Supplemental Computational Methods

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 10^{th} and 90^{th} 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).

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.

References

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

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

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

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

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

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.

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.

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.

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.

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

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.

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.

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.

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

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.