then performed on this set of t-statistics, where consensus promoters with p-values below 0.01 (after Benjamini & Hochberg FDR correction) were deemed to be significantly differentially-expressed (DE). DE promoters from each comparison were retrieved for subsequent analysis.

Visualization of differentially-expressed genes

Heatmaps: The normalized expression levels (in all CAGE libraries) of promoters classified as differentially-expressed were extracted and plotted as a hierarchically-clustered heatmaps in R using the gplots package (Warnes et al. 2015).

Analysis of functional enrichment

Gene Ontology

Consensus promoters were associated with genes (Frozen Gene Catolog) using their genomic coordinates. The complete gene ontology (GO) dataset for *D. pulex* (`http://genome.jgi.doe.gov/cgi-bin/ToGo?species=Dappu1`) was downloaded and GO terms were associated with the gene annotation. We applied the Fisher's Exact Test in the topGO package (Alexa and Rahnenfuhrer 2010) in R, asking which GO terms were over-represented among genes shown to have differentially-regulated promoters (see previous section). Enrichment analysis was performed separately using terms from the GO categories *Molecular Function* and *Biological Process*, respectively. GO Terms with p-values less than 0.01 were classified as "significantly enriched".

Pathway Analysis

We extracted the KEGG (Kyoto Encyclopedia of Genes and Genomes; `http://www.genome.jp/kegg/`) pathway identifier, using the same promoter-to-gene-annotation dataset described for the GO analysis.

Using the set of terms for differentially-expressed consensus promoters, we performed a test for statistical enrichment of KEGG pathways using the Python tool PEAT (C. Jackson et al. 2016, In Preparation). KEGG terms with p-values below 0.01 were considered significantly enriched.

Data Access

CAGE sequence data in this manuscript have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) (`http://www.ncbi.nlm.nih.gov/geo`) and will be made public immediately after publication of the submitted manuscript.

Acknowledgments

We thank Peter Cherbas and Sen Xu for critical comments to the manuscript, and Teresa Crease for feedback regarding the methodology. We would like to thank Kim Young for her work culturing *Daphnia* collections. We thank Xiangyu Yao for contributing to our motif discovery workflow. This work was supported by a grant-in-aid to ML from the NIH (Identifier: 1R01GM101672-01A1). This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303.

Author Contributions

RTR co-conceived the idea, performed the computational analyses and experimental work and wrote the paper. KS performed experimental and culturing work and contributed to the paper. VPB developed the idea, contributed to the paper and edited the paper. ML conceived the idea and edited the paper.

Disclosure Declaration

The authors declare that they have no competing interests.

References

Alexa A and Rahnenfuhrer J. 2010. *topGO: topGO: Enrichment analysis for Gene Ontology.* R package version 2.20.0.

Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, and Sandelin A. 2014. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Communications* **5**: 5336.

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al.. 2015. Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science (New York, NY)* **347**: 1010–1014.

Bailey TL, Johnson J, Grant CE, and Noble WS. 2015. The MEME Suite. *Nucleic Acids Research* **43**: W39–W49.

Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, and van Nimwegen E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biology* **10**: R79.

Benaglia T, Chauveau D, and Hunter D. 2009. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* **32**: 1—-29.

Butler JEF and Kadonaga JT. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development* **16**: 2583–2592.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al.. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**: 626–635.

Chen ZX, Sturgill D, Qu J, Jiang H, Park S, Boley N, Suzuki AM, Fletcher AR, Plachetzki DC, Fitzgerald PC, et al.. 2014. Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Research* **24**: 1209–1223.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al.. 2011. The ecoresponsive genome of Daphnia pulex. *Science (New York, NY)* **331**: 555–561.

Cosma MP. 2002. Ordered recruitment: Gene-specific mechanism of transcription activation. *Molecular Cell* **10**: 227–236.

Crane R, Gadea B, Littlepage L, Wu H, and Ruderman JV. 2004. Aurora A, meiosis and mitosis. *Biology of the Cell* **96**: 215–229.

Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, et al.. 2008. Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nature Methods* **5**: 629–635.

Down TA, Bergman CM, Su J, and Hubbard TJP. 2007. Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Computational Biology* **3**: e7.

Ebert D. 2005. Ecology, epidemiology, and evolution of parasitism in Daphnia. `http://www.ncbi.nlm.nih.gov/books/NBK2036/`. Online; accessed 12 April 2016.

Edgar R, Domrachev M, and Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**: 207–210.

FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

Fitzgerald PC, Shlyakhtenko A, Mir AA, and Vinson C. 2004. Clustering of DNA sequences in human promoters. *Genome Research* **14**: 1562–1574.

Fitzgerald PC, Sturgill D, Shyakhtenko A, Oliver B, and Vinson C. 2006. Comparative genomics of Drosophila and human core promoters. *Genome Biology* **7**: R53.

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, and Sandelin A. 2008. A code for transcription initiation in mammalian genomes. *Genome Research* **18**: 1–12.

Gorr TA, Rider CV, Wang HY, Olmstead AW, and LeBlanc GA. 2006. A candidate juvenoid hormone receptor cis-element in the Daphnia magna hb2 hemoglobin gene promoter. *Molecular and cellular endocrinology* **247**: 91–102.

Gupta S, Stamatoyannopoulos JA, Bailey TL, and Noble WS. 2007. Quantifying similarity between motifs. *Genome Biology* **8**: R24.

Haag CR, McTaggart SJ, Didier A, Little TJ, and Charlesworth D. 2009. Nucleotide polymorphism and within-gene recombination in Daphnia magna and D. pulex, two cyclical parthenogens. *Genetics* **182**: 313–323.

Haberle V, Forrest AR, Hayashizaki Y, Carninci P, and Lenhard B. 2015. CAGEr: precise tss data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Research* **43**: e51.

Haberle V, Li N, Hadzhiev Y, Plessy C, Previti C, Nepal C, Gehrig J, Dong X, Akalin A, Suzuki AM, et al.. 2014. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**: 381–385.

Hanson SJ, Stelzer CP, Welch DBM, and Logsdon JM. 2013. Comparative transcriptome analysis of obligately asexual and cyclically sexual rotifers reveals genes with putative functions in sexual reproduction, dormancy, and asexual egg production. *BMC Genomics* **14**: 412.

Hebert PDN. 1981. Obligate Asexuality in Daphnia. *The American Naturalist* **117**: 784–789.

Hedges SB, Dudley J, and Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics (Oxford, England)* **22**: 2971–2972.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, and Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**: 576–589.

Hiruta C, Nishida C, and Tochinai S. 2010. Abortive meiosis in the oogenesis of parthenogenetic Daphnia pulex. *Chromosome Research* **18**: 833–840.

Homer H, Gui L, and Carroll J. 2009. A spindle assembly checkpoint protein functions in prophase I arrest and prometaphase progression. *Science (New York, NY)* **326**: 991–994.

Hoskins RA, Hoskins RA, Landolin JM, Landolin JM, Brown JB, Brown JB, Sandler JE, Sandler JE, Takahashi H, Takahashi H, et al.. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research* **21**: 182–192.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, et al.. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12**: 115–121.

Jin F, Hamada M, Malureanu L, Jeganathan KB, Zhou W, Morbeck DE, and van Deursen JM. 2010. Cdc20 Is Critical for Meiosis I and Fertility of Female Mice. *PLoS genetics* **6**: e1001147.

Juven-Gershon T and Kadonaga JT. 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology* **339**: 225–229.

Kadonaga JT. 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**: 40–51.

Kilham SS, Kreeger DA, Lynn SG, Goulden CE, and Herrera L. 1998. COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia* **377**: 147–159.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al.. 2006. CAGE: cap analysis of gene expression. *Nature Methods* **3**: 211–222.

Kurosawa J, Nishiyori H, and Hayashizaki Y. 2011. Deep cap analysis of gene expression. *Methods in Molecular Biology (Clifton, NJ)* **687**: 147–163.

Law CW, Chen Y, Shi W, and Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**: R29.

Lawrence M and Morgan M. 2014. Scalable Genomics with R and Bioconductor. *Statistical Science* **29**: 214–226.

Lenhard B, Sandelin A, and Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* **13**: 233–245.

Li H and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis GR, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078–2079.

Lynch M, Spitze K, and Crease T. 1989. The Distribution of Life-History Variation in the Daphnia pulex Complex. *Evolution* **43**: 1724–1736.

Mahony S and Benos PV. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research* **35**: W253–8.

Murata M, Nishiyori-Sueki H, Kojima-Ishiyama M, Carninci P, Hayashizaki Y, and Itoh M. 2014. Detecting Expressed Genes Using CAGE. In *Transcription Factor Regulatory Networks*, pp. 67–85. Springer New York, New York, NY.

Nepal C, Hadzhiev Y, Previti C, Haberle V, Li N, Takahashi H, Suzuki AMM, Sheng Y, Abdelhamid RF, Anand S, et al.. 2013. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Research* **23**: 1938–1950.

Ohler U, Liao Gc, Niemann H, and Rubin GM. 2002. Computational analysis of core promoters in the Drosophila genome. *Genome Biology* **3**: research0087.1–0087.12.

Pahlavan G, Polanski Z, Kalab P, Golsteyn R, Nigg EA, and Maro B. 2000. Characterization of Polo-like Kinase 1 during Meiotic Maturation of the Mouse Oocyte. *Developmental Biology* **220**: 392–400.

Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, and Sandelin A. 2009. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* **38**: D105–D110.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics* **47**: 11.12.1–11.12.34.

Rach EA, Yuan HY, Majoros WH, Tomancak P, and Ohler U. 2009. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biology* **10**: R73.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* p. gkv007.

Saito TL, Hashimoto Si, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, and Morishita S. 2013. The transcription start site landscape of C. elegans. *Genome Research* **23**: 1348–1361.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, and Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics* **8**: 424–436.

Saxonov S, Berg P, and Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 1412–1417.

Schult DA and Swart P. 2008. Exploring network structure, dynamics, and function using NetworkX. In *7th Python in Science Conferences (SciPy)* (eds. G Varoquaux, T Vaught, and J Millman), pp. 11–16.

Schurko AM, Logsdon JM, and Eads BD. 2009. Meiosis genes in Daphnia pulex and the role of parthenogenesis in genome evolution. *BMC Evolutionary Biology* **9**: 78.

Shannon CE. 1948. A mathematical theory of communication. *The Bell System Technical Journal* **27**: 379–423, 623–656.

Takahashi H, Kato S, Murata M, and Carninci P. 2012a. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods in Molecular Biology (Clifton, NJ)* **786**: 181–200.

Takahashi H, Lassmann T, Murata M, and Carninci P. 2012b. 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols* **7**: 542–561.

Tucker AE, Ackerman MS, Eads BD, Xu S, and Lynch M. 2013. Population-genomic insights into the evolutionary origin and fate of obligately asexual Daphnia pulex. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 15740–15745.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, et al.. 2015. *gplots: Various R Programming Tools for Plotting Data.* R package version 2.17.0.

Figure Legends

Figure 1: TSS profiling in *D. pulex* using CAGE. **A**. Schematic of CAGE annotations. i) Individual sequenced CAGE tags (represented by short, horizontal black lines) are aligned to the genome in a strand-specific manner, ii) defining distinct CTSSs (represented by dark blue vertical lines). iii) CTSSs with CAGE tag support above 2 tpm are spatially clustered into TSRs (indicated with red lines). CTSSs (gray vertical lines) below the 2 tpm threshold are ignored during this clustering step and are not included in the eventual TSRs. iv) TSRs with evidence across three states are classified as consensus promoters. **B**. A summary of the developmental stages surveyed in this study. We sequenced CAGE-adapted cDNA libraries originating in i) sexual females, ii) males, and iii) asexual females. The life cycle of *D. pulex* is summarized (left panel), showing the parthenogenic (ameiotic) and sexual (meiotic) cycles. A representative visualization of CAGE tag densities for a single promoter region across the three states is presented at right. The illustration of the *Daphnia* life cycle in the left panel is adapted from an illustration by Dita B. Vizoso (Freiburg University) in (Ebert 2005), and is used with permission. **C**. Proportions of CAGE annotations by genomic location. Locations of all aligned CAGE tags, CTSSs and TSRs by genome segment are shown, including 1kb upstream of the CDS (orange), 1kb downstream of CDS (red), within the CDS (light yellow), CDS introns (light blue) and within intergenic (*i.e.* exclusive to the other categories) regions (dark blue).

Figure 2: **A**. Distributions of consensus promoter width and shape in the *D. pulex* promoter atlas. A histogram representing the distribution of calculated consensus promoter (n=10,580) widths is shown in orange (outer figure). Each bin width represents 5 bp. *Inset*: Consensus promoter shapes have a bimodal distribution. A histogram representing the shapes (measured with the Shape Index (SI)) of all consensus promoters (n=10,580) is shown in white, with each bin indicating 0.1 of a SI. The densities of broad (coral) and peaked (royal blue) consensus promoter shapes were fitted from the overall distribution of SI values (see Methods). **B**. Distinct dinucleotide preferences at transcription initiation sites in *D. pulex.* The dinucleotide frequencies at CTSS ([-1,+1]; aqua) are compared to background (coral). CTSSs show a two-fold or greater preference for the dinucleotides CA, GA, GC and GT, and are similarly depleted for AA, AT and TT. **C**. Representative examples of canonical CAGE tag distribution patterns observed in *D. pulex* consensus promoters. Peaked consensus promoters (above) exhibit narrow CAGE tag distributions, whereas broad consensus promoters (below) are typified by more a dispersed distribution of CAGE tags. **D**. Consensus promoter expression correlates with shape more strongly than width. Consensus promoter expression (measured according to total number of CAGE tags) is plotted against TSR width in base pairs (bp). Peaked (SI >1), broad (SI <-1.5) and unclassified (all other) consensus promoters are identified by green, red, and blue circles, respectively. **E**. Broad consensus promoters have greater expression than peaked consensus promoters. A significantly greater number of CAGE tags are observed in broad consensus promoters relative to peaked consensus promoters (*p <0.0005; Tukey's HSD). Box-and-whisker plots representing the distributions of the consensus promoter expression in three shape classes (broad: red, peaked: green, and unclassified: blue) are shown.

Figure 3: *De novo* discovery of core promoter elements in *D. pulex.* The *D. pulex* core promoter motifs identified in this study are listed. For each identified motif (n=8) we show a logo representing the PWM of each motif, its frequency relative to regions surrounding major CAGE peaks (-200,+50) (see Methods), observed motif enrichment E-value, and the E-value of the most similar motif within the JASPAR database (Portales-Casamar et al. 2009). The motif enrichment E-value represents the probability that a motif of equal length would be discovered in an equivalent number of randomly-derived sequences with the same underlying nucleotide frequencies with equal or lower likelihood.

Figure 4: The co-occurrence and distribution of identified *D. pulex* core promoter motifs within promoter regions. **A**. Heatmap of co-occurrence frequencies among identified *D. pulex* motifs. The log of each p-value is plotted within the heatmap. The frequency distributions of *Dpm2* and *Dpm3* (**B**), *Dpm1* and *Dpm4*, (**C**) and *Dpm5* and *Dpm6* (**D**) relative to identified promoters (TSRs) are shown. (The distributions of *Dpm7* and *Dpm8* are not shown.) **E**. Current model of core promoter composition in *D. pulex* derived from the evidence in this study. A cartoon illustration of the *Daphnia* core promoter motifs that exhibit strong positional distributions are shown, with their approximate locations relative to the TSS (+1). **F**. Model representing the positions and consensus sequences of canonical core promoter elements between *D. pulex* and *D. melanogaster*. The four major core promoter elements in *D. melanogaster* are displayed, along with their typical positions relative to the TSS (+1). The consensus sequence of each element, if present, is shown for *D. melanogaster* (Dm; red) and *D. pulex* (Dp; dark blue). Note that an individual core promoter may have none, all, or some of the elements listed in the illustration. Graphic adapted from (Butler and Kadonaga 2002). **G**. Comparison of promoter shape between TATA and Inr-containing promoters and those lacking TATA. The box-and-whisker plots representing the distributions of calculated shape index (SI) values for consensus promoters with Inr (coral), TATA (green) and those lacking TATA (blue) are shown. Initiator (**) and TATA-containing (*) consensus promoters possess a significantly more peaked shape ($p < 0.001$) than TATA-less promoters.

Figure 5: Differential expression analysis of *D. pulex* consensus promoters. **A**. Representation of consensus promoter expression among the states surveyed in this study. A scatterplot of consensus promoter expression (in tpm) within all three states measured within our study is shown, with the value for asexual females (x-axis) plotted against sexual females (y-axis). Corresponding expression values for males are represented according to a color gradient in log-scale. A small number of consensus promoters (n=145) that lie outside the area of the are not shown. **B**. Barplot representing the number of differentially-expressed (p <0.01) consensus promoters observed within each of five comparisons (see Methods). Bars representing upregulated consensus promoters are shown in red; down-regulated promoters are yellow in color (below). **C**. Mean-average (MA) plot of consensus promoter expression of within asexual compared to sexual females. Mean average expression of consensus promoters (x-axis) is plotted against the log fold-change (FC) of the ratio of the expression of consensus promoters between asexual females and sexual females (y-axis). Differentially-expressed consensus promoters (p <0.01 are represented by red dots; all others are colored in black. Upper and lower blue lines on the plot indicate the log(FC) of 2 and -2, respectively. **D**. Heatmap of the expression of differentially-expressed (p <0.01) consensus promoters between asexual females and sexual females. **E**. Heatmap grid of relative expression of consensus promoters of *D. pulex* meiosis genes within two selected comparisons: males vs. females and asexual females vs. sexuals. Cells are shaded according to the calculated t-statistic of a given comparison. Instances of significant differential expression (p <0.01) are labeled with two asterisks (**).

Figure 6: Extreme upregulation observed at the putative promoter of a hemoglobin gene in *D. pulex* sexual females. **A**. Mapped CAGE tags from each of the three surveyed states to an annotated hemoglobin gene (ID:315053) on scaffold 13 are shown. The frequency of CAGE tags observed at each genomic coordinate (x-axis) are indicated by the y-axes of each plot. Note that larger y-axis scales are applied for the sexual females plot due to the dramatically higher number of mapped CAGE tags observed at the same locus. **B**. Consensus promoter expression (number of CAGE tags in tpm) at the same genomic locus as Part A across all three states is presented in the left panel; in the right panel only the values for males and asexual females are shown to provide perspective. The standard error of the mean of all replicates is shown for each individual plot. **C**. Schematic illustration of the core and proximal promoter region of the hemoglobin gene (ID:315053). The major CTSS (+1) is identified by the blue arrow, and the TATA consensus sequence is represented by the red rectangle. The purple line represents the consensus promoter region identified by CAGE. The genomic coordinates for the sequences (all on Scaffold 13) are shown in black. Note that the sequence for the negative strand is shown; the illustration was flipped to improve legibility. The drawing was not made to scale.

Supplementary Figure Legends

Figure S1: Correlation between within our CAGE experiment. A matrix containing pairwise comparisons of individual CAGE libraries (n=8) is shown. Multiple individual scatterplots are presented (lower-left), which compares CAGE tag count per CTSS within each comparison. In the upper-right portion of the matrix the Pearson correlation coefficient of each individual comparison is shown. Individual experimental samples are colored and labeled along the diagonal of the matrix.

Figure S2: Multi-dimensional scaling (MDS) plot of the CAGE samples within this experiment. Distances between each sample are reported in terms of leading log-fold-changes between each pair of samples. The identity of each CAGE sample is labeled directly on the plot.

Figure S3: Additional heatmaps of differentially-expressed (p <0.01) consensus promoters are shown. **A**. Males vs. sexual females. **B**. Males vs. asexual females. **C**. Males vs. females (*i.e.* asexual females and sexual males). **D**. Asexuals vs. sexuals (*i.e.* males and sexual females).

Figure S4: State-specific upregulation of the gene of the precursor egg protein vitellogennin (VTG) in asexual females. The number of aligned CAGE tags observed upstream of the VTG gene (ID:322419) are plotted within the genomic region that surrounds the VTG gene (scaffold_47:116142-127656) for all three states. The number of CAGE tags at each position are represented by the y-axis of each plot, respectively.

Figure S5: Gene Ontology (GO) categories that are enriched among genes whose consensus promoters are significantly (p <0.01) i) upregulated in asexual females and ii) upregulated in sexual females within our study. Gene Ontology IDs and Pathway Names are shown.

Figure S6: KEGG pathways enriched among the genes of consensus promoters upregulated within asexual females (vs. sexual females). Meiosis-related pathways are shaded in gray, along with number of genes expected and observed within each pathway, and the corresponding odds ratio and p-value.

Figure S7: Subgraphs of the most enriched GO terms found in differentially-expressed genes found sexual and asexual females. Rectangles represent the five most significantly enriched GO terms, and are color-coded from least significant (yellow) to most significant (red). Circular nodes represent GO terms within the GO semantic hierarchy. General information about each node is printed within each node, including the GO ID, a brief descriptor, the calculated p-value and the number of genes containing each individual term. *A*. GO Molecular Function (MF) categories enriched in asexual females. *B*. GO Biological Process (BP) categories enriched in asexual females. *C*. GO Molecular Function (MF) categories enriched in sexual females. *D*. GO Biological Process (BP) categories enriched in asexual females.

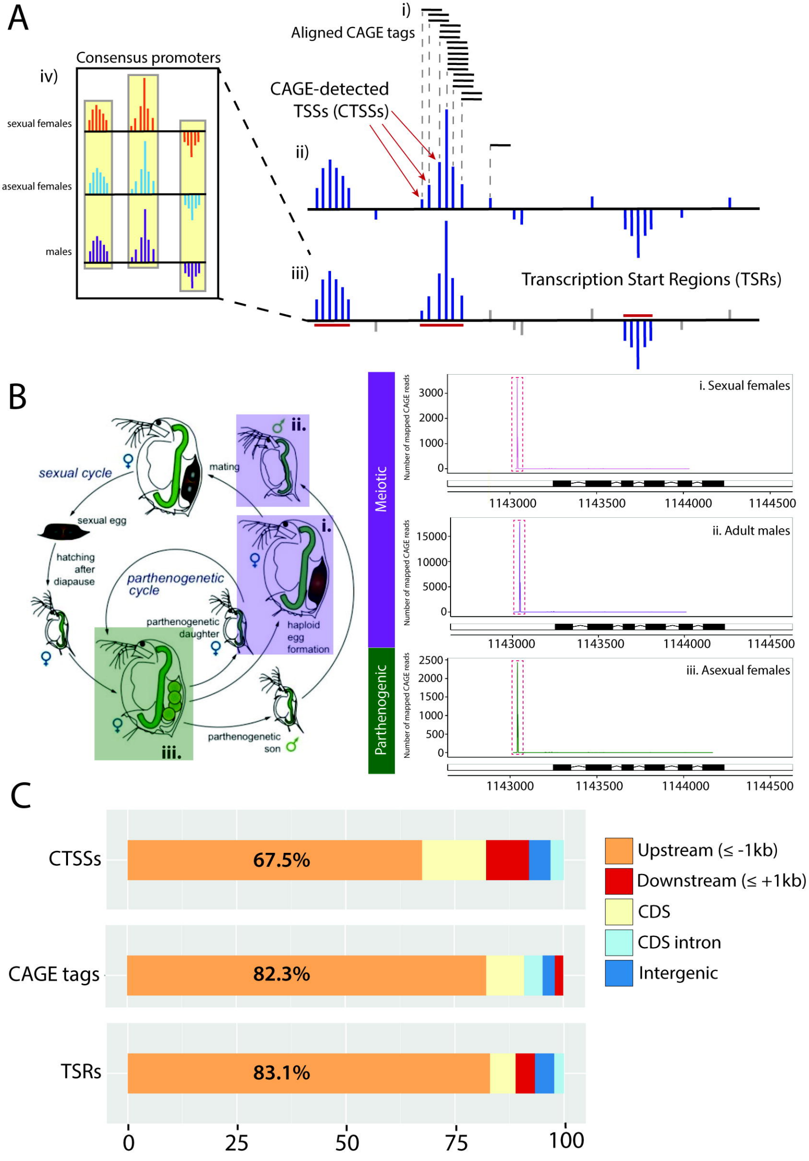
Tables

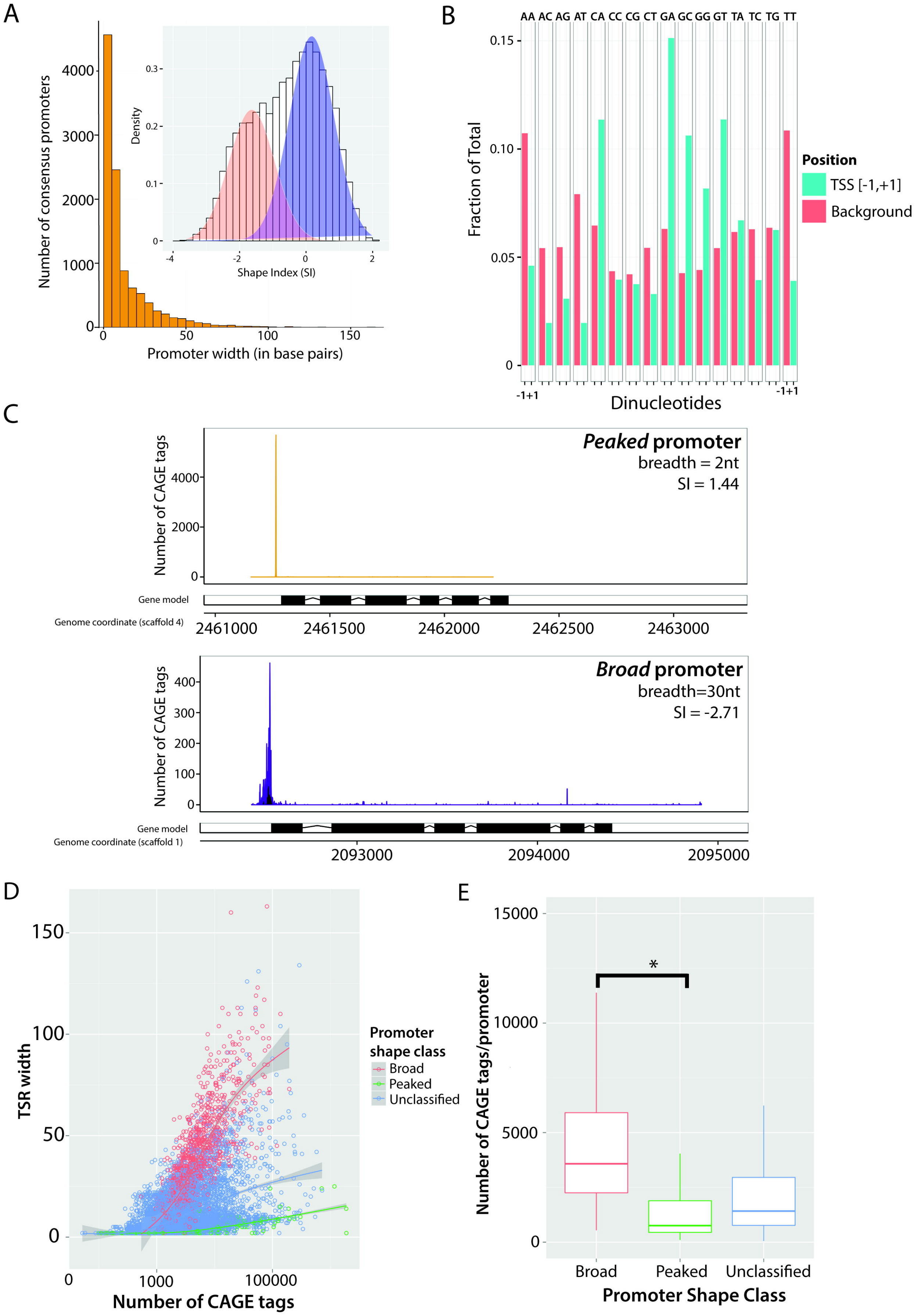
Table 1: Summary of CAGE libraries in this study. The value at the end of each library name refers to the biological replicate number.

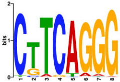
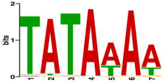
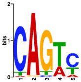
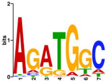
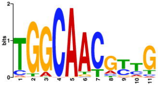
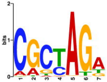
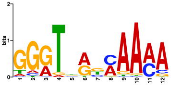
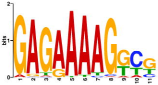
Number	Library Name	Number of Sequenced CAGE Tags	Number of Mapped CAGE Reads
1	Asexual females-1	28,803,508	18,601,744
2	Asexual females-2	16,701,216	10,839,287
3	Asexual females-3	29,786,273	20,754,759
4	Sexual females-1	24,076,420	15,861,581
5	Sexual females-2	24,567,545	15,163,393
6	Sexual females-3	15,115,501	9,621,093
7	Males-1	18,512,317	12,412,516
8	Males-2	24,655,373	16,960,704
Total	–	182,218,153	120,215,077

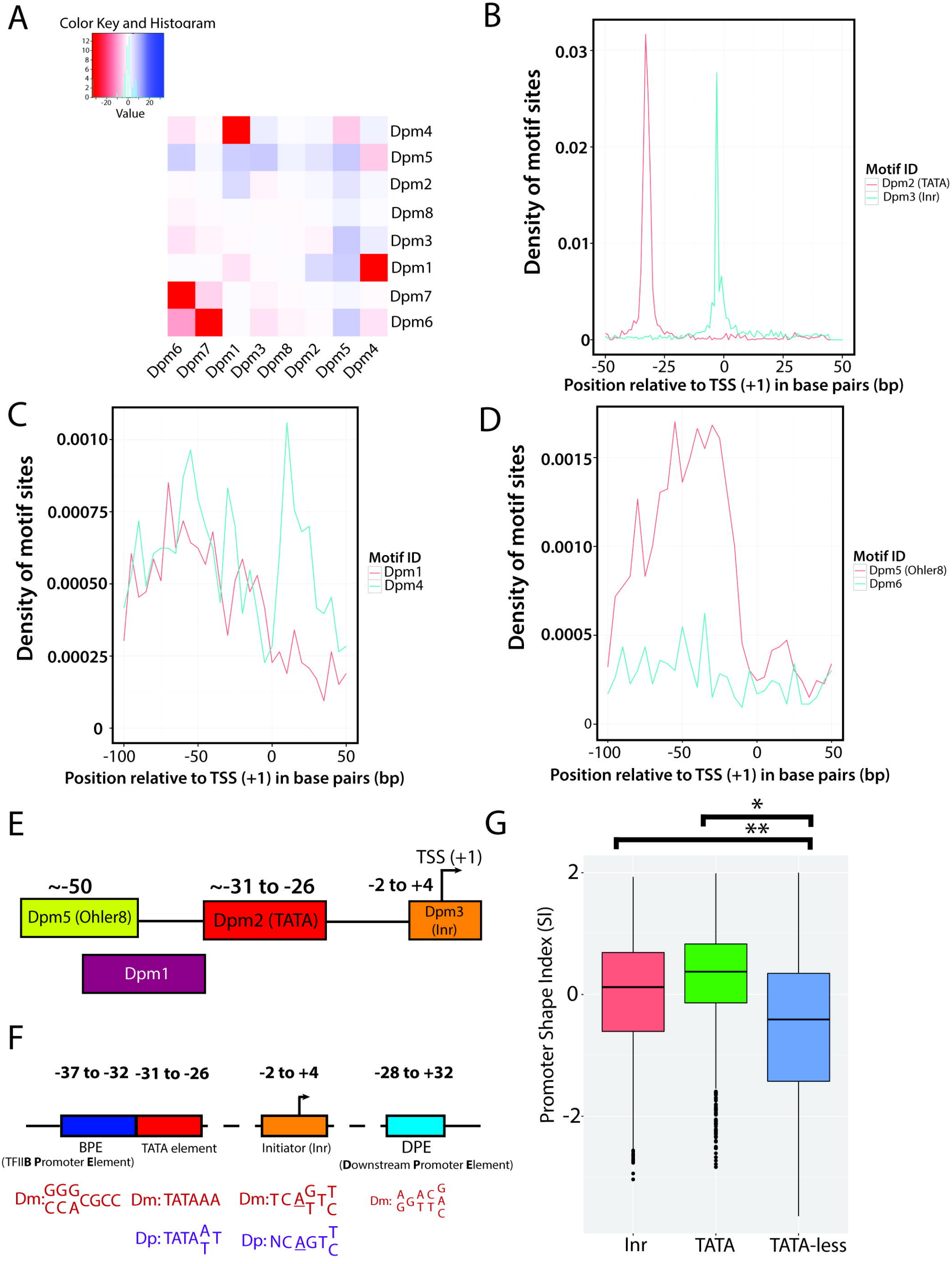
Table 2: Summary of CAGE evidence generated in this study.

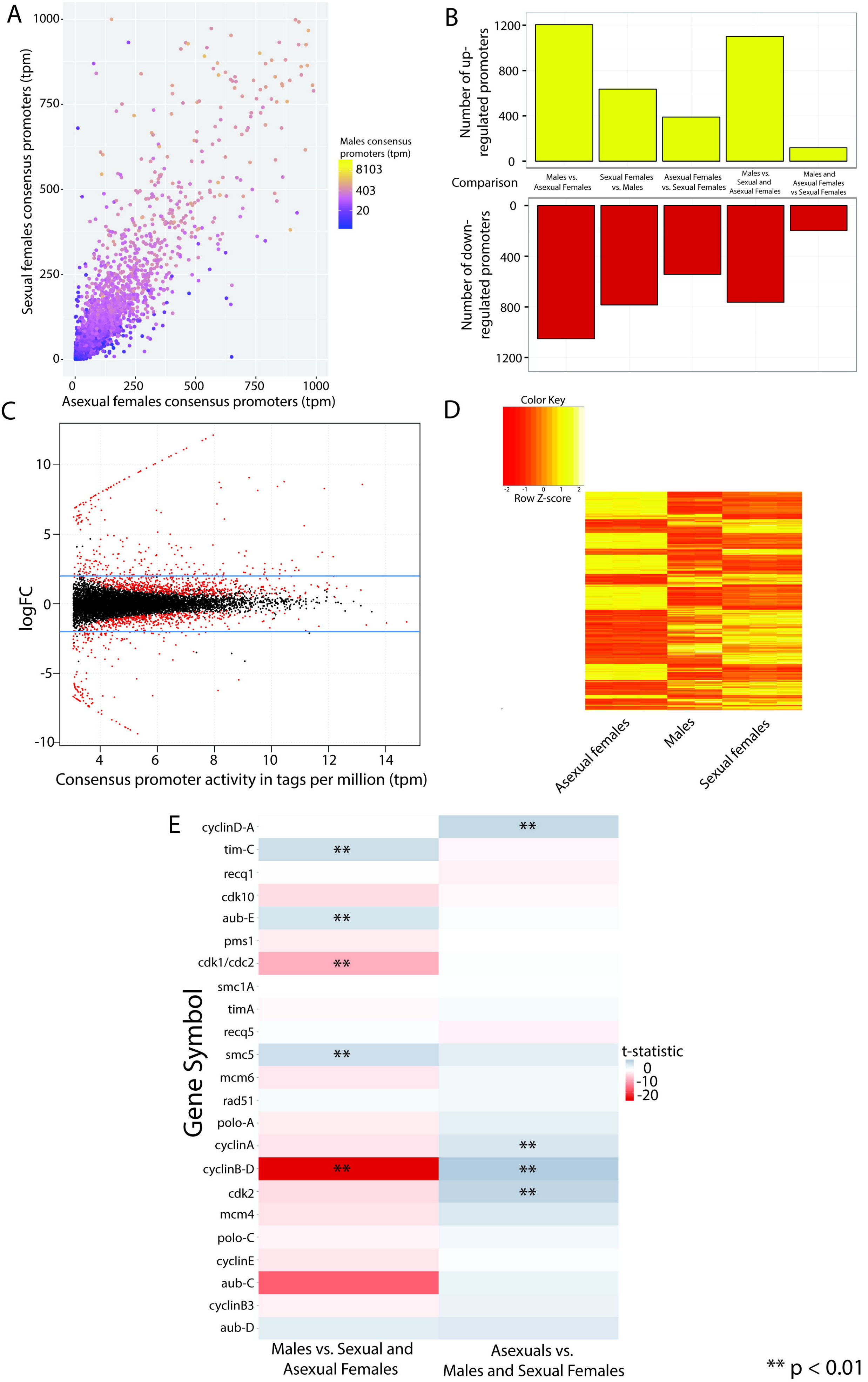
Sample Name	Number of Mapped Reads	TSRs (unique)	Consensus Promoters
Asexual females	50,195,790	11,496 (316)	–
Sexual females	40,646,067	11,289 (231)	–
Males	29,373,220	11,558 (557)	–
Total	120,215,077	12,662	10,665

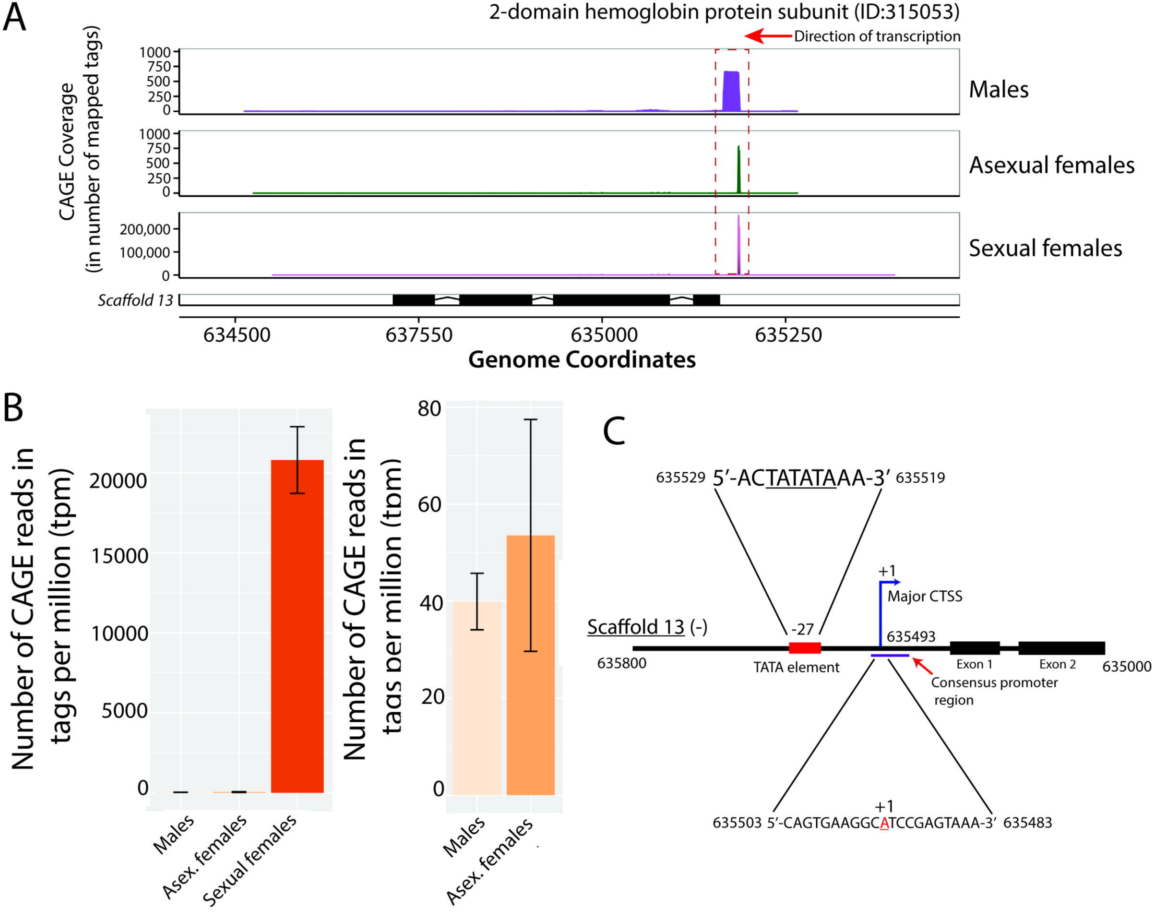

A
i)
Aligned CAGE tags
Consensus promoters
iv)
sexual females
asexual females
males
CAGE-detected
TSSs (CTSSs)
ii)
iii)
Transcription Start Regions (TSRs)
B
sexual cycle
mating
sexual egg
hatching
after
diapause
parthenogenetic
cycle
parthenogenetic
daughter
haploid
egg
formation
parthenogenetic
son
i.
ii.
iii.
Meiotic
Parthenogenic
Number of mapped CAGE reads
i. Sexual females
3000
2000
1000
0
1143000
1143500
1144000
1144500
ii. Adult males
15000
10000
5000
0
iii. Asexual females
2500
2000
1500
1000
500
C
CTSSs
67.5%
CAGE tags
82.3%
TSRs
83.1%
0
25
50
75
100
Upstream (≤ -1kb)
Downstream (≤ +1kb)
CDS
CDS intron
Intergenic

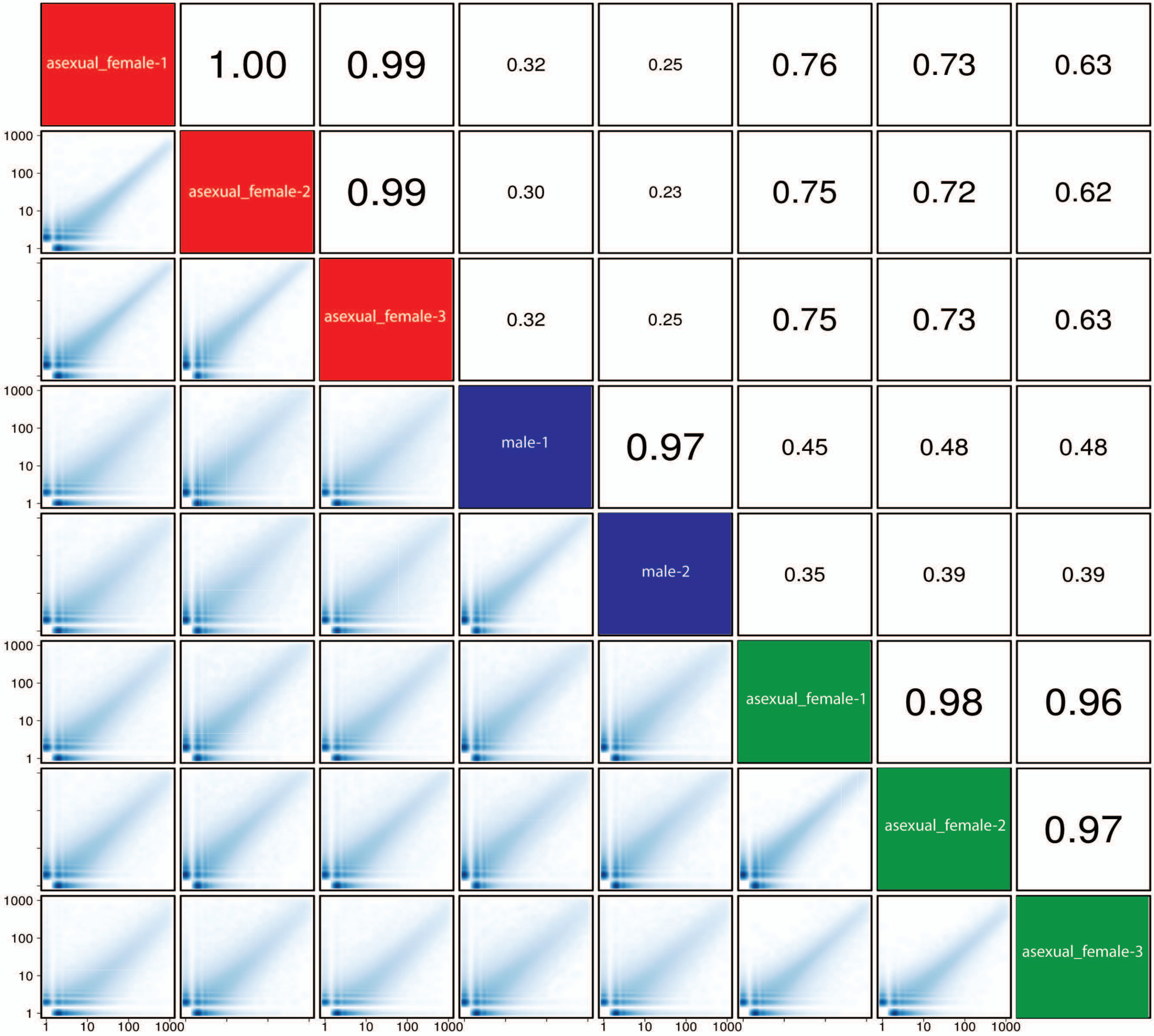

A
Number of consensus promoters
0
1000
2000
3000
4000
0
50
100
150
Promoter width (in base pairs)
Density
0
0.1
0.2
0.3
-4
-2
0
2
Shape Index (SI)
B
AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT
Fraction of Total
0
0.05
0.10
0.15
-1+1
-1+1
Dinucleotides
Position
TSS [-1,+1]
Background
C
Number of CAGE tags
0
2000
4000
Peaked promoter
breadth = 2nt
SI = 1.44
Gene model
Genome coordinate (scaffold 4)
2461000
2461500
2462000
2462500
2463000
Number of CAGE tags
0
100
200
300
400
Broad promoter
breadth=30nt
SI = -2.71
Gene model
Genome coordinate (scaffold 1)
2093000
2094000
2095000
D
TSR width
0
50
100
150
0
1000
100000
Number of CAGE tags
Promoter shape class
Broad
Peaked
Unclassified
E
Number of CAGE tags/promoter
0
5000
10000
15000
*
Broad
Peaked
Unclassified
Promoter Shape Class

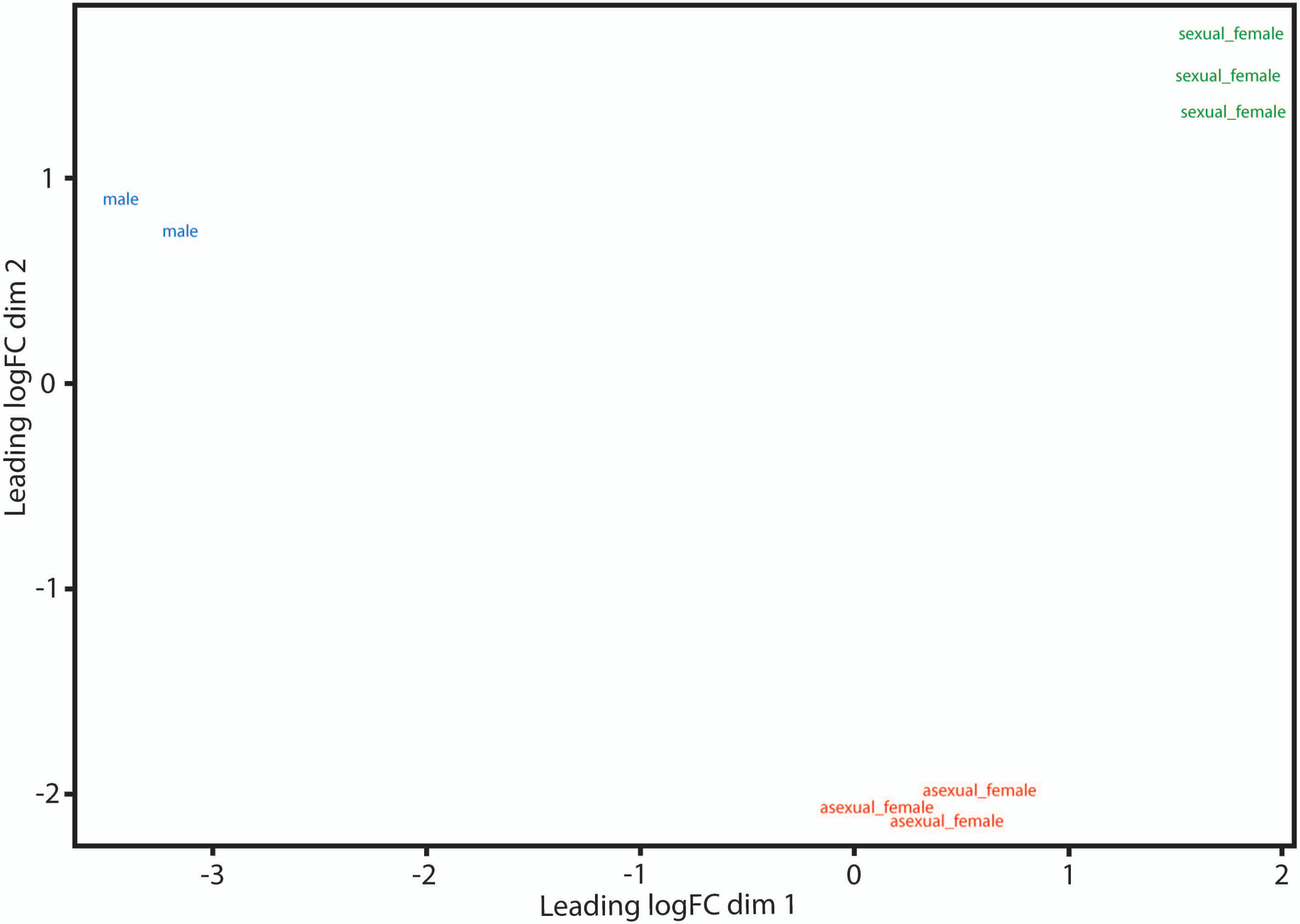
Motif ID	**Motif logo**	**Occ. (%)**	**Enrichment**	**Best match** (JASPAR)	**E-value**
Dpm1		9.48	$1e^{-17}$	MA0154.1_EBF1	**2.67×10^{-5}**
Dpm2		22.62	$1e^{-308}$	MA0108.1_TBP	**6.19×10^{-9}**
Dpm3		12.04	$1e^{-283}$	MA0092.1_Hand1_T	**2.36×10^{-3}**
Dpm4		11.95	$1e^{-84}$	MA0139.1_CTCF	**4.12×10^{-5}**
Dpm5		15.28	$1e^{-147}$	MA0365.1_RFX1	**2.12×10^{-6}**
Dpm6		6.86	$1e^{-45}$	MA0266.1_ABF2	**5.51×10^{-5}**
Dpm7		4.11	$1e^{-33}$	MA0009.1_T	**1.74×10^{-4}**
Dpm8		4.63	$1e^{-33}$	MA0326.1_MAC1	**2.38×10^{-4}**

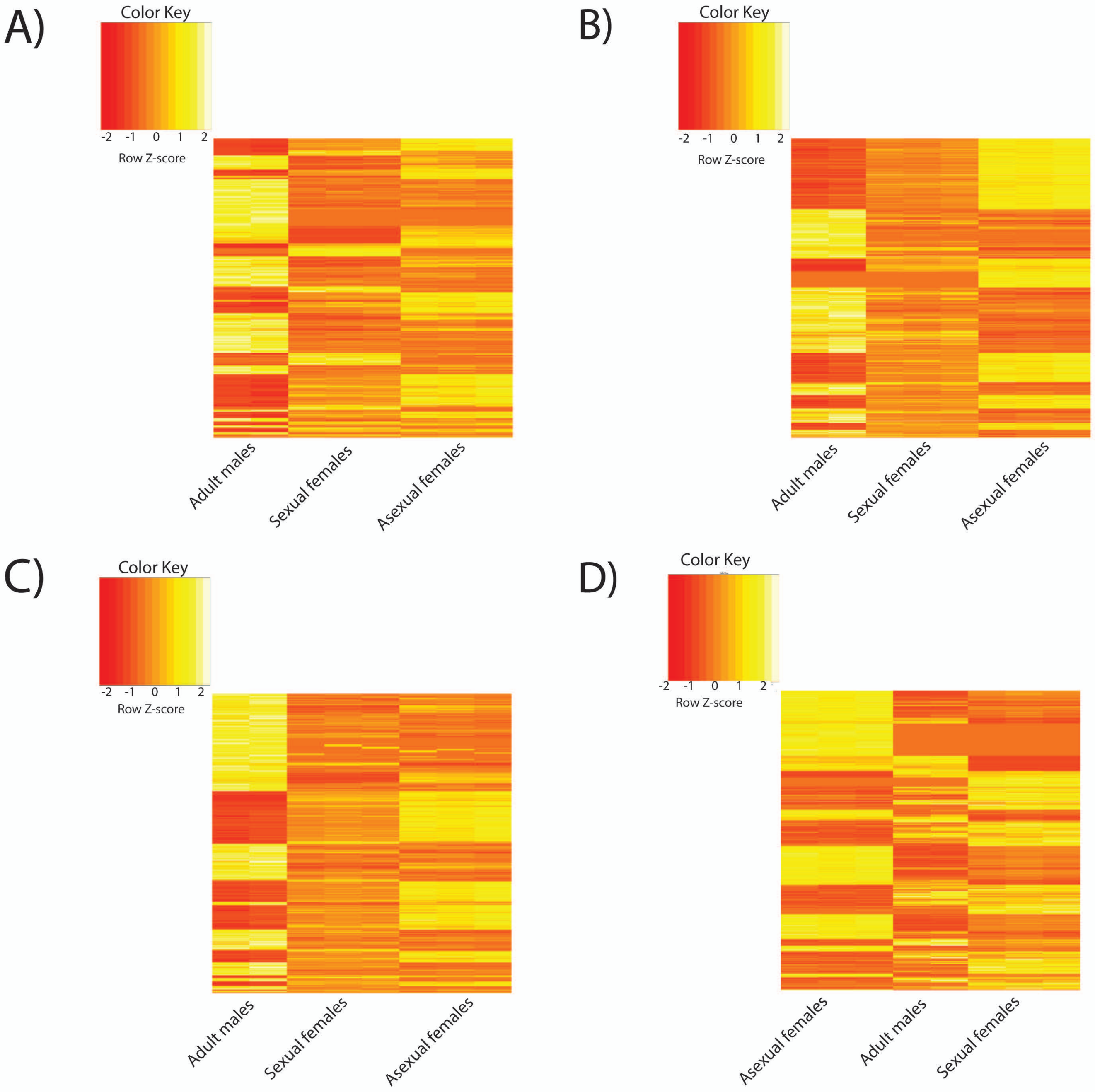

A
Color Key and Histogram
Value
Dpm4
Dpm5
Dpm2
Dpm8
Dpm3
Dpm1
Dpm7
Dpm6
Dpm6 Dpm7 Dpm1 Dpm3 Dpm8 Dpm2 Dpm5 Dpm4
B
Density of motif sites
Position relative to TSS (+1) in base pairs (bp)
Motif ID
Dpm2 (TATA)
Dpm3 (Inr)
C
Density of motif sites
Position relative to TSS (+1) in base pairs (bp)
Motif ID
Dpm1
Dpm4
D
Density of motif sites
Position relative to TSS (+1) in base pairs (bp)
Motif ID
Dpm5 (Ohler8)
Dpm6
E
~-50
~-31 to -26
-2 to +4
TSS (+1)
Dpm5 (Ohler8)
Dpm2 (TATA)
Dpm3 (Inr)
Dpm1
F
-37 to -32
-31 to -26
-2 to +4
-28 to +32
BPE
(TFIIB Promoter Element)
TATA element
Initiator (Inr)
DPE
(Downstream Promoter Element)
Dm: GGG/CCA CGCC
Dm: TATAAA
Dm: TCA G/T T T/C
Dm: A/G G A/T C/T G/A/C
Dp: TATA A/T T
Dp: NCAGT T/C
G
Promoter Shape Index (SI)
*
**
Inr
TATA
TATA-less

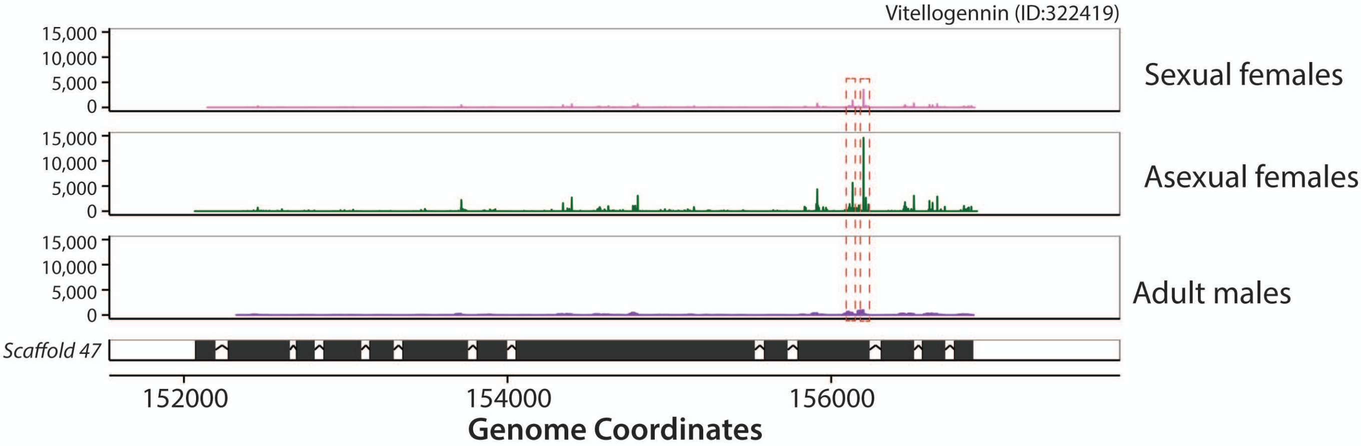

A
Sexual females consensus promoters (tpm)
1000
750
500
250
0
Asexual females consensus promoters (tpm)
0
250
500
750
1000
Males consensus promoters (tpm)
8103
403
20
B
Number of up-regulated promoters
1200
800
400
0
Comparison
Males vs. Asexual Females
Sexual Females vs. Males
Asexual Females vs. Sexual Females
Males vs. Sexual and Asexual Females
Males and Asexual Females vs Sexual Females
Number of down-regulated promoters
0
400
800
1200
C
logFC
10
5
0
-5
-10
4
6
8
10
12
14
Consensus promoter activity in tags per million (tpm)
D
Color Key
-2
-1
0
1
2
Row Z-score
Asexual females
Males
Sexual females
E
Gene Symbol
cyclinD-A
tim-C
recq1
cdk10
aub-E
pms1
cdk1/cdc2
smc1A
timA
recq5
smc5
mcm6
rad51
polo-A
cyclinA
cyclinB-D
cdk2
mcm4
polo-C
cyclinE
aub-C
cyclinB3
aub-D
**
t-statistic
0
-10
-20
Males vs. Sexual and Asexual Females
Asexuals vs. Males and Sexual Females
** p < 0.01


A
2-domain hemoglobin protein subunit (ID:315053)
Direction of transcription
CAGE Coverage
(in number of mapped tags)
1000
750
500
250
0
Males
1000
750
500
250
0
Asexual females
200,000
100,000
0
Sexual females
Scaffold 13
634500
637550
635000
635250
Genome Coordinates
B
Number of CAGE reads in
tags per million (tpm)
20000
15000
10000
5000
0
Males
Asex. females
Sexual females
Number of CAGE reads in
tags per million (tpm)
80
60
40
20
0
Males
Asex. females
C
635529
5'-ACTATATAAA-3'
635519
+1
Major CTSS
Scaffold 13 (-)
-27
635493
635800
TATA element
Exon 1
Exon 2
635000
Consensus promoter
region
+1
635503
5'-CAGTGAAGGCATCCGAGTAAA-3'
635483


asexual_female-1
1.00
0.99
0.32
0.25
0.76
0.73
0.63
1000
100
10
1
asexual_female-2
0.99
0.30
0.23
0.75
0.72
0.62
asexual_female-3
0.32
0.25
0.75
0.73
0.63
1000
100
10
1
male-1
0.97
0.45
0.48
0.48
male-2
0.35
0.39
0.39
1000
100
10
1
asexual_female-1
0.98
0.96
asexual_female-2
0.97
1000
100
10
1
asexual_female-3
1
10
100
1000
1
10
100
1000
1
10
100
1000
1
10
100
1000


sexual_female
sexual_female
sexual_female
male
male
asexual_female
asexual_female
asexual_female
1
0
-1
-2
-3
-2
-1
0
1
2
Leading logFC dim 2
Leading logFC dim 1


A)
Color Key
-2 -1 0 1 2
Row Z-score
Adult males
Sexual females
Asexual females
B)
Color Key
-2 -1 0 1 2
Row Z-score
Adult males
Sexual females
Asexual females
C)
Color Key
-2 -1 0 1 2
Row Z-score
Adult males
Sexual females
Asexual females
D)
Color Key
-2 -1 0 1 2
Row Z-score
Asexual females
Adult males
Sexual females



Vitellogennin (ID:322419)
15,000
10,000
5,000
0
Sexual females
15,000
10,000
5,000
0
Asexual females
15,000
10,000
5,000
0
Adult males
Scaffold 47
152000
154000
156000
Genome Coordinates

i) Upregulated in asexual females

Molecular Function

GO ID	Pathway Name (KEGG ID)	*Corr. p*-value
GO:0003735	structural constituent of ribosome	0.015

Biological Process

GO ID	Pathway Name (KEGG ID)	*Corr. p*-value
GO:0006807	nitrogen compound metabolic process	1.2x10^{-7}
GO:0044281	small molecule metabolic process	6.8x10^{-6}
GO:0044763	single-organism cellular process	3.0x10^{-5}

ii) Upregulated in meiotic females

Biological Process

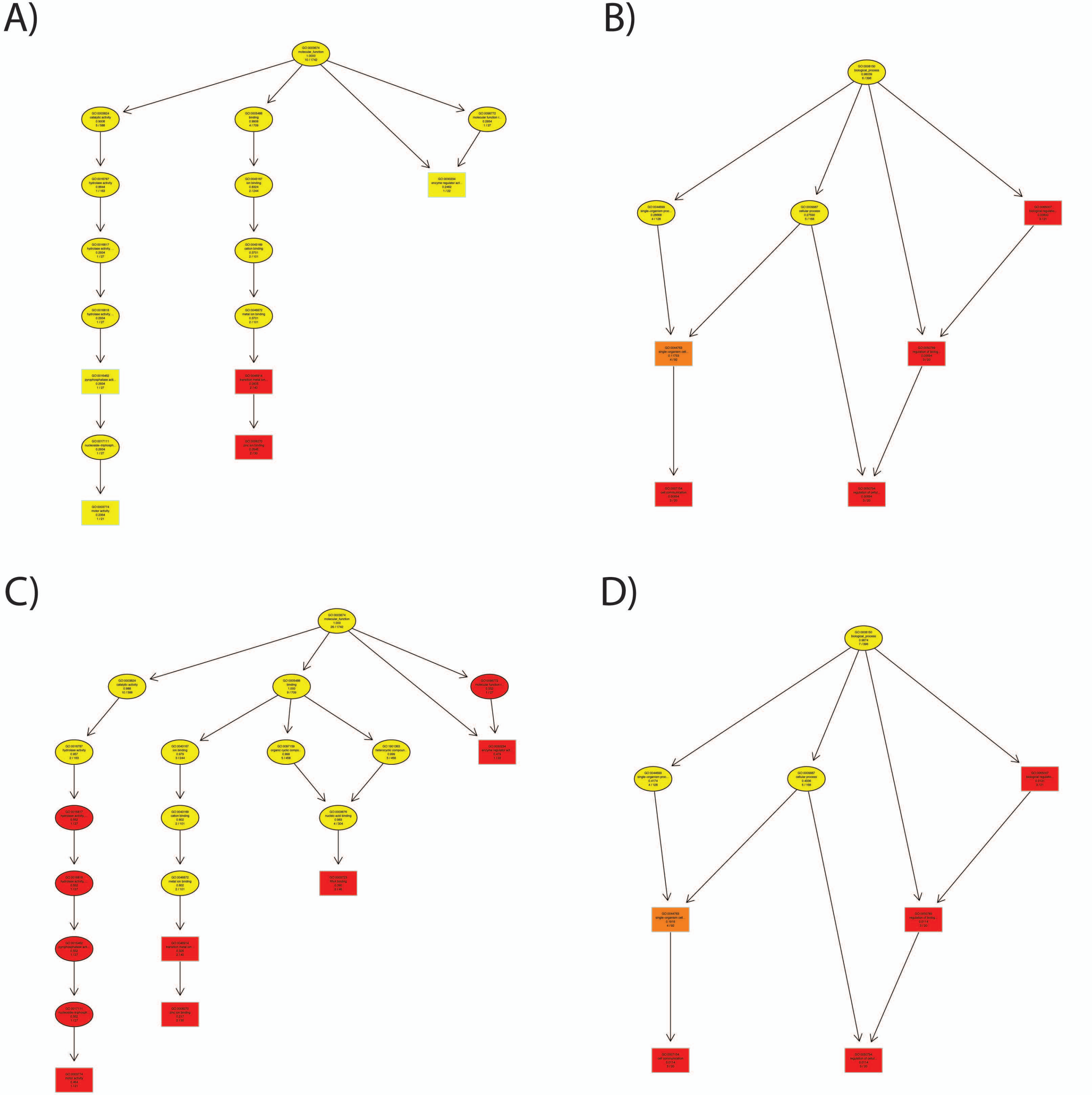
GO ID	Pathway Name (KEGG ID)	*Corr. p*-value
GO:0006139	nucleobase-containing compound metabolic process	1.4 x10^{-5}
GO:0006725	cellular aromatic compound metabolic process	2.6x10^{-5}
GO:0034641	cellular nitrogen compound metabolic process	9.3x10^{-6}
GO:1901360	organic cyclic compound metabolic process	2.9x10^{-6}
GO:0044700	single organism signaling	0.0071

Molecular Function

GO ID	Pathway Name (KEGG ID)	*Corr. p*-value
GO:0003735	hormone activity	0.014
GO:0004857	enzyme inhibitor activity	0.035
GO:0005102	receptor binding	0.045

KEGG pathways enriched in asexual females (vs. sexual females)

Pathway Name (KEGG ID)	Sig. Genes (Expected)	Sig. Genes (Observed)	Odds Ratio	*p*-value
Cell cycle (04110)	2.99	12	5.04	1.57×10^{-5}
Spliceosome (03040)	2.80	11	4.89	4.37×10^{-5}
Viral carcinogenesis (05203)	3.92	11	3.39	8.29×10^{-4}
Prion diseases (05220)	0.530	4	10.0	1.17×10^{-3}
Alcoholism (05034)	3.9	10	3.06	2.77×10^{-3}
p53 signaling pathway (04115)	1.16	5	5.27	3.80×10^{-3}
RNA transport (03013)	3.67	9	3.06	5.94×10^{-3}
Progesterone-mediated oocyte maturation (04914)	3.14	8	3.02	6.88×10^{-3}
Oocyte meiosis (04114)	3.82	9	2.79	6.88×10^{-3}
Systemic lupus erythematosus (04114)	2.55	7	3.26	7.50×10^{-3}


A)
GO:0003674
molecular_function
1.0000
10 / 1742
GO:0003824
catalytic activity
0.9006
5 / 588
GO:0005488
binding
0.9908
4 / 709
GO:0098772
molecular function r...
0.2934
1 / 27
GO:0016787
hydrolase activity
0.8844
1 / 163
GO:0043167
ion binding
0.8324
2 / 244
GO:0030234
enzyme regulator act...
0.2462
1 / 22
GO:0016817
hydrolase activity, ...
0.2934
1 / 27
GO:0043169
cation binding
0.3701
2 / 101
GO:0016818
hydrolase activity, ...
0.2934
1 / 27
GO:0046872
metal ion binding
0.3701
2 / 101
GO:0016462
pyrophosphatase acti...
0.2934
1 / 27
GO:0046914
transition metal ion...
0.0905
2 / 40
GO:0017111
nucleoside-triphosph...
0.2934
1 / 27
GO:0008270
zinc ion binding
0.0546
2 / 30
GO:0003774
motor activity
0.2364
1 / 21
B)
GO:0008150
biological_process
0.98039
6 / 396
GO:0044699
single-organism proc...
0.28668
4 / 128
GO:0009987
cellular process
0.27590
5 / 168
GO:0065007
biological regulatio...
0.00800
3 / 21
GO:0044763
single-organism cell...
0.11753
4 / 92
GO:0050789
regulation of biolog...
0.00694
3 / 20
GO:0007154
cell communication
0.00694
3 / 20
GO:0050794
regulation of cellul...
0.00694
3 / 20
C)
GO:0003674
molecular_function
1.000
26 / 1742
GO:0003824
catalytic activity
0.988
10 / 588
GO:0005488
binding
1.000
9 / 709
GO:0098772
molecular function r...
0.552
1 / 27
GO:0016787
hydrolase activity
0.957
2 / 163
GO:0043167
ion binding
0.979
3 / 244
GO:0097159
organic cyclic compo...
0.999
5 / 458
GO:1901363
heterocyclic compoun...
0.999
5 / 458
GO:0030234
enzyme regulator act...
0.479
1 / 22
GO:0016817
hydrolase activity, ...
0.552
1 / 27
GO:0043169
cation binding
0.802
2 / 101
GO:0003676
nucleic acid binding
0.983
4 / 304
GO:0016818
hydrolase activity, ...
0.552
1 / 27
GO:0046872
metal ion binding
0.802
2 / 101
GO:0003723
RNA binding
0.390
2 / 46
GO:0016462
pyrophosphatase acti...
0.552
1 / 27
GO:0046914
transition metal ion...
0.326
2 / 40
GO:0017111
nucleoside-triphosph...
0.552
1 / 27
GO:0008270
zinc ion binding
0.217
2 / 30
GO:0003774
motor activity
0.464
1 / 21
D)
GO:0008150
biological_process
0.9874
7 / 396
GO:0044699
single-organism proc...
0.4174
4 / 128
GO:0009987
cellular process
0.4336
5 / 168
GO:0065007
biological regulatio...
0.0131
3 / 21
GO:0044763
single-organism cell...
0.1916
4 / 92
GO:0050789
regulation of biolog...
0.0114
3 / 20
GO:0007154
cell communication
0.0114
3 / 20
GO:0050794
regulation of cellul...
0.0114
3 / 20