doi:10.1038/nature09990

Supplementary Methods

I. ChIP-chip and ChIP-seq methods

Chromatin Immunoprecipitation

ChIP experiments were performed as described previously (Negre, Lavrov et al. 2006) with whole *Drosophila melanogaster* animals from the following developmental stages: embryonic stages 0-4h, 4-8h, 12-16h, 16-20h, 20-24h, larval stages L1, L2, and L3, pupal stage and adult male. Briefly, the biological material was homogenized in the presence of 1.8% of formaldehyde. The cross-linked chromatin was sonicated using a Bioruptor (Diagenode) to an average size of 500bp. Pre-cleared chromatin extract was incubated overnight at 4C with the specific antibody and immunoprecipitated with protein-A Sepharose beads.

ChIP-chip Microarray details

After purification of the DNA and amplification of the libraries by linker-mediated PCR, the samples were then labeled by incorporation of Cy3-dCTP (for the Input sample) or Cy5-dCTP (for the IP sample) using the Invitrogen Random Priming Labeling Kit. Samples were then hybridized on dual-color Agilent 244K tiling microarrays. Three arrays were used to cover the genome and each ChIP was performed in triplicate. Image analysis was carried out using Agilent Feature Extraction Software Version 9.1.2 (Agilent, CA). Dye-normalized log2-ratios of ChIP-signal to input-signal were calculated. For the HMM-segmentation, the data were quantile normalized (Ji, Jiang et al. 2008). For each replicate the log-ratio data of the set of three Agilent tiling were concatenated. For each probe, the mean of the dye-normalized log-ratio was calculated throughout the three respective replicates (i.e. one value per probe, per stage, per histone modification).
For Hybridizations on Affymetrix Tiling arrays v2.0 (MR), the IP sample and the Input samples are both labeled and hybridized separately according to Affymetrix protocols after amplification of the material by Linker mediated PCR.

ChIP-Sequencing details

For each time-point, one replicate of ChIP and its corresponding Input sample were sequenced on one lane of Illumina each. The native IPs were used to produce the Illumina libraries. The double-stranded DNA ends were repaired with T4DNA Polymerase, Klenow Fragment and T4 PNK enzymes. After a second purification step, an adenine-residue was added with Klenow [3'>5' exo-] enzyme and again purified on Quiaquick columns. Adapters from Illumina for LM-PCR were then ligated to the end of the DNA molecules. The product of the reaction was then run on an Agarose gel (2% NuSieve) and a band corresponding to 200 bp was extracted and purified. 20 cycles of PCR were performed using phusion polymerase (Finnzyme F-530S) and the Illumina oligos. The PCR product was then purified by gel electrophoresis. High throughput sequencing was performed on an Illumina Genome Analyzer with standard Illumina 36 cycles reaction kit. The quality-filtered 36-bp short sequence reads were aligned to the reference sequence consisting of dmel5 (NCBI Build 5, March 2006) *D.melanogaster* genome using ELAND (Efficient Local Alignment of Nucleotide Data) software as implemented in the Illumina Genome Analyzer software 1.3.2, allowing up to two mismatches with the reference sequence. Only successfully mapped unique monoclonal reads were used in subsequent analyses.

Peak calling – ChIPseq

Standard peak calling ChIPseq experiments was performed with MACS (Zhang, Liu et al. 2008) and Peakseq (Rozowsky, Euskirchen et al. 2009). MACS analyses were performed with the following parameters: tagSize = 36, mfold = 2, genomeSize = 120000000, bandwidth = 100,

pvalue = 1e-5. Peakseq analyses were run with the following parameters: L =200, window size = 30000, max threshold = 100, max gap = 200, FDR = 0.05, number of sims = 10, bin size = 10000, bin sizeM = 1000, max count = 3, extended region size = 2000, Pf = 0, pval threshold = 0.05. Identification of large, enriched domains was perfomed using a custom modification of HGGSEG (available upon request), run using the following parameters: num states = 2, smooth = 32000, num starts = 3.

Peak calling – ChIPchip

Peak calling for ChIP-chip experiments performed on agilent arrays was performed using CisGenome (Ji, Jiang et al. 2008), with the following parameters: Method to compute FDR = 0, W = 5, Window Boundary = 300, Standardize MA Statistics = 1, Region Boundary Cutoff, MA = 3, Expected Hybridization Length = 12, Posterior Probability Cutoff, P> 0.5, G0 Selection Criteria, p% = 0.01, G1 Selection Criteria, q% = 0.05, Selection Offset = 6, Grid Size = 1000, Number of Permutations = 10, Echangeable Groups = 1, Max Gap within aRegion = 250, Max Run of Insignificant Probes within a Region = 3, Min Region Length = 150, Min No. of Significant Probes Within a Region = 5.

Peak calling for ChIP-chip experiments performed on Affymetrix arrays was performed using MAT (Johnson, Li et al. 2006), with the following parameters: BandWidth = 125, MaxGap = 100, MinProbe = 3, Pvalue = 1e-5, FDR = 0.05.

H3K27me3 Domain Finding - Segmentation Based of ChIP-chip Data Based on Hidden Markov Model (HMM).

Fold changes for every probe on the array was calculated and intensity information at every 5000 base pair was identified with window based smoothing. This continuous intensity information was used as input to HMMSeg, Hidden Markov Model based segmentation for parameter learning and region calling (Day et al. 2007). HMMSeg was run with a wavelet smoothing window size of 32,000 bp and 2 states. The post-processing of the result was carried out a using Perl scripts (available on demand). Depleted peaks were removed by checking each called HMM-peak against the mean smoothed tag density scores in that peak region. Only HMM-peaks with a mean greater than 0 were retained. HMMSeg's smoothing technique, which allows for calling large domains as peaks, introduced an artifact into the results: Called peaks often extended beyond the actual binding domain of the antibody. To correct for this, bins were trimmed from the 5' and 3' ends of the HMM-peaks by assessing their significance. Z-scores for each bin were calculated independently for each HMM-peak. Those bins with z-scores below a defined threshold were excluded from the respective HMM-peak. To account for outliers, an additional parameter was defined to specify the number of contiguous bins with significant z-scores that must be found before terminating the trimming algorithm. Finally, whole HMM-peaks were assessed for significance. A cutoff value was determined for each chromosome by a defined percentile. All peaks which have a maximum probe value that falls below the cutoff threshold were excluded from the results.

H3K27me3 Domain Finding - Segmentation of ChIP-chip and ChIP-seq Data Based on Summed Squared Z-Scores (SSZS).

This segmentation method is based on a method used to reduce the search space for finding chromatin signatures (Hon et al. 2008). Each chromosome was divided up into bins of 100 bp. For ChIP-chip data, within each bin, the mean of all probes was calculated, for ChIP-seq data the background-corrected read count. Values for bins were interpolated if missing and data for neighboring bins was present. For each combination of stage s and histone modification h, the

mean μh,s and standard deviation σh,s was calculated over all chromosomes. For each bin j the z-scores zh,s,j = (logRh,s,j- μh,s)/ σh,s was calculated. The z- and z2 –scores were summed over a sliding window of size w = 400 bins. The following c2-statistics was used to determine the significance at a p-value cutoff of 10-5:

$$y_{h,j} = \sum_{k=1}^{w} z^2_{h,j+k} \sim \chi^2_w$$

Those significant summed z2 –scores were discarded that originate from negative summed z-scores. Furthermore, only contiguous spans of significant bins of size more than 4Kb were called "domains" and were investigated further.

H3K27me3 Domain Finding - Combining the HMM and SSZS-bases domain definitions. Finally, we defined H3K27me3 domains in this study as those chromosomal regions, in which both methods indicated the presence of a domain. Further, the HMM approach was used to refine the position of boundaries of the domains, as HMM boundary definitions were in general considered more precise than the SSZS boundary definitions.

II. Antibodies

Description

The antibodies used in his study were:
Commercially purchased:
H3K9Ac ab4441 Abcam
H3K9me3 ab8898 Abcam
RNA Pol II 8wG16 Covance
H3K4me1 ab8895 Abcam
H3K4me3 ab8580 Abcam
H3K27Ac ab4729 Abcam
H3K27me3 07-449 Upstate
H3K36me3 ab9050 Abcam
HP1 ab25726 Abcam
HP1 Covance
a-H3K4me3-LP LP Bio
a-end300 Santa Cruz

gifts from the community:
a-rbCBP-MM Mattias Mannervik
a-Bab1-SC Sean Carroll/Thomas M. Williams
a-bks-MM Mattias Mannervik
a-brm-AD Andrew Dingwall
a-cad55-JR John Reinitz
a-CBP-MM Mattias Mannervik
a-chinmo-EB Erika Bach
a-dll-SC Sean Carroll/Thomas Williams
a-en-FM Florence Maschat
a-gsbn-FM Florence Maschat
a-sens-HB Hugo Bellen
a-Sin3A-RC Ross Cagan/Tirtha Das

a-snr1-AD	Andrew Dingwall
a-STAT92E-EB	Erika Bach
GAF3558	Carl Wu
CTCF-C	Rainer Renkawitz
CTCF-N	Rainer Renkawitz
Su(Hw)-1	Victor Corces
Su(Hw)-2	Pam Geyer
CP190	David M. Glover
BEAF-32	Ulrich Laemmli
Mod(mdg4)	Victor Corces
HDAC-492	Dan Garza/Marc Hild/Novartis
HDAC-493	Dan Garza/Marc Hild/Novartis
HDAC-494	Dan Garza/Marc Hild/Novartis
HDAC-495	Dan Garza/Marc Hild/Novartis
HDAC-496	Dan Garza/Marc Hild/Novartis
HDAC-497	Dan Garza/Marc Hild/Novartis
HDAC-498	Dan Garza/Marc Hild/Novartis
HDAC-499	Dan Garza/Marc Hild/Novartis
HDAC-500	Dan Garza/Marc Hild/Novartis
HDAC-501	Dan Garza/Marc Hild/Novartis

Custom-made and available to the community:

a-Ubx1-MK	Kevin White/Max Kauer
a-Ubx2-MK	Kevin White/Max Kauer
KW0-CNC	Kevin White/Nicolas Negre
KW0-D	Kevin White/Nicolas Negre
KW0-dCtBP7667	Kevin White/Nicolas Negre
KW0-GRO	Kevin White/Nicolas Negre
KW0-INV7657	Kevin White/Nicolas Negre
KW0-KN7697	Kevin White/Nicolas Negre
KW0-RUN7659	Kevin White/Nicolas Negre
KW0-TTK7691	Kevin White/Nicolas Negre
KW0-UBX7701	Kevin White/Nicolas Negre
KW0-ZFH17684	Kevin White/Nicolas Negre
KW3-D-D2	Kevin White/Nicolas Negre
KW3-disco-D2	Kevin White/Nicolas Negre
KW3-h-D1	Kevin White/Nicolas Negre
KW3-hkb-D1	Kevin White/Nicolas Negre
KW3-jumu-D2	Kevin White/Nicolas Negre
KW3-kni-D2	Kevin White/Nicolas Negre
KW3-Kr-D2	Kevin White/Nicolas Negre
KW3-Trl-D2	Kevin White/Nicolas Negre
KW4-E(z)-D2	Kevin White/Nicolas Negre
KW4-GATAe-D1	Kevin White/Nicolas Negre
KW4-Pcl-D2	Kevin White/Nicolas Negre
KWG-GFP	Kevin White/Ralf Kittler
a-FTZ-F1	Kevin White/JiangLiu

III. Gene Expression experiments on whole Drosophila embryos

RNAseq library construction

Solexa libraries for cDNA sequencing were constructed similarly to previously described methods (Marioni, Mason et al. 2008). Briefly, matching total RNA was collected from each time-point of this study in TRIzol reagent and isolated according to the manufacturer instructions. DNAse I treated RNA is then purified and concentrated using the RNeasy MinElute Cleanup Kit. PolyA RNAs are then purified using the micropolyA purist Kit from Ambion and converted into single stranded DNA after Reverse Transcription using random hexamers. The second strand synthesis is then carried out by adding to the reaction the RNAseH (Invitrogen #18021014) and DNA Polymerase II (NEB #M0209S) enzymes. At this stage, the double stranded DNA is then cleaned up on Qiagen Quiaquick columns and the ends are repaired by using the T4DNA Polymerase, Klenow Fragment, and T4 PNK enzymes. After another round of Qiaquick purification, an A residue was added with Klenow [3'>5' exo-] enzyme and the product was again purified on Quiaquick columns. Adapters from Illumina for LM-PCR are then ligated to the end of the DNA molecules. The product of the reaction was then run on an Agarose gel (2% NuSieve) and a band corresponding to 300 bp was then extracted and purified. 20 cycles of PCR reaction were then performed using phusion polymerase (Finnzyme F-530S) and the Illumina oligos. The product was then purified by gel electrophoresis. Solexa sequencing as then performed on Genome Analyzer with standard Illumina 36 cycles reaction kit.

RNAseq alignment, transcript assembly

RNAseq reads were aligned to the Drosophila melanogaster reference genome (dm3) using TopHat (v. 1.0.12) (Trapnell, Pachter et al. 2009). SAM formatted alignment files were used for transcript assembly using cufflinks (v0.8.3) (Trapnell, Williams et al.) with RefSeq transcript annotations as a gene model reference (i.e., allowing for multiple transcript isoforms per gene). FPKMs were extracted from cufflinks output for each assembled transcript. Unassembled transcript models were assigned an FPKM of 0. One pseudocount was added to all estimates prior to $\log_2$ transformation.

RNAseq timecourse clustering

Data from this study were merged with an independent RNAseq timecourse performed by the modENCODE transcription group (Graveley et al. submitted). Prior to study merging, robust regression models were built for each transcript to estimate, and remove, a lab of origin effect. Studies were then merged and experiments were sorted by developmental stage and a three timepoint moving average was applied to smooth expression estimates. The resulting merged transcript expression measurements were then clustered by k-means clustering (k=28), as implemented in the R function *clara.*

IV. Gene expression – chromatin marks classification

Binary classification

The longest annotated transcript was selected to represent each gene. Genes were classified as 'marked' when there was at least one peak within the range -1000 bp upstream to TSS and min[2000bp, length of transcript] downstream to TSS, otherwise it was classified as 'unmarked'. The RNA-Seq time course was used to identify gene activity. A gene was considered active if at least one exon has RNA coverage >=20%, and the entire transcript has an average coverage >=10%.

For each one of H3K4Me1, H3K4Me3, H3K9Ac, H3K27Ac, CBP and PolII, we classified each gene, based on ChIP and RNAseq signals, into four categories: (1) Category11: Marked and

Activate; (2) Category10: Marked and Inactivate; (3) Category01: Unmarked and Activate; (4) Category00: Unmarked and Inactive.
The same analysis was done on H3K4Me3 mapping in Kc cells.

Regression classifier

We implemented a supervised learning approach to distinguish active promoters from inactive promoters based on their chromatin modification and transcription factor binding properties. The transcription start site (TSS) was used as an estimate of the promoter location and RNAseq data was used to define active and inactive promoters based on an RPKM threshold. As many transcripts start sites (TSSs) are shared, or are in close proximity, we grouped TSSs by surrounding each TSS with 200 bp regions, merging overlapping regions, and then took the max transcript value within a merged region to reduce redundancy in the data used in the classifier. The distribution of RPKM values for the different developmental stages are shown in Supplementary Fig. 5a.

The features used to distinguish active from inactive promoters correspond to the ChIP-chip and ChIP-seq data generated for the 6 chromatin modifications and binding site profiles for PolII and CBP. A 1kb region was centered on each TSS and split into 100bp bins. ChIP-chip and ChIP-seq values were mapped to each bin producing a vector of values for each dataset at each TSS. If multiple values were present across a bin, the average of the values was used, weighted by the faction of overlapping signal. We required that at least 5 of the bins had values and filtered out regions that did not have consistent signal across all the dataset for a given developmental stage.

We implemented a strategy developed for integrating chromatin signatures into promoter prediction models in order to represent the values as features for the classifier (Wang, Xuan et al. 2009). For the set of positives we first computed an average vector for each mark and then for each individual promoter we calculated the Pearson correlation and the dot product with the average of the positives. The Pearson correlation was selected to represent the shape of the signal and the dot product was used to represent the intensity.

We learned a classifier for each development stage separately using a logistic regression model implemented in the WEKA machine learning software suite (Frank, Hall et al. 2004). We selected a number of RPKM values (0.1, 0.5, 1, 1.5, 2, 2.5) as thresholds to define the positives. Performance evaluation was carried out using 10 fold cross-validation, where the chromosomal distribution and RPKM distribution of transcripts were matched in each fold. A RPKM threshold of 1 was selected based on the receiver operating characteristic (ROC) curves (Supplementary Fig. 8d), using the area under the curve (AUC) as a performance metrix (Supplementary Fig. 8b) and looking at the recall values for a false discovery rate (FDR) of 0.10 (Supplementary Fig. 8c).

The binary classifier predicts transcripts as marked or unmarked, resulting in 4 possible outcomes. Transcripts can be marked and active (MA), marked but inactive (MI), unmarked and inactive (UI), and unmarked but active (UA). The UA and MI classes contradict the expectation that transcriptional status is consistent with the chromatin signature profiles at the promoter. In order to investigate the UA and MI classes, we first compared the transcript level distributions (RPKM values) between the MA and UA classes and between the MI and UA classes (supplementary Fig. 9b). Median RPKM score for the MA transcripts is significantly greater than the median RPKM score for the UA transcripts, suggesting that highly expressed transcripts have a greater success rate a being classified as marked. An exception is in the AdultMale. In contrast, the MI and UI classes often have similar medians or the UI class is greater.

Next we quantified how predictable the transcripts are across the 12 developmental stages. We define the predictability of active transcripts to be the number of stages where the transcripts were classified as marked normalized by the number of stages the transcripts were detected as active using RNAseq. Similarly, the predictability of inactive transcripts is the number of stages a transcript is classified as unmarked normalized by the number of stages the transcript is inactive. We then binned the predictability values into quartiles and show the distribution for the numbers of stages the transcripts are active or inactive (Supplementary Fig. 10b). We find that most transient active transcripts have low predictability. In contrast the predictability of inactive transcripts is more stable.

V. Characterization of genes associated with H3K27me3 domains during fly development

Binary Clustering of H3K27me3 Domain Genes.

To prepare for binary clustering of H327me3 genes, those genes were extracted that were defined as H3K27me3 domain genes in at least one developmental stage. A binary matrix was established taking into account all those genes over all developmental stages. Whenever a gene was part of an H3K27me3 domain in a distinct stage, the respective position in the matrix was filled in by 1, otherwise by 0. This binary matrix was used for binary hierarchical clustering using the statistical software R (R Development Core Team 2008) and the bioconductor package collection (Gentleman et al. 2004). In R the "binary" distance measure of the "dist" function was employed. In brief, the distance between two genes based on their individual vector of 0 and 1 was calculated by comparing each position of the vector for gene A with the respective position of the vector for gene B. The distance was calculated by the number of times vector A differed from vector B divided by the number of positions in which at least one gene showed a "1". This distance matrix was clustered using the hierarchical clustering function "hclust" with average linkage option. The resulting dendrogram was used to define cluster membership for each gene. Using the "rect.hclust" function in R, the dendrogram was cut such that a chosen number (either 20 or 50) of clusters was produced.

Functional Classification of H3K27me3 Domain Genes using Gene Ontology.

For clusters of H3K27me3 domain genes, an overrepresentation analysis for Gene Ontology (GO) entries was carried out. To this end, a Fisher's test was employed as implemented in the DAVID Bioinformatics Resource (Dennis et al., 2003). As a reference gene list, all RefSeq (Pruitt et al,. 2007) entries were used. The resulting *P*-values were adjusted using Benjamini-Hochberg's multiple testing correction method.

Comparison of H3K27me3 domain gene clusters with clusters of similar time- and tissue-specific embryonic in situ gene expression patterns

Tomancak et al. (Tomancak et al. 2007) grouped genes into clusters based on pattern annotations derived from *in situ* hybridization experiments at different stages of embryogenesis. They established a set of "all" clusters in which each gene could be a member of several clusters, as well as a set of "core" clusters in which each gene could be a member of only one cluster. Here, lists of genes names of individual "all" and "core" clusters from the study carried out by Tomancak et al. (2007; http://www.fruitfly.org/insitu/) were compared to list of gene names contained by individual H3K27me domain clusters. Over- and underrepresentation of Tomancak cluster genes in each H3K27me3 domains was tested by a Fisher's exact test ($p < 0.05$).

VI. Site specific transcription factors

Peak annotation

For multiple datasets for the same factor and same stage or cell type (0-12h as early embryo, pupae, larva and Kc-167 cell type), we merged the peaks and used union part for following analysis. The genome annotation from FlyBase 5.24 was used to annotate peaks mapped by the

site-specific factors. The peaks are sequentially annotated as 5'UTR, 3'UTR, CDS, intron and intergenic if they overlapped with the region annotated by the gff file.

TFBS Complexity

To quantify the interaction between transcription factors binding sites, we counted overlapping binding sites for each pair of TFs. For each pairwise combination of transcription factors, binding site overlap enrichment was calculated using Fisher's exact test. –Log10 transformed p-values were used to hierarchically cluster all TFs. To identify transcription factor binding site hotspots, we combined TFBS from all TFs assayed at early embryo stages. The number of TFs in each merged TFBS was defined as 'TFBS complexity'. The TFBS complexity categories were annotated based on FlyBase 5.24 for *Drosophila melanogaster*. As TFBS complexity may overlap more than one annotation type, we sequentially marked regions as 5'UTR, 3'UTR, CDS, intron and intergenic. Genes were assigned a TFBS complexity based on the maximum TFBS complexity within 2kb of the annotated TSS. Spearman correlation coefficient was calculated between FPKM values and TFBS complexity of each gene (genes with FPKM<1 were ignored). Regions annotated as tissue-specific enhancers by CAD , CBP binding sites, and H3K4Me1 enriched regions were used as enhancer-related regions. To quantify enrichment between each putative enhancer class and TFBS complexity categories, Fisher's exact tests were performed. The p-value was transformed by –log10, and then scaled to generate the heatmap, with TFBS complexity as rows and enhancers as columns.

Binding site enrichment calculation

We calculated the similarity between chromatin modifications and protein profiles (Fig 3a) by first taking the union of the ChIP-chip and ChIP-seq data, then determining the number of overlapping base pairs for regions R1 and R2 (the 2 datasets being compared), and using the following formula to calculate the enrichment:

$$\frac{\text{Size}(R1 \Lambda R2) \times \text{Size}(\text{background})}{\text{Size}(R1) \times \text{Size}(R2)}$$

In this case the background is the regions represented on the ChIP-chip array, and for ChIP-seq data only those regions that overlap the background were considered. Similarly, we used this approach to calculate the enrichment of CBP or H3K4me1 with known enhancers (Fig 1c) and the enrichment of TF binding with known enhancers (Fig 1g).

VII. Clustering CBP bound regions

We developed an approach to group CBP regions by their patterns of overlapping transcription factor binding events. For this analysis we focused on the distal CBP bound regions using the ChIP-seq and filtered out peaks that overlapped a 1kb span surrounding all annotated TSSs. CBP ChIP-seq peaks were merged across all developmental stages and regions greater than 1kb were segmented into 1kb regions by first identifying the location that is bound across the most developmental stages and then adding a 1kb span. The 1kb segment was removed and the procedure was repeated until there were no 1kb fragments remaining.

In order to represent the transcription factor (TF) binding events associated with CBP regions, we required, for TF peaks that were shorter than the overlapping CBP regions, that at least 50% of the TF peak overlap in the CBP region or, for TF peaks longer than the overlapping CBP region,

that at least 50% of the CBP region overlap the TF (Supplementary Fig. 20a). Through this approach we assembled a binary vector for each CBP region corresponding to a range of TFs, insulators, and chromatin remodeling datasets generated in this study and from the literature. A set of random genomic regions we included as a control and were selected matching the length distribution and total number of associated experiments. We did not include the CBP data as a feature, and used this to validate the clustering and see if subsets of the regions were enriched for CBP binding.

The regions were clustered using a finite mixture model of multivariate Bernoulli distributions. Mixture weights and component distributions learned from data using the Expectation Maximization algorithm using the Bernoulli mix software package (Myllykangas, Tikka et al. 2008). We learned the component distributions for a range of cluster sizes and selected a model using the Bayesian information criterion (BIC) score, a metric that balances the model fit with the complexity of the model.

$$BIC = -2L + N\ln(n),$$

where L is the log-likelihood of the model, N is the number of estimated parameters, and n is the number of datasets used. A model with 22 clusters was selected for further analysis (Supplementary Fig. 20b, Supplementary Table 14).

In order to determine the clusters enriched for CBP and at what developmental stages these regions are bound, we calculated the enrichment and statistical significance of CBP binding in each cluster. For the enrichment calculation, we used procedure described above where the background included the regions selected for the analysis. The statistical enrichment for CBP was determine using a hypergeometric test, where we calculated the probability of observing the number of CBP regions within a cluster given the size of the cluster, the total number of CBP regions and the total number of region used in the analysis. The p-values were converted to a false discovery rate (FDR) as a multiple hypothesis correction (only FDR<0.01 are shown in Fig 1d, Supplementary Table 15). In addition we calculated the enrichment of enhancers (Fig 1e, Supplementary Table 16) and chromatin profiles and PolII binding (Supplementary Fig 20c) were carried out as described above, where the background includes the regions used in the analysis. The statistical significance of the overlap was determined using the Genome Structure Correlation analysis tool.

VIII. Motif discovery

We collected experimental datasets annotating transcription factor binding from both modENCODE and the literature (Negre, Brown et al. ; Moses, Pollard et al. 2006; Sandmann, Jensen et al. 2006; Georlette, Ahn et al. 2007; Jakobsen, Braun et al. 2007; Sandmann, Girardot et al. 2007; Zeitlinger, Zinzen et al. 2007; Kwong, Adryan et al. 2008; Lee, Li et al. 2008; Bushey, Ramos et al. 2009; Gambetta, Oktaba et al. 2009; Liu, Jakobsen et al. 2009; MacArthur, Li et al. 2009; Schuettengruber, Ganapathi et al. 2009; Zinzen, Girardot et al. 2009).

Each peak dataset was randomly partitioned into two subsets and +/- 200bp from the center of each peak was taken. From one of the two random partitions, the top 250 peaks in terms of intensity (randomly selected if no intensity values were available) were used in motif discovery. Motif discovery was performed independently using five tools (AlignACE (Hughes, Estep et al. 2000), MDscan (Liu, Brutlag et al. 2002), MEME (Bailey and Elkan 1994), Weeder (Pavesi, Mereghetti et al. 2004), and Trawler (Ettwiller, Paten et al. 2007)).

The resulting motifs were used to scan the genome for motif instances using a PWM threshold corresponding to a p-value of $4x10^{-8}$ as determined by TFM-Pvalue (Touzet and Varre 2007).

Enrichment of each motif was computed as the fraction of instances found in the second random partition of peaks divided by the fraction for instances of shuffled control motifs (Wilson's confidence interval (Wilson 1927) at Z=1.5 was used on the ratios to give a conservative enrichment). We ignored from all our analyses all motifs predicted within coding exons, repeats, transposons, 3' untranslated regions and non-coding RNAs (from FlyBase version 5.28).

For each factor, we selected up to 5 motifs in descending order of enrichment in their original dataset while not permitting any two motifs with a correlation greater than 0.75. We also selected known motifs for each factor from the literature (Sen, Stultz et al. ; Matys, Fricke et al. 2003; Wasserman and Sandelin 2004; Down, Bergman et al. 2007; Ivan, Halfon et al. 2008; Noyes, Christensen et al. 2008; Noyes, Meng et al. 2008; Reed, Huang et al. 2008; MacArthur, Li et al. 2009) and jointly with the discovered motifs evaluated their enrichment. We also took all other known motifs with similarity at least 0.75 to any of the known/discovered motifs and display them. This resource can be browsed at (http://www.broadinstitute.org/~pouyak/fly-motif-disc/www/).

IX. Promoter validation

Activities of the predicted promoters were tested by transient transfection of luciferase reporter plasmids and dual luciferase assays. The firefly luciferase reporter plasmids were constructed by replacing the SV40 promoter in pGL3P-df (Heintzman, Stuart et al. 2007) by Drosophila melanogaster genomic sequences of about 1 kb in sizes containing the predicted promoters. The inserts were generated by PCR amplification of Drosophila melanogaster genomic DNA with primer pairs with 15-base extensions of either 'CCCGGGCTCGAGATC' or 'CCGGAATGCCAAGCT' added 5' to the region-specific sequences. The PCR products were inserted into pGL3P-df and digested with Bgl II and Hind III using the In-Fusion Dry-Down PCR Cloning Kits (Clontech). Transient transfection of Kc167 cells was carried out in 96-well plates using the Effectene transfection reagent (Qiagen). The cell culture was diluted to a density of 1 million cells per ml one day before transfection. On the day of transfection, 100 µl of the cell culture was added to each well and 320 ng of a firefly luciferase reporter construct was co-transfected with 80 ng of the Pol III-Renilla luciferase reporter (obtained from Dr. J. T. Kadonaga at UCSD) per well. Cells were harvested 1 day after transfection and the luciferase activities were measured using the Dual-Luciferase Reporter Assay system (Promega). To correct for transfection efficiencies, the firefly luciferase activity of each sample was normalized to the corresponding Renilla luciferase activity. In a single experiment, a reporter plasmid was always transfected in triplicate. Every experiment included four positive controls and eight negative controls. If a reporter construct showed a normalized luciferase activity at least two standard deviations above the mean of the eight negative controls in an experiment, it was scored as positive. A predicted promoter that scored positive in at least two of three independent experiments was considered to be active.

X. Enhancer validation

Peaks from CBP ChIP-chip and ChIP-seq experiments have been filtered for two parameters. We kept distal binding sites by excluding peaks falling between -500bp and +500bp of any annotated promoter. Peaks occuring only at one stage out of the 12 stages studied have also been removed. We generated two list of binding sites of CBP, occuring at all stages and occuring on embryonic stages only after merging of all CBP peaks. The list of peaks was visually inspected using the Integrated Genome Browser (Affymetrix) and 1.5 kb regions were selected approximately centered on peak maxima. Primers were designed in mass using the BatchPrimer3 program (http://probes.pw.usda.gov/batchprimer3/index.html), optimized to generate 24mers. PCR

products were cloned in to the pBPGUw vector (Pfeiffer, Jenett et al. 2008) using the Gateway system (Invitrogen) via a TOPO/pCR8/GW intermediate. Injections were performed by Genetic Services, Inc. into the phi-C31 compatible docking site attP2. 24 hr collections of embryos for each construct were fixed and subjected to in situ hybridization using a GAL4 anti-sense RNA probe generated as in (Pfeiffer, Jenett et al. 2008).

References

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bushey, A. M., E. Ramos, et al. (2009). "Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions." Genes Dev **23**(11): 1338-50.

Day, N., Hemmaplardh, A., et al. (2007). "Unsupervised segmentation of continuous genomic data". Bioinformatics **23**(11): 1424-6.

Dennis G Jr, Sherman BT, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5):P3.

Down, T. A., C. M. Bergman, et al. (2007). "Large-scale discovery of promoter motifs in Drosophila melanogaster." PLoS Comput Biol **3**(1): e7.

Ettwiller, L., B. Paten, et al. (2007). "Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation." Nat Methods **4**(7): 563-5.

Frank, E., M. Hall, et al. (2004). "Data mining in bioinformatics using Weka." Bioinformatics **20**(15): 2479-81.

Gambetta, M. C., K. Oktaba, et al. (2009). "Essential role of the glycosyltransferase sxc/Ogt in polycomb repression." Science **325**(5936): 93-6.

Gentleman, R.C., Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Georlette, D., S. Ahn, et al. (2007). "Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells." Genes Dev **21**(22): 2880-96.

Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet **39**(3): 311-8.

Hon, G., Ren, B., et al. (2008). "ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome." PLoS Comput Biol **4**(10): e1000201.

Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." J Mol Biol **296**(5): 1205-14.

Ivan, A., M. S. Halfon, et al. (2008). "Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs." Genome Biol **9**(1): R22.

Jakobsen, J. S., M. Braun, et al. (2007). "Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network." Genes Dev **21**(19): 2448-60.

Ji, H., H. Jiang, et al. (2008). "An integrated software system for analyzing ChIP-chip and ChIP-seq data." Nat Biotechnol **26**(11): 1293-300.

Johnson, W. E., W. Li, et al. (2006). "Model-based analysis of tiling-arrays for ChIP-chip." Proc Natl Acad Sci U S A **103**(33): 12457-62.

Kwong, C., B. Adryan, et al. (2008). "Stability and dynamics of polycomb target sites in Drosophila development." PLoS Genet **4**(9): e1000178.

Lee, C., X. Li, et al. (2008). "NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila." Mol Cell Biol **28**(10): 3290-300.

Liu, X. S., D. L. Brutlag, et al. (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nat Biotechnol **20**(8): 835-9.

Liu, Y. H., J. S. Jakobsen, et al. (2009). "A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development." Dev Cell **16**(2): 280-91.

MacArthur, S., X. Y. Li, et al. (2009). "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions." Genome Biol **10**(7): R80.

Marioni, J. C., C. E. Mason, et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-17.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-8.

Moses, A. M., D. A. Pollard, et al. (2006). "Large-scale turnover of functional transcription factor binding sites in Drosophila." PLoS Comput Biol **2**(10): e130.

Myllykangas, S., J. Tikka, et al. (2008). "Classification of human cancers based on DNA copy number amplification modeling." BMC Med Genomics **1**: 15.

Negre, N., C. D. Brown, et al. "A comprehensive map of insulator elements for the Drosophila genome." PLoS Genet **6**(1): e1000814.

Negre, N., S. Lavrov, et al. (2006). "Mapping the distribution of chromatin proteins by ChIP on chip." Methods Enzymol **410**: 316-41.

Noyes, M. B., R. G. Christensen, et al. (2008). "Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites." Cell **133**(7): 1277-89.

Noyes, M. B., X. Meng, et al. (2008). "A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system." Nucleic Acids Res **36**(8): 2547-60.

Pavesi, G., P. Mereghetti, et al. (2004). "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." Nucleic Acids Res **32**(Web Server issue): W199-203.

Pfeiffer, B. D., A. Jenett, et al. (2008). "Tools for neuroanatomy and neurogenetics in Drosophila." Proc Natl Acad Sci U S A **105**(28): 9715-20.

Pruitt, K.D., Tatusova, T., et al. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-65.

R Development Core Team. (2008). "R: A Language and Environment for Statistical Computing." In: R Foundation for Statistical Computing, Vienna, Austria.

Reed, D. E., X. M. Huang, et al. (2008). "DEAF-1 regulates immunity gene expression in Drosophila." Proc Natl Acad Sci U S A **105**(24): 8351-6.

Rozowsky, J., G. Euskirchen, et al. (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." Nat Biotechnol **27**(1): 66-75.

Sandmann, T., C. Girardot, et al. (2007). "A core transcriptional network for early mesoderm development in Drosophila melanogaster." Genes Dev **21**(4): 436-49.

Sandmann, T., L. J. Jensen, et al. (2006). "A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development." Dev Cell **10**(6): 797-807.

Schuettengruber, B., M. Ganapathi, et al. (2009). "Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos." PLoS Biol **7**(1): e13.

Sen, A., B. G. Stultz, et al. "Odd paired transcriptional activation of decapentaplegic in the Drosophila eye/antennal disc is cell autonomous but indirect." Dev Biol **343**(1-2): 167-77.

Touzet, H. and J. S. Varre (2007). "Efficient and accurate P-value computation for Position Weight Matrices." Algorithms Mol Biol **2**: 15.

Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-11.

Trapnell, C., B. A. Williams, et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-5.

Wang, X., Z. Xuan, et al. (2009). "High-resolution human core-promoter prediction with CoreBoost_HM." Genome Res **19**(2): 266-75.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet **5**(4): 276-87.

Wilson, E. B. (1927). "Probable inference, the law of succession, and statistical inference." Journal of the American Statistical Association **22**: 209-212.

Zeitlinger, J., R. P. Zinzen, et al. (2007). "Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo." Genes Dev **21**(4): 385-90.

Zhang, Y., T. Liu, et al. (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.

Zinzen, R. P., C. Girardot, et al. (2009). "Combinatorial binding predicts spatio-temporal cis-regulatory activity." Nature **462**(7269): 65-70.

Supplementary Table 1 - Summary of datasets produced. (A) Chromatin developmental time-course. Number of binding sites (BS) for each chromatin-associated mark or factor are indicated. E0-4, embryos 0-4 hours (h) after egg lay (AEL); E4-8, embryos 4-8h AEL; E8-12, embryos 8-12h AEL; E12-16, embryos 12-16h AEL; E16-20, embryos 16-20h AEL; E20-24, embryos 20-24h AEL; L1, first larval instar; L2, second larval instar; L3, third larval instar. Binding sites for pupae, adult females and adult males are also indicated, as are the cumulative total of unique BS across all stages. (B-C-D) Datasets produced for TFs (B), HDACs (C) and Insulator proteins (D). "Factor" indicates the protein target. Gene symbols have been kept to distinguish between homonyms. Many factors were assayed at multiple stages. "Stage/Cell Type" indicates the material used for the ChIP experiments. If performed in whole animals, the developmental stage is indicated as in (A). Kc167 and S2 represent cell lines. "Ab" indicates which particular antibody has been used for ChIP assays. All antibody names starting with KW indicate antibodies produced by the modENCODE project. Other names indicate antibodies donated by the community. The KWG-GFP antibody is used on transgenic animals where the target factor is fused with GFP via BAC recombineering into the P[acman] system [28] (see Supplementary Text 1). "Platform" indicates whether Affymetrix or Agilent tiling-arrays were used for ChIP-chip, or if Solexa (Illumina) sequencing was used for ChIP-seq. "NumofTFBS" indicates the number of binding sites for each factor reported in this study. "Peak Feature" describes whether the target factor has a binding profile comprised typical peaks (punctate), larger domains (broad) or both (mixed).

Supplementary Table 2 - Chromatin time-course datasets. This table indicates for each dataset of the chromatin time-course the number of peaks and their median length in base pairs.

Supplementary Table 3 - Promoter validation results (see separate file). This table is listing the coordinates of the novel promoters assayed for their activity. The coordinates of each fragment is indicated as well as the result for each orientation tested. "Validated" means that in two out of three independent experiments, the average of the triplicate transfections was greater than 2 standard deviations (SD) above the mean of the negative controls. "Supported" means that only one out of the three independent experiments had the average of the triplicates for that experiment greater than 2 SDs above the mean of the negative controls. "Unsupported" means that none of the experiments had the average of the triplicates greater than 2 SDs above the mean of the negative controls. "Incomplete" means that for that orientation all three experiments have not yet been performed.

Supplementary Table 4 - Enhancer validation summary. This table is listing the different DNA fragments (e.g. EO001) tested for enhancer activity. It provides information on the fragment localization in the genome and indicates whether any enhancer activity was detected ("observed" in the "embryonic expression" column).

Supplementary Table 5 - TF complexity percentages. This table indicates for each complexity category the total amount of genome covered, the number of TF associated to each category and the median length of the merged binding sites. It also indicates the number of transcripts

associated to each binding region (+/- 1kb from an annotated TSS), their mean RPKM value and the number and percentage of active genes associated to each TF complexity category.

Supplementary Table 6 - TSS class annotation at FDR 0.05 (see separate file). This table provides the result of a classifier of each transcript in the genome as a result of expression prediction based on ChIP data and RNA data. In the last column, TN indicates a non marked, non expressed transcript; TP indicates a marked expressed transcript; FN indicates a non marked, expressed transcript and FP a marked, non expressed transcript. The FDR of the classifier is set a 0.05.

Supplementary Table 7 - TSS class annotation at FDR 0.1 (see separate file). Legend as in Table 6, but the DFR of the classifier is set at 0.1.

Supplementary Table 8 - Novel promoter prediction based on co-occurence of H3K4me3, PolII and RNA in embryos (see separate file). This .bed file provides the coordinates of novel promoter predictions in embryos.

Supplementary Table 9 - Insulator validation. This table is indicating the result of the enhancer-blocking assay for the different DNA fragments tested in this study.

Supplementary Table 10 - Insulators Class I (see separate file). This .bed file provides the genomic coordinates of the Class I insulators (CTCF/CP190/BEAF-32).

Supplementary Table 11 - Insulators Class II (see separate file). This .bed file provides the genomic coordinates of the Class II insulators (SU(HW)).

Supplementary Table 12 - HDAC associated PREs (see separate file). This bed file provides the genomic coordinates of the putative PREs defined by the localization of HDAC1 and HDAC4a within H3K27me3 domains but not overlapping H3K4me3 domains.

Supplementary Table 13 - CBP embryo only enhancer predictions (see separate file). This .bed file provides the genomic coordinates of the enhancer predictions based on the presence of CBP binding sites in embryos only.

Supplementary Table 14 -TF driven clustering of CBP bound regions (see separate file). This table is listing the different genomic features that have been clustered based on the presence of CBP binding sites or TF binding sites. The first column indicates each of the 20 clusters studied.

Supplementary Table 15 - Enrichment of CBP developmental stages within CBP clusters (see separate file). This table indicates the enrichment of CBP binding sites at different developmental stages within each of the 20 clusters previously defined.

Supplementary Table 16 - Enrichment of enhancers within CBP clusters. This table indicates the enrichment of CAD enhancers categories within each of the 20 clusters previously defined.

A

Factor \ Stage	E0-4	E4-8	E8-12	E12-16	E16-20	E20-24	L1	L2	L3	Pupae	AdultFemale	AdultMale	Unique # of BS Of Each Factor
CBP	5 072	8 876	329	9 026	6 027	6 042	1 273	-	3 843	8 508	7 735	6 594	17 463
PolII	4 482	6 138	1 333	6 208	12 103	2 413	3 024	5 285	7 143	6 238	2 400	-	15 644
H3K4Me1	8 810	7 403	8 827	14 849	12 610	12 471	3 362	11 290	10 488	10 759	7 669	3 610	21 726
H3K4Me3	7 607	4 160	5 816	5 667	8 558	10 429	6 500	5 743	5 548	5 347	3 841	7 328	11 288
H3K9Ac	6 914	5 243	7 585	8 555	8 603	14 604	4 838	6 907	5 975	5 703	7 512	8 899	21 480
H3K9Me3	974	807	1 440	700	495	449	470	501	315	425	40	466	1 380
H3K27Ac	6 066	5 654	7 720	9 886	12 366	8 674	2 321	6 061	4 774	6 862	7 678	9 725	17 406
H3K27Me3	810	908	1 174	813	422	493	522	428	2 590	441	207	2 202	2 513

B

Factor	Stage/Cell Type	Ab	Platform	NumOfTFBS	Peak Feature	Reproducibility	Signal/Noise
bab1	E0-12h	a-Bab1-SC	Affymetrix v2.0	1 227	punctate	0,96	3,03
bks	E0-4h	Affymetrix v2.0	a-bks-MM	1 706	broad	0,96	2,03
brm	Pupae	Affymetrix v2.0	a-brm-AD	270	punctate	0,98	4,27
cad	E4-8h	Affymetrix v2.0	KWG-GFP	1 070	punctate	0,96	4,13
cad	AF(3d)	Solexa	KWG-GFP	3 457	punctate	No replicate	78,46
cad	E0-4h	Affymetrix v2.0	a-cad55-JR	724	punctate	0,96	3,01
cad	E0-4h	Solexa	KWG-GFP	2 207	punctate	0,61	77,49
cad	E4-8h	Solexa	KWG-GFP	5 626	punctate	No replicate	33,00
cad	AF	Solexa	KWG-GFP	7 579	punctate	0,04	129,32
cad	E0-4h	Affymetrix v2.0	KWG-GFP	1 700	punctate	0,45	4,13
chinmo	E0-12h	Affymetrix v2.0	a-chinmo-EB	7 054	broad	0,96	1,12
cnc	E0-12h	Affymetrix v2.0	KW0-CNC	699	punctate	0,89	2,70
CtBp	E0-12h	Affymetrix v2.0	KW0-dCtBP7667	4 947	punctate	0,95	2,19
D	E0-8h	Affymetrix v2.0	KW3-D-D2	2 979	punctate	0,90	3,39
Dfd	E16-24h	Affymetrix v2.0	KW0-dCtBP7667	581	punctate	0,41	47,99
Dfd	L3	Solexa	KWG-GFP	3 159	punctate	0,38	123,84
disco	E0-8h	Solexa	KW3-disco-D2	1 723	punctate	0,00	135,12
disco	E8-16h	Solexa	KW3-disco-D2	616	punctate	0,45	69,50
disco	E0-8h	Affymetrix v2.0	KW3-disco-D2	2 628	punctate	0,94	2,41
Dll	E0-12h	Affymetrix v2.0	a-dll-SC	75	broad	0,93	4,56
E(z)	E8-16h	Agilent 1M	KW4-E(z)-D2	1 927	punctate	0,60	3,93
EcR	Pupae	Solexa	KWG-GFP	483	mixed	0,36	430,24
EcR	L3	Solexa	KWG-GFP	603	mixed	0,51	150,51
EcR	Pupae	Solexa	KWG-GFP	508	mixed	0,63	225,02
en	E0-12h	Affymetrix v2.0	KWG-GFP	3 568	punctate	0,95	1,77
en	E7-24h	Affymetrix v2.0	a-end300	286	punctate	0,99	2,38
en	E12-24h	Affymetrix v2.0	KWG-GFP	1 502	punctate	0,97	2,31
eve	E1-6h	Agilent 1M	KWG-GFP	1 738	punctate	0,24	1,78
exd	E0-8h	Affymetrix v2.0	KWG-GFP	4 483	punctate	0,75	3,41
GATAe	E0-8h	Affymetrix v2.0	KW4-GATAe-D1	901	punctate	0,95	2,51
Gro	E0-12h	Affymetrix v2.0	KW0-GRO	626	broad	0,97	3,46
Gro	E0-12h	Affymetrix v2.0	KW0-GRO	1 338	broad	0,94	5,94
gsb-n	E7-24h	Affymetrix v2.0	a-gsbn-FM	765	punctate	0,98	2,36
h	E0-8h	Affymetrix v2.0	KW3-h-D1	1 944	punctate	0,95	2,53
hkb	E8-16h	Solexa	KW3-hkb-D1	1 623	punctate	0,55	97,55
hkb	E0-8h	Affymetrix v2.0	KW3-hkb-D1	1 269	punctate	0,94	2,50
HP1b	E16-24h	Affymetrix v2.0	a-HP1-Covance	3 396	mixed	0,94	2,79
HP1b	E16-24h	Affymetrix v2.0	a-HP1-Abcam	6 967	mixed	0,90	2,17
inv	E0-12h	Affymetrix v2.0	KW0-INV7657	3 222	mixed	0,86	2,09
jumu	E0-8h	Affymetrix v2.0	KW3-jumu-D2	943	punctate	0,94	2,74
kn	E0-12h	Affymetrix v2.0	KW0-KN7697	792	punctate	0,97	1,64
kni	E8-16h	Solexa	KW3-kni-D2	662	punctate	0,58	310,90
Kr	Kc167	Affymetrix v2.0	KW3-Kr-D2	2 809	punctate	0,81	3,58
Kr	E0-8h	Affymetrix v2.0	KW3-Kr-D2	869	punctate	0,97	3,71
Pl	E0-8h	Solexa	KW4-Pcl-D2	2 457	punctate	0,36	135,12
run	E0-12h	Affymetrix v2.0	KW0-RUN7659	333	punctate	0,96	1,55
sens	E4-8h	Affymetrix v2.0	a-sens-HB	11 773	mixed	0,99	13,60
sens	E4-8h	Affymetrix v2.0	KWG-GFP	16 070	mixed	0,98	22,04
Sin3A	E0-12h	Solexa	a-Sin3A-RC	4 046	punctate	0,87	114,77
Snr1	Pupae	Affymetrix v2.0	a-snr1-AD	280	mixed	0,96	4,03
Stat92E	E0-12h	Affymetrix v2.0	a-STAT92E-EB	105	punctate	0,96	2,18
tll	E0-4h	Affymetrix v2.0	KWG-GFP	97	punctate	0,97	3,98
Trl	E3-8h	Affymetrix v2.0	GAF3558	6 438	mixed	0,94	1,11
Trl	Kc167	Affymetrix v2.0	KW3-Trl-D2	7 692	mixed	0,86	4,44
Trl	E16-24h	Solexa	KW3-Trl-D2	5 195	mixed	0,87	539,95
Trl	Kc167	Solexa	KW3-Trl-D2	3 842	mixed	No replicate	147,97
ttk	E0-12h	Affymetrix v2.0	KW0-TTK7691	384	punctate	0,94	2,37
Ubx	E3-8h	Affymetrix v2.0	KW0-UBX7701	729	mixed	0,97	4,45
Ubx	E3-8h	Affymetrix v2.0	a-Ubx2-MK	161	mixed	0,96	5,00
Ubx	E0-12h	Affymetrix v2.0	KW0-UBX7701	1 300	mixed	0,94	3,98
zfh1	E0-12h	Affymetrix v2.0	KW0-ZFH17684	895	punctate	0,95	1,67

C

Factor	Stage/Cell Type	NumOfTFBS
HDAC1	E0-12h	4 468
HDAC11	E0-12h	2 301
HDAC3	E0-12h	2 588
HDAC4a	E0-12h	5 960
HDAC6	E0-12h	4 983
CTCF_N_Kc	Kc	2 024
CTCF_N	E0-12h	2 534
CTCF_N_S2	S2	2 254
CTCF_C	E0-12h	3 156
CP190	E0-12h	6 654
Mod(modg4)	E0-12h	3 062
Trl	E0-12h	3 906
BEAF-32	E0-12h	4 711
su(Hw)-1	E0-12h	3 422
su(Hw)-2	E0-12h	3 632

Supplementary Table 1

Supplementary Table 2

	E0-4		E4-8		E8-12		E12-16		E16-20		E20-24	
ChIP-Seq	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**
CBP	3 276	168	2 276	329	0	0	3 996	334	3 486	742	3 696	218
PolII	1 376	155	5 151	1 504	1 277	2 424	2 885	982	10 811	852	2 410	168
H3K4Me1	6 158	541	2 056	498	5 260	759	17 047	1 063	12 323	1 507	9 998	218
H3K4Me3	3 803	1 128	4 040	875	5 144	767	4 502	781	7 069	1 221	9 124	508
H3K9Ac	5 508	617	756	1 586	4 831	1 679	7 520	368	7 937	1 267	13 390	490
H3K27Ac	3 654	1 050	1 119	3 676	4 402	865	10 592	531	11 250	600	7 500	753
H3K9Me3	304	16 282	328	15 082	340	15 232	327	38 464	340	18 932	307	47 064
H3K27Me3	427	31 964	242	16 082	230	19 632	147	42 600	292	21 032	321	39 764

	E0-4		E4-8		E8-12		E12-16		E16-20		E20-24	
ChIP-chip	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**	**# of BS**	**median. length**
CBP	2 215	846	7 719	1 027	329	195	7 704	719	4 368	422	4 787	916
PolII	3 196	385	5 029	998	0	0	5 786	641	1 584	90	0	0
H3K4Me1	5 202	528	6 140	590	9 781	1 131	9 097	1 310	6 582	583	5 199	1 080
H3K4Me3	12 923	31	3 499	997	6 849	1 116	6 507	1 270	7 173	1 188	6 899	1 298
H3K9Ac	5 111	1 004	5 836	570	6 427	882	6 528	562	2 364	433	3 215	929
H3K27Ac	5 409	1 130	6 922	614	8 087	1 194	6 958	622	2 903	923	3 211	989
H3K9Me3	2 958	122	805	12 000	5 522	523	725	12 000	432	10 500	361	21 000
H3K27Me3	4 425	1 005	944	4 000	5 822	537	850	5 500	433	20 000	443	19 000

L1		L2		L3		Pupae		AdultFemale		AdultMale	
# of BS	median. length	# of BS	median. length	# of BS	median. length	# of BS	median. length	# of BS	median. length	# of BS	median. length
657	1 068	0	0	2 176	529	5 312	359	6 046	554	4 645	2 078
3 015	1 627	987	721	5 566	779	5 888	742	0	0	0	0
1 409	1 631	10 818	553	5 110	368	6 854	644	2 787	255	1 117	1 578
5 517	1 549	4 482	1 060	4 819	2 102	2 320	808	3 614	1 063	5 886	482
1 864	470	4 667	1 536	2 769	2 261	1 154	907	4 585	563	2 757	209
1 984	678	3 755	1 335	4 160	1 013	3 113	1 884	7 672	494	3 776	305
306	20 782	400	19 532	0	0	244	20 700	0	0	0	0
208	28 150	351	47 664	174	25 532	276	31 282	0	0	123	86 664

L1		L2		L3		Pupae		AdultFemale		AdultMale	
# of BS	median. length	# of BS	median. length	# of BS	median. length	# of BS	median. length	# of BS	median. length	# of BS	median. length
653	189	0	0	1 838	101	6 385	947	7 060	504	6 770	478
0	0	4 822	765	6 110	528	4 010	455	2 390	409	0	0
2 454	453	748	352	7 889	1 349	5 477	1 010	5 535	1 045	2 887	1 095
5 665	1 355	6 404	703	6 281	1 390	5 505	1 099	2 560	436	6 547	1 134
3 413	959	4 931	1 124	5 525	987	5 574	419	4 894	462	6 071	1 108
452	368	2 342	471	1 253	839	6 676	585	0	0	7 936	510
391	18 000	257	21 000	315	18 000	311	25 000	40	4 000	466	6 500
562	8 000	308	14 000	2 683	8 000	468	11 500	207	31 000	3 998	541

Supplementary Table 4

Fragment Designation	Chromosome	Start	End	5' gene	3' gene	Prediction Criteria	phiC31 docking site	Embryonic Expression	Other Expression, if tested	# lines assayed	Notes	Forward Primer	Reverse Primer
EO001	chr2L	518500	520000	cbt	ush	CBP peak in embryonic stages only	attP2	Observed	NA	1		GCAGTTCAGCGTCATCGTCATCGT	CCTTGGCTACTCCACCGCTTCAGA
EO002	chr2L	2164250	2165750	aop	aop	CBP peak in embryonic stages only	attP2	Observed	NA	1		TGGCATCCAGCTCGCATATCACTT	AGCGGGTACAGGCACACAGACACA
EO004	chr2L	2181200	2182400	aop	CG10874	CBP peak in embryonic stages only	attP2	Observed	NA	1		AATGCGCCTGTCTGGAGCGTATGT	TGTACTCCGATTCACCCCGACCAC
EO005	chr2L	2182400	2183600	aop	CG10874	CBP peak in embryonic stages only	attP2	Observed	NA	3		TCGCAGATCGAAGCAATCCACAAG	CACCCGATCTTCAAAAAGACGCACA
EO006	chr2L	2193100	2194300	CG34172	CG31668	CBP peak in embryonic stages only	attP2	no pattern	NA	2		GCAAAAGGGAATGCCAGAGAAATGC	AACCGATCCGAGCGTAAGCGTTTT
EO010	chr2R	2492800	2494200	CG15233	CG15234	CBP peak in embryonic stages only	attP2	Observed	NA	1	in intron of jing	TGCGACCAATCGAAGAGAATTCAA	TCATACGGTCGACAGATGGACATGG
EO013	chr2R	3046200	3047600	nec	pk	CBP peak in embryonic stages only	attP2	Observed	NA	2	in intron of pk	AGTATTCCCCGGTGGCTGGAGAGA	ACGGACAGCAATCGCTCCTCGTTT
EO015	chr2R	4530200	4531600	CG8635	ptc	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of ptc	TGATAAATGGGCGCACGCCTAGAG	AACACCCCACTGGCAACGAGAAGA
EO017	chr2R	5790800	5792000	CG1441	Fmrf	CBP peak in embryonic stages only	attP2	Observed	NA	3		TGCCACTGAAATCGTCTTCGGATTC	GTCAACTGACCAACAGGCCGCTTC
EO019	chr2R	6579000	6580200	CG12934	stan	CBP peak in embryonic stages only	attP2	Observed	NA	3		CCGCTGCGTGCGGTAATGTGTAAT	GCGACCACAAAGAGCTCGAATGCT
EO021	chr2R	7908000	7909200	otk	CG8964	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of otk	GGGAGGAAATCCATTGGTGGCTTG	ACGATTGCGAGGCGTTGAGTTACG
EO023	chr2R	12527000	12528400	Alk	gprs	CBP peak in embryonic stages only	attP2	Observed	NA	1		CGGGCCCGTAAAATGTTTAGGGATG	GGGCTATCGGACCACTTGACATGC
EO027	chr2R	15990100	15991600	CG16898	18w	CBP peak in embryonic stages only	attP2	Observed	NA	3	Pattern partial overlap with published 18w expression; likely enhancer of 18w	ATCCGGAGCAACTGCCACTTCAAA	GCCTACCGCCATTTCTGCTTTGGT
EO029	chr2R	18569200	18570400	RYBP	ppa	CBP peak in embryonic stages only	attP2	Observed	NA	1		GCCGCGATTAGTCATCGAATGCTT	CCAACATTTTCCATCAGTTTTCGGCTTA
EO032	chr3L	612900	614400	Reg-2	CG12030	CBP peak in embryonic stages only	attP2	Observed	NA	3		GACAGACATTGGCCAGACGCAGAA	ACCACCCATTCGTGACTTCAGCAG
EO034	chr3L	6771600	6772800	CG32392	vvl	CBP peak in embryonic stages only	attP2	Observed	NA	2	likely enhancer of vvl	GGAAATCGGTTGTCGCATTGAAAA	GGATGCCTTGCCAACCTGTCTCAC
EO036	chr3L	6792000	6793200	vvl	Prat2	CBP peak in embryonic stages only	attP2	no pattern	NA	1		TGCCCAACTAAAATGTAGCCGCTTG	GCTTGACTCCGGGATGTGCTGTTC
EO039	chr3L	10835600	10836800	tna	tna	CBP peak in embryonic stages only	attP2	Observed	NA	1		TCGGCTGGGAAGTACCCTCTAACGA	GAAAAAGCCGTCAACCACCAGCAC
EO042	chr3L	13074200	13075600	CG34429	CG17300	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of trn	TGGGACTTTGTTCGATGGCATGGT	GCTGTGCTGGAATGGGGAGATTTG
EO044	chr3L	13856000	13857200	CG8745	CG8745	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of CG8745	GACTCACACACGCCCCCGATAAGA	GGGGCTCATCAACAGGGTCGAAAA
EO046	chr3L	14506600	14507800	bbg	CG9592	CBP peak in embryonic stages only	attP2	no pattern	NA	1		TGGCTTCAAACCAAACAAAATGATCG	GCGATGGCAAACGAGTCAAGTCAG
EO050	chr3L	20915000	20916400	CG11458	fng	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of fng	CTTTGGCGGCCTTTGTTTTTGTGA	GAGGGGACTGCATCTCCGATCTCA
EO053	chr3R	2549300	2550700	pb	pb	CBP peak in embryonic stages only	attP2	Observed	NA	1	contains aspects of both pb and zen2 expression	CCCGGAGCGGCACAATTAGTCTTG	CGGTAATGCTGAATGAACCTTTCAA
EO055	chr3R	4633400	4634600	CG8359	CG9837	CBP peak in embryonic stages only	attP2	Observed	NA	1		GCCACTACTTTCTTGGGCCGGATG	ATGGGTCGACACAATCGCTGCAC
EO058	chr3R	5731600	5733000	CG34360	CG34360	CBP peak in embryonic stages only	attP2	Observed	NA	2		CACACTCGCACACACACACAAGCA	TGCCCAAACGATTTCAGCTTTTGC
EO060	chr3R	6179600	6180800	jumu	jumu	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of jumu	TGTCCTCGTTTTCCCCCATTTGAA	TTGCTTTTCGAATGCAGCCACCAT
EO079	chr4	400400	401800	CG2052	CG2052	CBP peak in embryonic stages only	attP2	Observed	NA	1		CGGACATGGTCAGATCGACTTGGA	GCAGCTGAGTCGGCACTGCAAATA
EO087	chrX	2990900	2992000	kirre	kirre	CBP peak in embryonic stages only	attP2	Observed	NA	2		TCCGTCGCCCATCACTCTCTCTCT	CCTGCGATTGGGAATGGGGTTAAA
EO089	chrX	3261400	3262800	dnc	dm	CBP peak in embryonic stages only	attP2	Observed	NA	2		CCCCGACGATAACCTCAAGTGCAA	CCGGGCGAAAAGCAAAACAAACTT
EO092	chrX	8941400	8942400	CG15364	CG10962	CBP peak in embryonic stages only	attP2	Observed	NA	1	in introns of CG10962 and CG42388	CTCGTTTCCGAGGCTTCACTGTGG	TCCGAAATAGCCATCCGTTATACCC
EO095	chrX	13143400	13144900	mew	mew	CBP peak in embryonic stages only	attP2	Observed	NA	1	contains aspects of both mew and CG15743 expression	TGCAGCTGGGTTTTACAGCAACGA	TCGGGGTGAACCAAATGAAGATCG
EO101	chrX	20360200	20361400	RhoGAP19D	RhoGAP19D	CBP peak in embryonic stages only	attP2	Observed	NA	1		TAAGAGCGGCAAGCGGAGGATGTA	ATCGGTATCGACCAACTGGCCACA
EO103	chrX	20560000	20561200	hydra	run	CBP peak in embryonic stages only	attP2	Observed	NA	1	likely enhancer of run	CGATGCCGATGATCACGAAAAGTG	AAAGCGACTGCCAATCGAGGAACA

Supplementary Table 5

TFBS Complexity Categories	Number of Binding Regions	Average Length of Binding Regions	Expression Median	Number of Transcripts	Number of Active Transcripts	Percentage of Active Transcripts
1	22 655	489	9	1 882	1 439	76,5%
2	5 295	986	10	1 113	937	84,2%
3	2 975	1 163	16	806	725	90,0%
4	2 076	1 259	14	669	600	89,7%
5	1 577	1 322	14	520	470	90,4%
6	1 227	1 360	13	388	350	90,2%
7	795	1 389	16	251	240	95,6%
8	643	1 422	11	180	166	92,2%
9	462	1 431	17	149	125	83,9%
10	298	1 425	13	84	76	90,5%
11	210	1 435	11	61	55	90,2%
12	142	1 452	7	46	40	87,0%
13	89	1 452	20	27	25	92,6%
14	53	1 457	14	18	15	83,3%
15	33	1 493	106	6	4	66,7%
16	18	1 503	92	4	4	100,0%
17	8	1 450	0	0	0	-
18	1	1 441	0	0	0	-
19	3	1 354	8	2	2	100,0%
20	1	1 529	5	1	1	100,0%
21	1	1 529	0	0	0	-

Supplementary Table 9

Panel	Donor	Class	Coordinates	Enhancer blocking
B	Rain7	Recipient Vector	3R:13373664	No
C	1A2	Positive control	X:255313..255772	Yes
D	SCS	Positive control	3R:7774458..7775524	Yes
E	Spacer	Negative control	2R:5863750..5864406	No
F	CP190-2894	CP190/CTCF	2R:5428851..5429464	Yes
G	CP190-11628	CP190/CTCF	X:13180486..13183206	Yes
H	CP190-4762	CP190/CTCF	2R:20199584..20201894	Yes
I	CP190-7635	CP190/CTCF	3R:2696003..2697000	Yes
J	CP190-9220	CP190/CTCF	3R:17231264..17234459	Yes
K	CP190-11742	CP190/CTCF	X:14794973..14797250	Strong
L	CP190-11742	CP190/CTCF	X:14794973..14797250	Weak
M	CP190-8562	CP190/CTCF	3R:11318599..11320846	Weak
N	CP190-11319	CP190/CTCF	X:9904070..9904636	Weak
O	GAF-Antp1	GAF	3R:2718919..2719623	No
P	GAF-Fab4	GAF	3R:12683547..12683918	No
Q	CP190-12404	CP190/Su(Hw)	X:20953657..20955060	No
R	CP190-8767	CP190/Su(Hw)	3R:12810442..12811667	No
S	CP190-3557	CP190/Su(Hw)	2R:10272132..10274060	No
T	CP190-2738	CP190/Su(Hw)	2R:4091291..4092348	No
U	CP190-4423	CP190/Su(Hw)	2R:17837219..17838527	No

Supplementary Table 16

Cluster	Enhancer-class	enrichment	z-score	p-value	FDR
Cluster_6	CAD	-2,64338	-3,744268	9,05E-05	0,001176
Cluster_7	CAD	-2,28832	-7,039395	9,68E-13	1,26E-11
Cluster_7	blastoderm	-5,61099	-5,496367	1,94E-08	1,26E-07
Cluster_7	ectoderm-agg	-2,92467	-3,068896	0,001074	0,004655
Cluster_9	CAD	1,52371	6,13716	4,20E-10	5,46E-09
Cluster_9	dorsal-mesothoracic-disc	2,71558	4,994863	2,94E-07	1,91E-06
Cluster_9	nervous-system-agg	1,78474	3,063193	0,001095	0,004745
Cluster_9	ventral-thoracic-disc	2,20168	2,745959	0,003017	0,009804
Cluster_10	CAD	-1,38573	-3,194582	0,0007	0,009102
Cluster_11	trunk-mesoderm-primordium	3,86704	7,757173	4,33E-15	5,63E-14
Cluster_11	mesoderm-agg	3,24349	6,490468	4,28E-11	2,78E-10
Cluster_11	somatic-muscle-primordium	3,92393	6,296446	1,52E-10	6,60E-10
Cluster_11	CAD	1,65151	4,955475	3,61E-07	1,17E-06
Cluster_11	muscle-agg	2,73569	4,868031	5,64E-07	1,41E-06
Cluster_11	embryonic-larval-somatic-muscle	3,28202	4,838823	6,53E-07	1,41E-06
Cluster_13	trunk-mesoderm-primordium	3,12974	4,975485	3,25E-07	4,23E-06
Cluster_14	CAD	2,4143	7,331396	1,14E-13	1,48E-12
Cluster_14	embryonic-larval-somatic-muscle	3,6358	3,764073	8,36E-05	0,000543
Cluster_15	embryonic-ventral-nervous-system	1,53787	3,345155	0,000411	0,005345
Cluster_16	CAD	1,92435	7,606604	1,41E-14	1,83E-13
Cluster_16	ectoderm-agg	2,61007	6,022039	8,61E-10	5,60E-09
Cluster_16	ectoderm	2,85319	5,05529	2,15E-07	9,31E-07
Cluster_16	dorsal-mesothoracic-disc	2,51839	3,899505	4,82E-05	0,000157
Cluster_16	blastoderm	1,60958	3,237883	0,000602	0,001464
Cluster_16	trunk-mesoderm-primordium	2,4207	3,204765	0,000676	0,001464
Cluster_16	mesoderm-agg	1,9847	2,979686	0,001443	0,002679
Cluster_18	embryonic-ventral-nervous-system	2,23467	4,231213	1,16E-05	0,000151
Cluster_18	nervous-system-agg	1,93424	3,738433	9,26E-05	0,000602
Cluster_22	CAD	2,94129	16,07457	0	0
Cluster_22	blastoderm	4,64349	29,03857	0	0
Cluster_22	ectoderm	3,66121	8,617994	0	0
Cluster_22	ectoderm-agg	3,55236	11,1538	0	0
Cluster_22	mesoderm-agg	2,24827	3,383356	0,000358	0,000931
Cluster_22	muscle-agg	1,98994	2,610632	0,004519	0,009791

Supplementary Figure 1. Pair-wise overlap enrichment between datasets (block bootstrap enrichment Z-score, from <-5 (blue) to >80 (red)) generated by our group, the BDTNP, and regulatory element predictions from CAD. The RNAseq time course was used to segregate all transcripts into 4 quartiles by FPKM. All factors studied for the chromatin time-course project (marked with 't') have been ordered per factor by developmental stage.

Supplementary Figure 2. Example of the distributions of the 8 chromatin marks studied. Below each ChIP-seq track, boxes indicate regions of enriched signal. These profiles all correspond to ChIP-seq data from the pupal stage. Note the striking difference between the distributions of H3K9me3 and H3K27me3 (in blue) and all other marks. Conversely H3K4me3, H3K9Ac, H3K27Ac and H3K4me1 (purple) all exhibit an occupancy profile similar to that of PolII (red). CBP (green) is also correlated to the RNAseq coverage (orange).

Supplementary Figure 3. Morphology of the TF binding data.
This browser shows different binding site distributions for different factors. While some factors mainly bind narrow peaks (ex. Bab1 and BRM), others mainly bind large domains (ex. DLL and GRO) while still others bind a combination of both (ex. BKS and CHINMO).

Supplementary Figure 4. (A) Distribution of the number of genes marked (y-axis) by 6 histone modifications of chromatin modifying enzymes (colors), plotted against the number of developmental stages the gene is marked in (x-axis). (B) Pair-wise overlap enrichment between non-TF datasets (block bootstrap enrichment Z-score, from <-5 (blue) to >80 (red)). The RNAseq time course was used to segregate all transcripts into 4 quartiles by FPKM. All factors studied for the chromatin time-course project (marked with 't') have been ordered per factor by developmental stage.

Supplementary Figure 5.(A) Prediction of novel promoters. Number of novel promoter predictions (y-axis) per developmental stage are depicted in grey bars, cumulative total of unique predictions in black dots. Distribution of H3K4Me3 (grey) and PolII (black) marks relative to gene TSSs depicted in inset. (B) Novel promoter prediction validation. Individual experiments, in triplicate, are represented as a single bar. Mean log10 transformed, normalized luciferase measurements from constructs (x-axis) with inserts in the forward (blue) and reverse (green) orientations (y-axis). Black lines depict standard error. The central portion of the graph depicts the validation of novel promoter predictions based on data from 0-12 hour embryos, while the right depicts validation of novel promoter prediction from Kc cell data.

Supplementary Figure 6. H3K9me3 defines heterochromatic regions. In our chromatin time-course experiments, H3K9me3 is largely overlapping with H3K27me3 domains. Using peptide competition assays followed by ChIP we were able to demonstrate that this overlap is resulting from an antibody cross-reactivity at this particular locus (data not shown). We detected the real H3K9me3 domains by comparison with HP1, a chromodomain protein that specifically binds this Histone modification. H3K9me3 is located in large domains at the centromeric end of chromosomes 2L, 2R, 3L and along the chromosome 4. Image shows example region of HP1 and

H3K9me3 binding overlapping in a heterochormatic region and H3K27me3 and H3K9me3 binding overlapping in a non-heterochromatic region.

Supplementary Figure 7. Percentage of genes associated with each factor conditional upon gene expression status during the time-course. The union of Agilent and Solexa peaks has been used for each factor to assign genes as "marked" or "unmarked" depending on the presence of a specific factor or Histone mark within -1kb to +1kb of the TSS. Genes have also been classified as "active" or "inactive" according to their sequencing coverage in the RNAseq experiments. The distribution of "marked" and "unmarked" genes is represented for (A,E) H3K27Ac, (B,F) H3K4me1, (C,G) H3K4me3, (D,H) H3K9Ac. On each graph, for each time point, the genes in red are active and the genes in blue are inactive. The genes in dark color are "marked" while the genes in light color are "unmarked". The red line separates the active genes from the inactive genes at all stages.

Supplementary Figure 8. Building a classifier of gene expression from chromatin marks. (A) Distribution of FPKM estimates for all12 developmental stages. (B) AUC values across a range of FPKM thresholds for models trained on each developmental stage separately. (C) Recovery of marked active genes across a range of FPKM thresholds for models trained on each developmental stage separately (FDR= 0.10). (D) ROC curves for the logistic regression classifier across multiple FPKM values for 12 developmental stages. Line colors correspond to different FPKM thresholds: red = 0.1, green = 0.5, blue = 1.0, cyan = 1.5, magenta = 2.0, black = 2.5.

Supplementary Figure 9. The classifier of gene expression detects an umarked active gene category. (A) Binary classifier outcome transcript distribution. Outcomes defined at FDR = 0.10. MA = marked active, MI = marked inactive, UA = unmarked active, UI = marked inactive. (B) Distribution of FPKM values for binary classier outcomes. **a,** MA vs. MI at FDR = 0.05. **b,** UA vs. UI at FDR = 0.05. **c,** MA vs MI at FDR = 0.10. **d,** UA vs UI at FDR = 0.10.

Supplementary Figure 10. Unmarked active genes have temporally restricted expression patterns. (A) Enrichment of FlyAtlas spatial expression terms for the unmarked active and marked active genes in the Adult male (a similar pattern is observed with adult female). Note that the marked active class is more enriched in tissue specific terms. (B) Predictability of active and inactive transcripts. **a-b,** Predictability of active transcripts is defined as the number of times a transcript is classified as marked (FDR = 0.05 (a), FDR = 0.10 (b)) normalized by the number of stages the transcript is active. **c-d,** Predictability of inactive transcripts is defined as the number of times a transcript is classified as unmarked (FDR = 0.05 (c), FDR = 0.10 (s)) normalized by the number of stages the transcript is inactive.

Supplementary Figure 11. Examples of active genes not associated to H3K4me3. This genome browser view shows the occupancy profile of H3K4me3 (purple) and the RNAseq coverage (orange) around the Trypsin gene complex on the chromosome 2R. Genes associated to H3K4me3 are highlighted. We can observe that they are all expressed at all stages investigated.

The Trypsin genes however, as well as the gene *sha* and *nompA* are transiently expressed and do not have H3K4me3 at their promoters.

Supplemental Figure 12. H3K4me3 unmarked, detected genes in Kc cells. (A-B) Seven representative examples of unmarked active genes observed in embryos and synchronized cell culture. (C) qPCR validation of unmarked active genes from synchronized cell culture.

Supplementary Figure 13. Domains of H3K27me3. (A) Genome browser example of the distribution of the repressive H3K27me3 mark along chromosome 2R over developmental time starting with embryos (turquoise) and progressing to adults (red) . Most domains appear to be present at all time-points, but a substantial fraction show some stage specificity (starred examples). (B) Genes within H3K27me3 domains have lower mean gene expression (RPKM) values than genes outside the domains, especially genes adjacent to H3K27me3 domains. (C) Clustering of 1264 H3K27Me3 domain genes by temporal dynamics. Domain genes (columns) are grouped into clusters based on the temporal pattern (y-axis) of Histone mark presence (blue) or absence (white) and are arranged along the x-axis.

Supplementary Figure 14. GO category analysis of H3K27me3 associated genes. (A) Summary of main enriched GO categories in the different clusters of H3K27me3 domains. (B) Example of GO terms enriched in a stage specific (pupae) cluster of H3K27me3 domain genes, with red indicating biological process GO terms, white indicating cellular components GO terms and blue indicating molecular function categories.

Supplementary Figure 15. Spatial restriction of genes stably associated with H3K27me3.(A) Example of Tomancak (2007) clusters of genes with similar expression profile from in situ experiments enriched or depleted in H3K27me3 domain gene cluster 89. H3K27me3 domain gene clusters are enriched for spatially restricted genes while the these clusters are depleted for ubiquitously expressed genes. This suggests that H3K27me3 is a default mechanism for the inhibition of these genes, which is lifted in gene-specific time and tissue-dependent manner. (B,C) Examples of genes with known expression pattern (FlyExpress, BDGP) within H3K27me3 domains showing spatially restricted expression (B, *midline*; C, *wingless*; green) and immediately adjacent genes showing ubiquitous expression (B, *CG6907*; C, *CG4567*; blue).

Supplementary Figure 16. HDACs are associated with TSSs, transcribed exons, and PREs. (A-B, D-E) Enrichment of HDAC binding sites (y-axis) around active (A; FPKM > 1) and inactive (D; FPKM < 1) metagenes (x-axis corresponding to 2000 bp upstream and 1000 bp downstream of the TSS and 1000bp upstream and 2000 bp downstream of the TES of genes). Each of five different HDACs is plotted as a seperate color, as labeled. FPKM estimates were derived from pooled RNAseq data from stages E0-4h, E4-8h and E8-12h. HDAC1, 4a, 6 and 11 show a strong enrichment at the TSS and depletion along the gene body. In contrast, HDAC3 shows a strong depletion at the TSS and an enrichment along the gene body. (B, E) Enrichment of 4 different Histone tri-methylations (y-axis) across active (B) and inactive (E) metagenes (x-axis). Note the striking similarity between (i) H3K4me3 and HDAC 1, 4a, 6 and 11 profiles and (ii) H3K36me3 and HDAC3 profiles. In contrast, genes defined as inactive have reduced enrichment of HDACs

at the promoter and a depletion along the gene body. Similar differences are also observed for the corresponding Histone tri-methylations. (C) HDAC enrichment (y-axis) is correlated with target gene expression level (x-axis). (F) HDAC enrichment (y-axis) at varying distances from PHO sites (x-axis). HDAC4a and 1 are strongly enriched in the proximity of PHO sites, while HDAC6 and 11 show a moderate enrichment and HDAC3 a strong depletion.

Supplementary Figure 17. Prediction of silencers. Flowchart of silencer prediction from HDAC1 and HDAC4a binding site data. The union of HDAC1 and HDAC4a binding sites (n=6191) was filtered for sites overlapping H3K4me3. Sites within the remaining 2521 sites that overlapped regions of H3K27me3 to predict 537 PREs.

Supplementary Figure 18. Examples of silencers. This IGB browser example is centered around the homeotic gene cluster ANT-C. The PC and PHO data are from [15]. Common binding regions for HDAC1 and HDAC4a are associated with either H3K4me3, GAF or PCL/PHO Binding Regions representing Polycomb Response Elements (blue squares).

Supplementary Figure 19. CBP and H3K4me1 are associated with enhancers.
This IGB browser example represents signal for CBP (green) and H3K4me1(pink) at three different time-points (Adult Male, Pupae and Embryos 0-4h) around the region of *even-skipped* that contains well characterized enhancers (represented by the REDFly track in purple). Also represented are the insulators, the blue dashed line representing Class I gene boundaries. In embryos, the several enhancers within the intergenic region around *eve* are bound by CBP and contain H3K4me1 signal. Note that both signals are not present later during development when these enhancers are not active.

Supplementary Figure 20. Clustering CBP bound regions. (A) Illustration of criteria used to associate experiments with CBP regions. (B) Bayesian information criterion score vs cluster number used in model training. (C) Enrichment of chromatin and PolII profiles within each CBP cluster. The rows of the enrichment map correspond to the CBP clusters 1-20, where the number of regions is indicated in the row label. Columns of the enrichment map correspond to chromatin time course experimental data. Each cell represents the enrichment (red) or depletion (blue) of each experimental binding site set within the binding sites of each CBP cluster.

Supplementary Figure 21. "CBP embryo only" enhancer validation examples. (A) CBP Chip-seq data, across the developmental time course, for genomic regions corresponding to enhancer predictions that were tested in Fig. 2f. (B) Additional examples of tested regions for which reporter expression overlaps aspects of available RNA *in situ* patterns for neighboring genes. EO017 overlaps the known expression of *CG1441*; EO050 overlaps the known expression of *fringe*; and EO060 overlaps the known expression of *jumeau* (known gene expression data from FlyExpress database).

Supplementary Figure 22. Insulator validation. (A) Diagram of the insulator validation strategy; A recipient *P* element integrated at 3R:13373664 containing the *even skipped* stripe 2 and 3 enhancer elements separated by an eye-expressed eGFP is used as a substrate for

Recombination Mediated Cassette Exchange, replacing the eGFP with a genomic DNA fragment. (B – U) Immunohistochemistry with an anti b-Galactosidase antibody to detect reporter expression. All stage 10/11 embryos are oriented anterior to the left dorsal to the top. (B – E) Control lines: (B) recipient construct showing strong expression in *eve* stripe 2 and 3 territories, with weaker expression in stripe 7 and cephalic territories. (C-D) The characterised *1A2* and *scs* insulators block stripe 3 expression. (E) A negative control spacer fragment from the *eve* locus shows no enhancer blocking activity. (F – N) Class I elements generally show enhancer blocking activity. Each fragment is associate with CTCF and CP190 binding. Strong: (F) 2894, (G) 11628, (H) 4762, (I) 7635, (J) 9220, (K &L) 11742 shows variable activity with some embryos showing strong enhancer blocking (K) and others weak (L). (M & N) 8562 (M) and 11319 (N) are class I elements that show weak enhancer blocking activity. (O & P) Two GAF positive regions show no enhancer blocking activity. (O) Antp1, (P) fab4. (Q – U) Class II elements that bind Su(hw) and CP190 show little or no enhancer blocking activity. (Q) 12404, (R) 8767, (S) 3557, (T) 2738, (U) 4432. See supplementary table 15 for full details of the assayed fragments.

Supplementary Figure 23. TF clustering, including HOT spots. Pair-wise enrichment for all transcription factor combinations, including TFBS overlapping HOT regions.

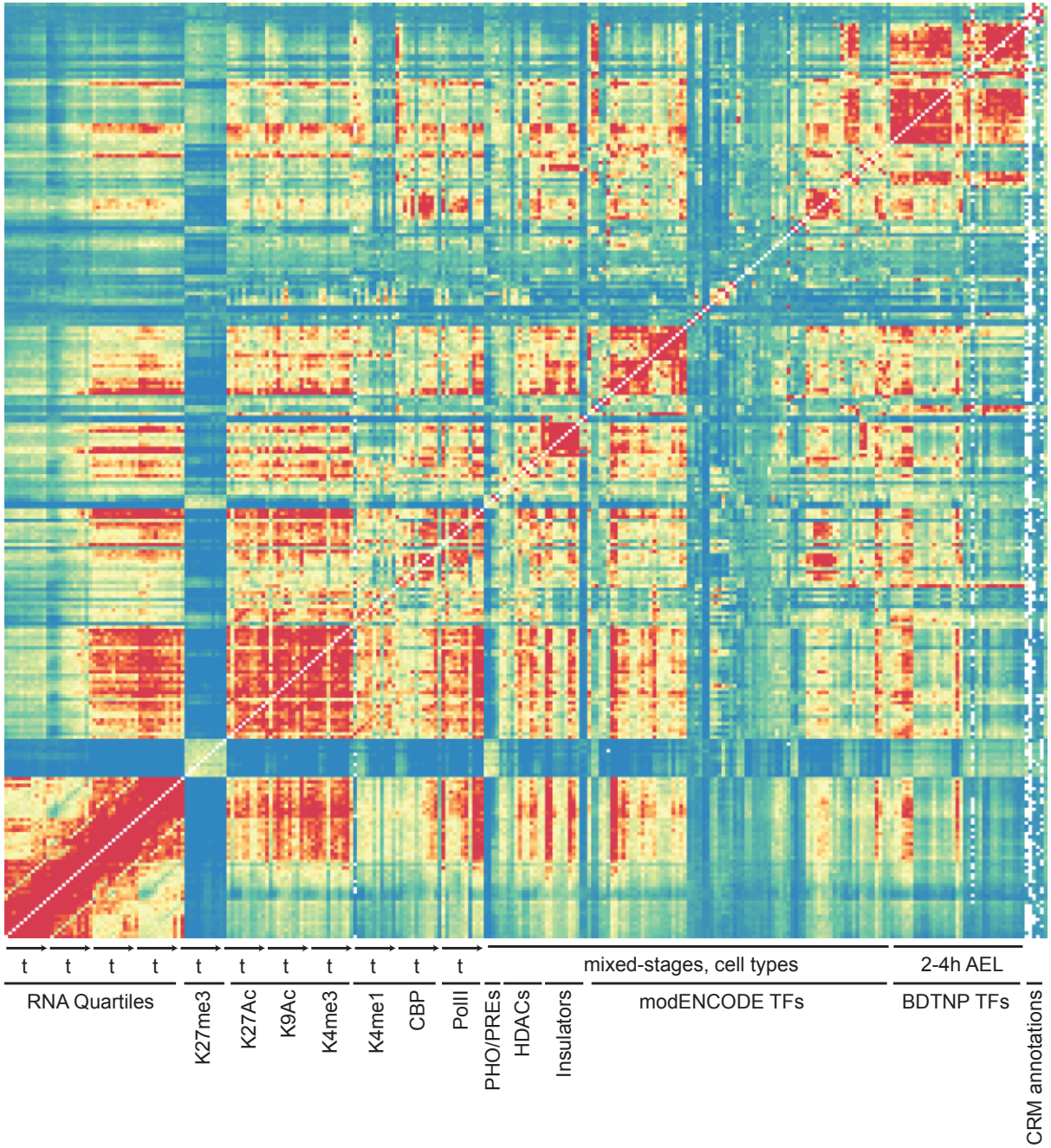
Supplementary Figure 24. Hotspot distributions. (A) Distribution of HOT regions (in red) over the genome in relation to GC content (gray scale). (B) Distribution of HOT regions (in red) over the genome in relation to gene density (gray scale). (C) The fraction of HOT regions that overlap with five classes of genomic annotation (5' UTR (dark blue), coding exon (orange), intron (purple), 3'UTR (green), and intergenic(ligght blue)), for each set of merged transcription factor binding sites, binned by complexity (x-axis). From Category 1 to 8, the proportion of intergenic and TSS regions covered increases at the expense of CDS and intron categories. (D) Heatmap depicting the –log transformed Fisher's exact test p-value quantifying the pairwise enrichment between each TF binding site set and merged binding site complexity categories.

Supplementary Figure 25. Transcription factor interactions and associated gene expression patterns. (A) Hierarchical transcriptional regulatory network defined by TFBS interactions between pairs of TFs in this and published data. Nodes (TFs) identified in this study in pink, those based on two previous studies in green and yellow. Previously identified edges (regulatory interactions) depicted in grey, those derived from this study in blue. Edges connecting factors whose binding sites significantly overlap (block bootstrap $Z > 10$) are depicted as red dashed lines. (B-C) Gene expression medoids (blue to red) for each of 64 and 18 k-means clusters (y-axis) derived from independent microarray (B) and RNAseq (C) transcription time courses, at each developmental stage (x-axis, labeled by stage). Metaclusters (described in main text) are boxed and labeled in roman numerals.

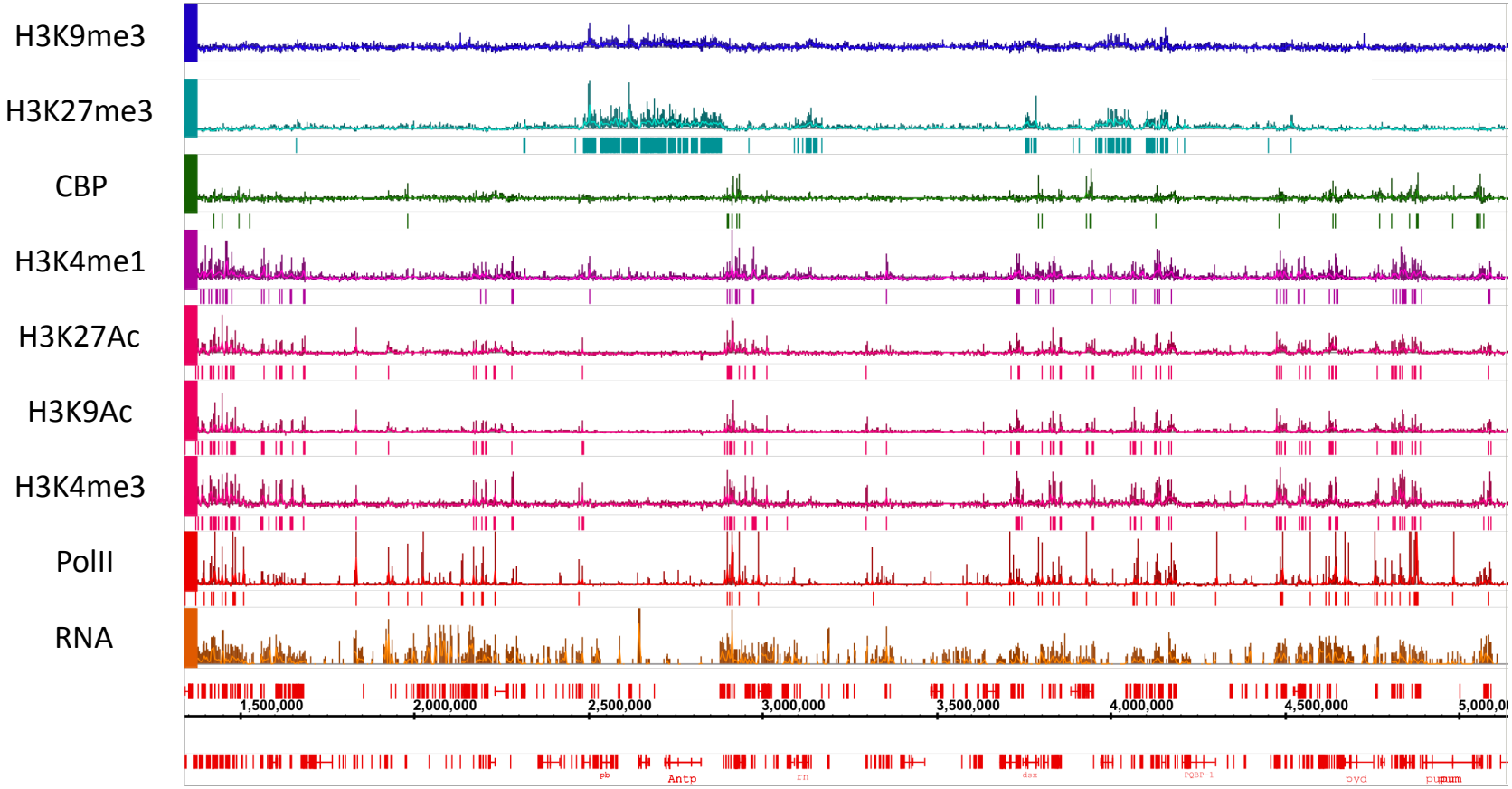
Supplementary Figure 26. TFBS interaction vignette.

Supplementary Figure 27. Networks constructed exclusively from Furlong et al. data (A) and BDTNP data (B).

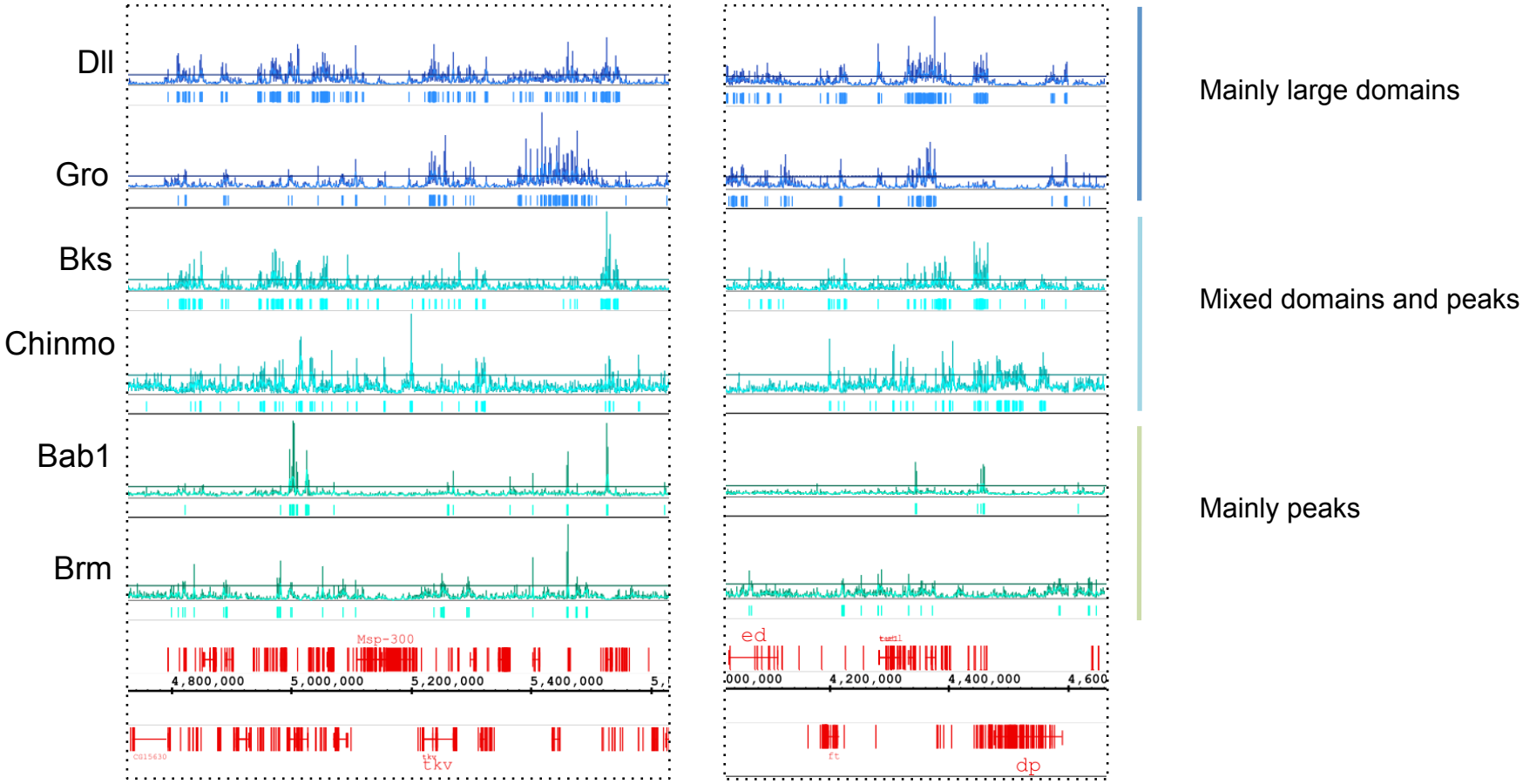
Supplementary Figure 28. Number of references found for each protein that we have studied in either PubMed (blue line) or FlyBase (red line).



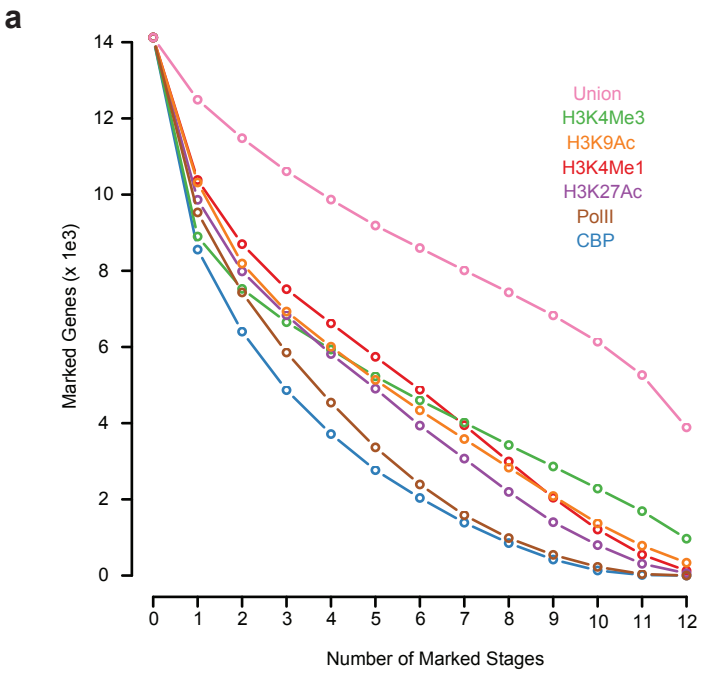
Supplementary Fig. 1

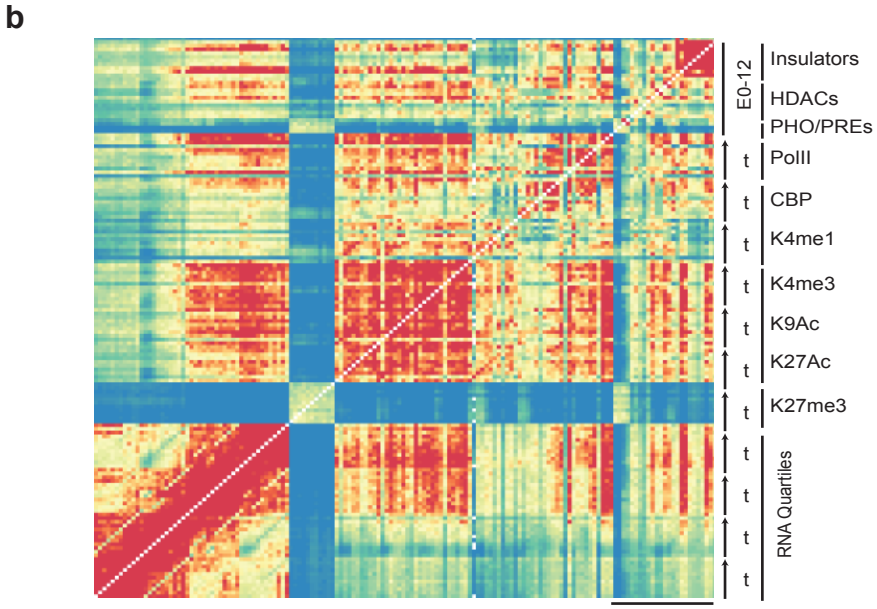


Supplementary Fig. 2

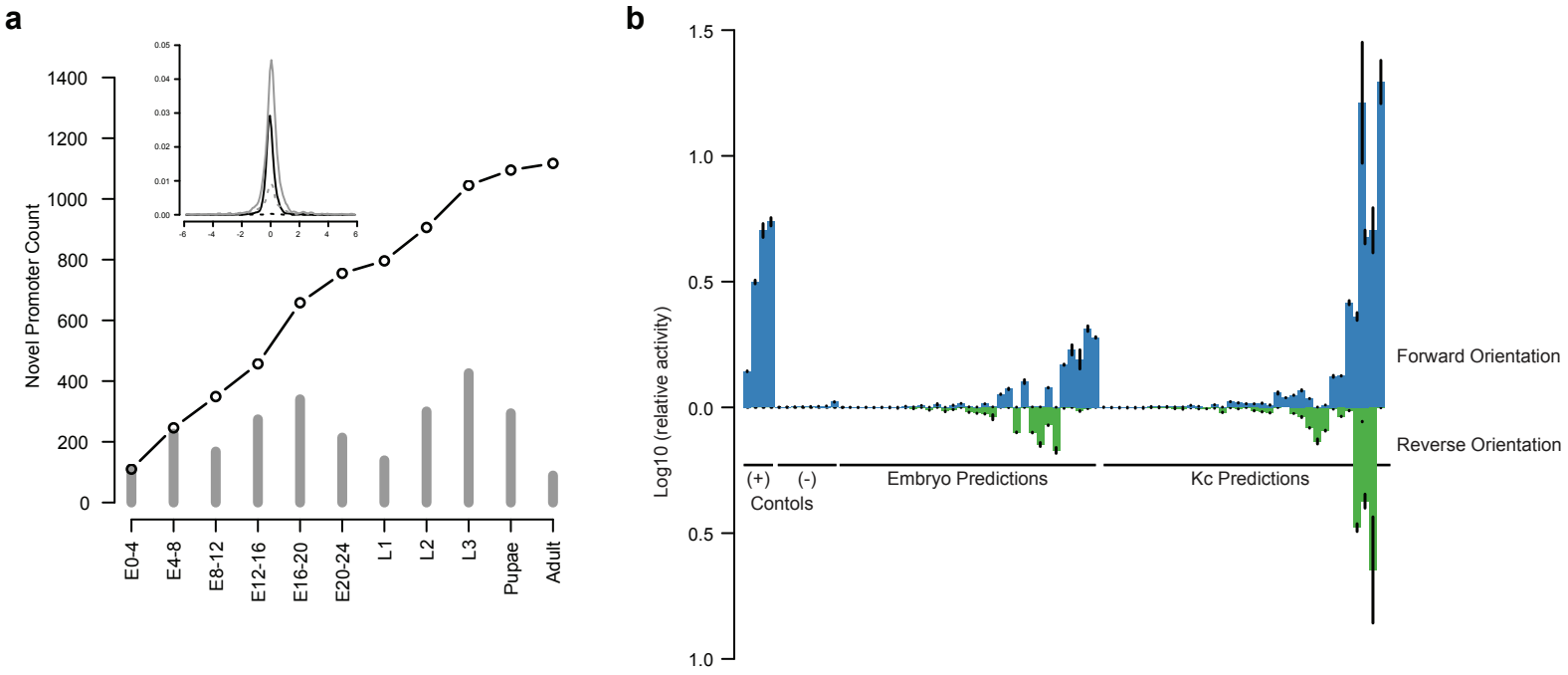


Supplementary Fig. 3

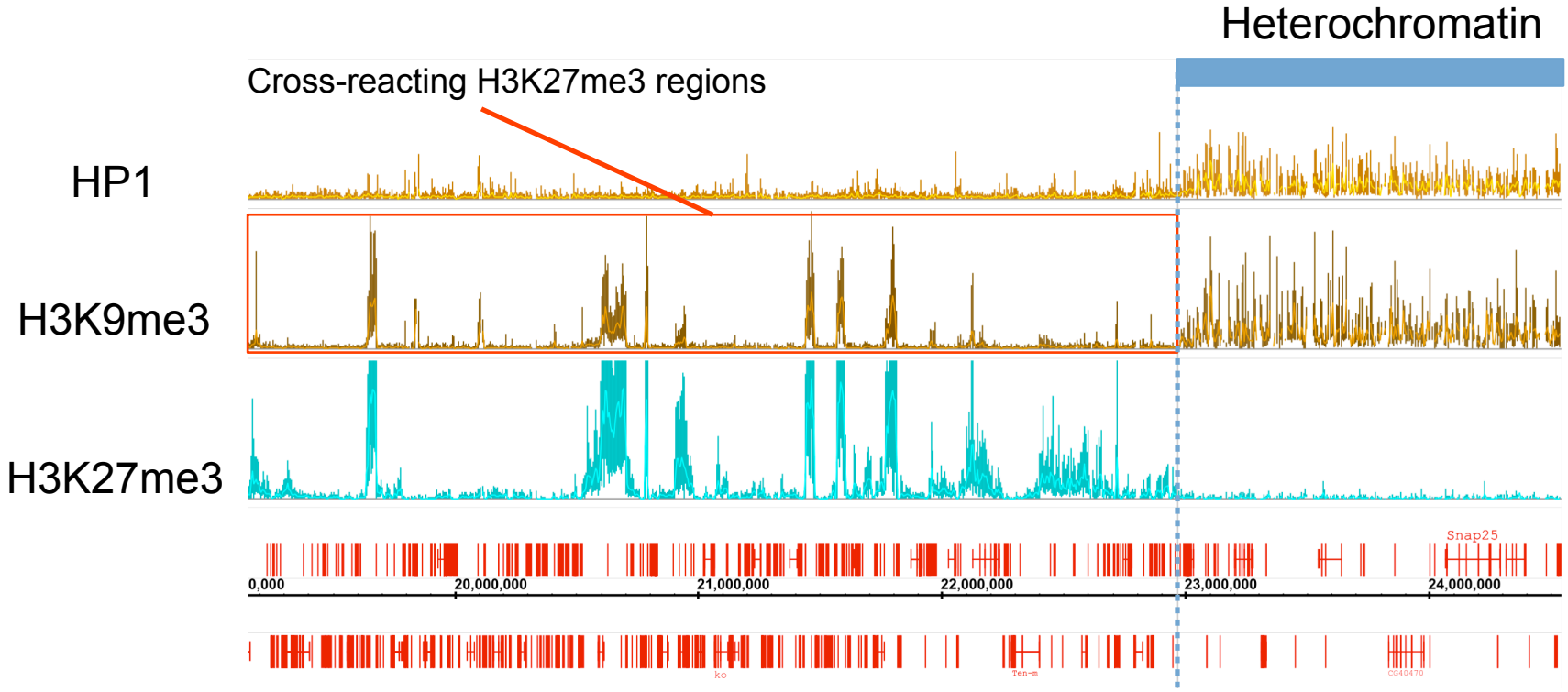




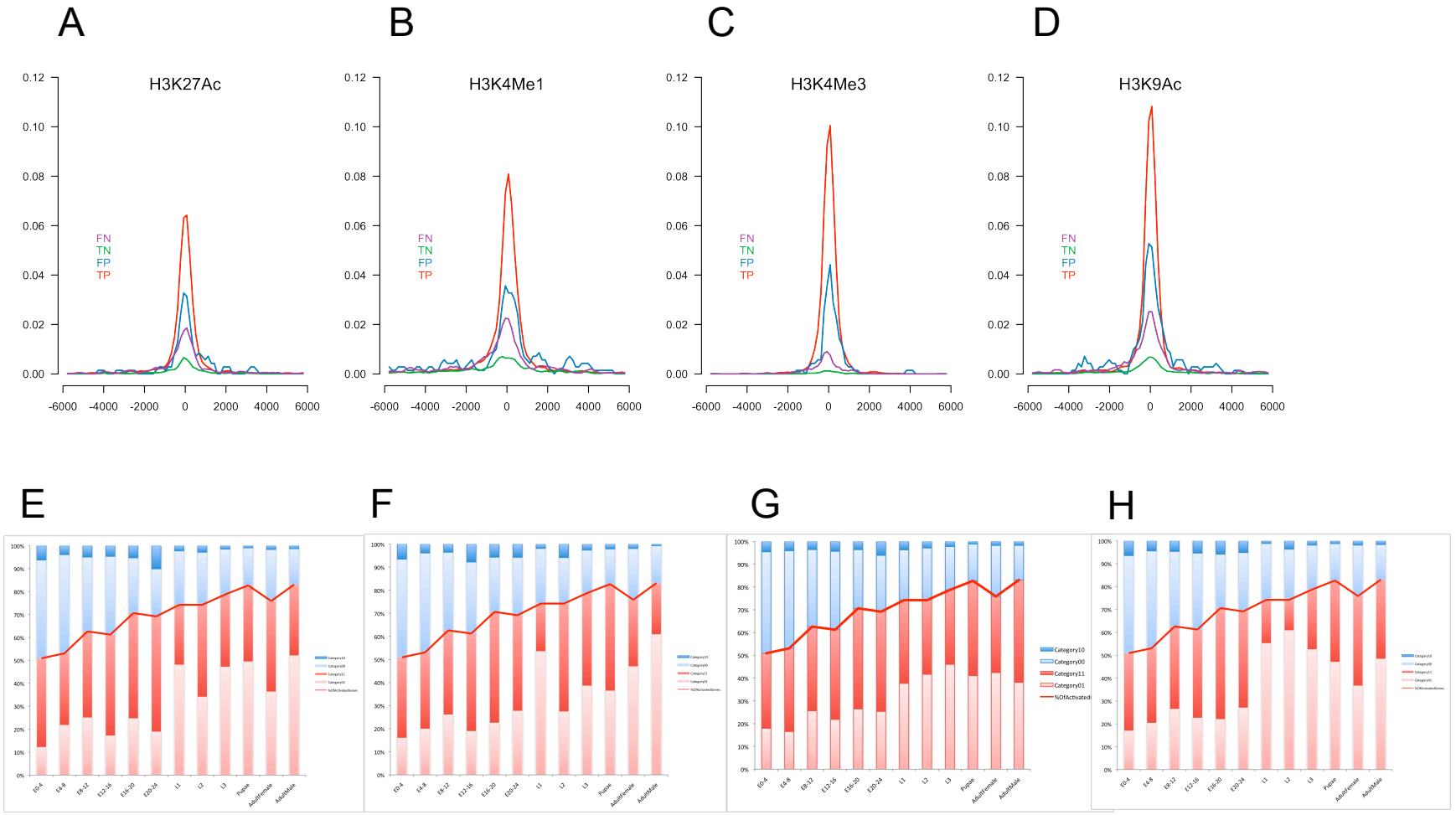
Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7

a

Developmental stage	0.001	0.01	0.05	0.1	0.5	1	5	10	100	1000	top	all
E0-4hr	5107	17	98	327	1240	437	925	715	4568	683	76	14193
E4-8hr	4236	10	48	200	1395	707	1754	1151	4068	545	79	14193
E8-12hr	2991	3	92	275	1427	742	2242	1451	4412	484	74	14193
E12-16hr	3284	9	33	125	1054	688	2442	1704	4313	469	72	14193
E16-20hr	1995	2	84	324	1325	635	2601	1864	4804	480	79	14193
E20-24hr	2626	11	30	111	775	594	2966	1936	4444	600	100	14193
L1	1747	6	123	341	1238	642	3001	2020	4351	610	114	14193
L2	2023	15	304	364	1057	760	3560	2138	3254	616	102	14193
L3	1451	8	79	213	1090	840	4138	2448	3353	478	95	14193
Pupae	1083	5	25	147	694	601	3135	2474	5369	602	58	14193
AdultMale	1033	4	20	123	693	604	3423	2498	5021	717	57	14193
AdultFemale	1520	4	32	173	1093	866	3158	2188	4516	531	112	14193

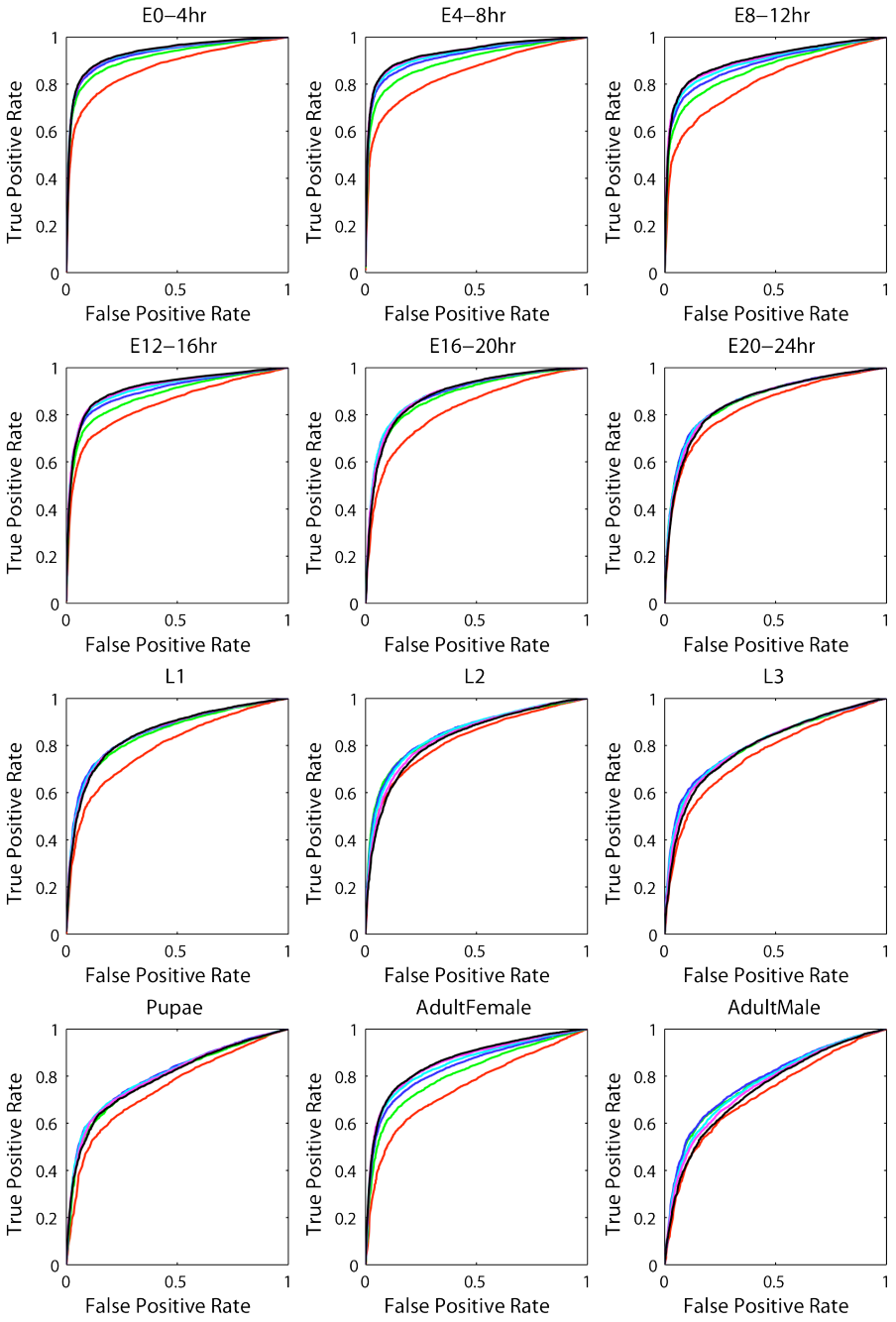
b

Developmental stage	RPKM threshold 0.1	0.5	1	1.5	2	2.5
E0-4hr	0.871	0.914	0.927	0.932	0.934	0.935
E4-8hr	0.847	0.896	0.918	0.927	0.932	0.933
E8-12hr	0.810	0.861	0.881	0.892	0.900	0.902
E16-20hr	0.821	0.878	0.889	0.893	0.890	0.887
E12-16hr	0.844	0.882	0.901	0.910	0.914	0.915
E20-24hr	0.833	0.862	0.866	0.863	0.860	0.857
L1	0.793	0.847	0.858	0.856	0.855	0.854
L2	0.819	0.852	0.854	0.851	0.840	0.831
L3	0.762	0.802	0.809	0.807	0.802	0.796
Pupae	0.749	0.792	0.805	0.803	0.799	0.792
AdultMale	0.722	0.770	0.779	0.768	0.757	0.745
AdultFemale	0.753	0.811	0.839	0.854	0.863	0.867

c

Developmental stage	RPKM threshold 0.1	0.5	1	1.5	2	2.5
E0-4hr	0.743	0.810	0.828	0.834	0.832	0.834
E4-8hr	0.735	0.796	0.831	0.839	0.845	0.844
E8-12hr	0.718	0.754	0.779	0.790	0.801	0.794
E16-20hr	0.853	0.844	0.837	0.821	0.789	0.757
E12-16hr	0.792	0.814	0.836	0.841	0.842	0.838
E20-24hr	0.836	0.824	0.802	0.775	0.738	0.678
L1	0.833	0.802	0.789	0.756	0.726	0.690
L2	0.816	0.788	0.748	0.696	0.613	0.504
L3	0.898	0.771	0.712	0.657	0.611	0.507
Pupae	1.000	0.870	0.801	0.756	0.711	0.669
AdultMale	0.999	0.869	0.778	0.683	0.594	0.497
AdultFemale	0.877	0.791	0.776	0.766	0.753	0.745

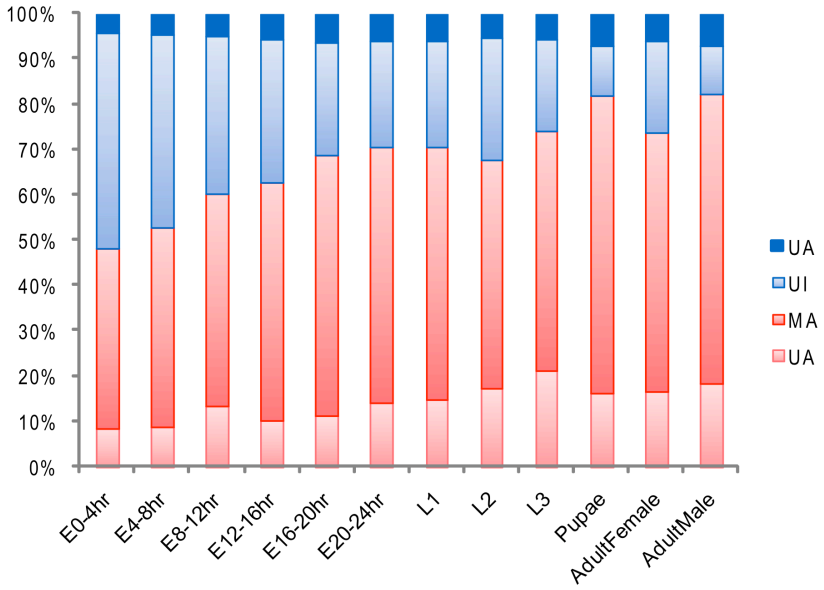
d



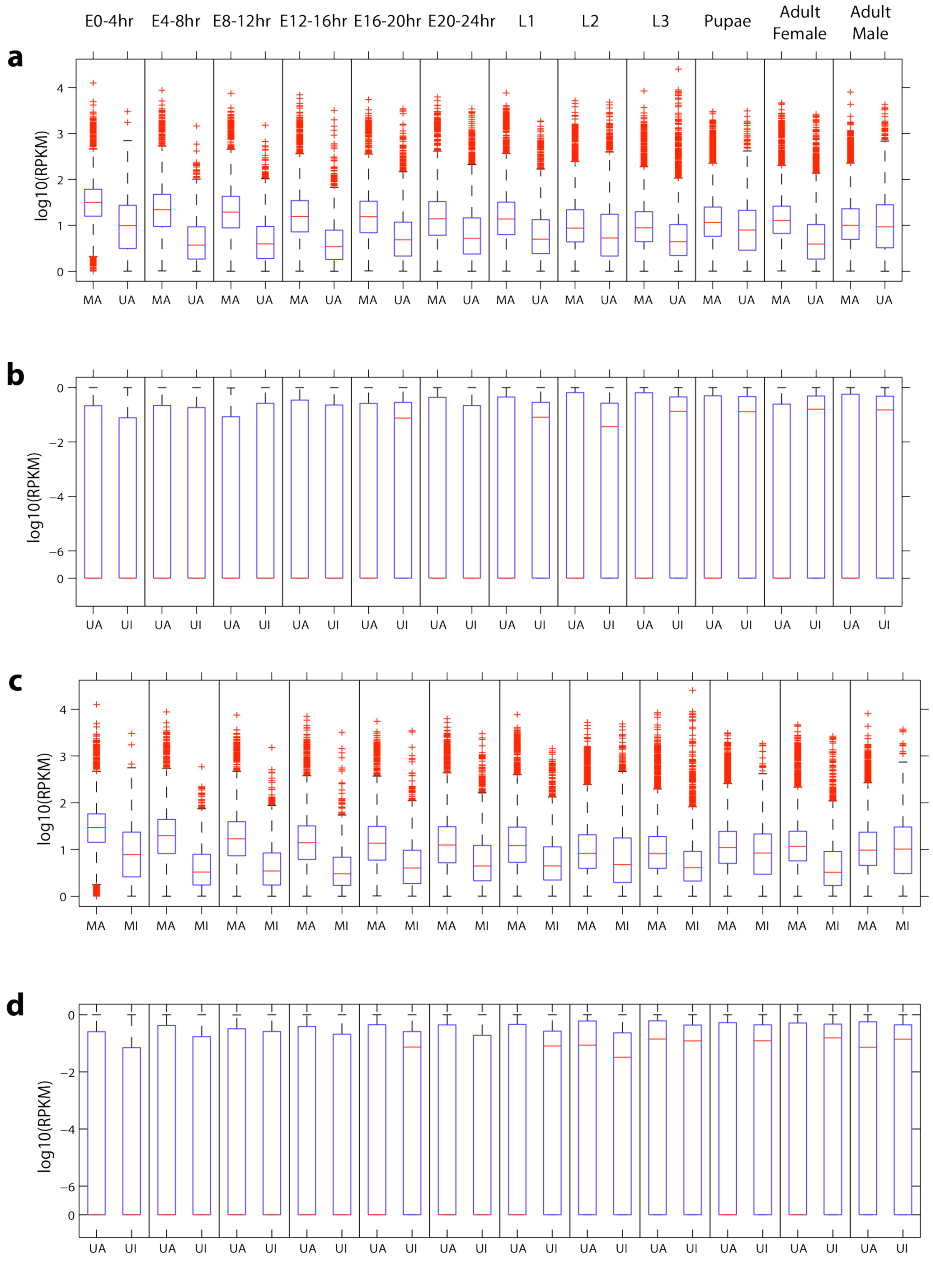
Supplementary Figure 8

A

Developmental stage	AUC	Recall	
		FDR = 0.05	FDR = 0.10
E0-4hr	0.93	0.73	0.83
E4-8hr	0.92	0.76	0.83
E8-12hr	0.88	0.69	0.78
E12-16hr	0.89	0.75	0.84
E16-20hr	0.90	0.71	0.84
E20-24hr	0.87	0.65	0.80
L1	0.86	0.66	0.79
L2	0.85	0.60	0.75
L3	0.81	0.58	0.71
Pupae	0.80	0.65	0.80
AdultFemale	0.78	0.66	0.78
AdultMale	0.84	0.58	0.78

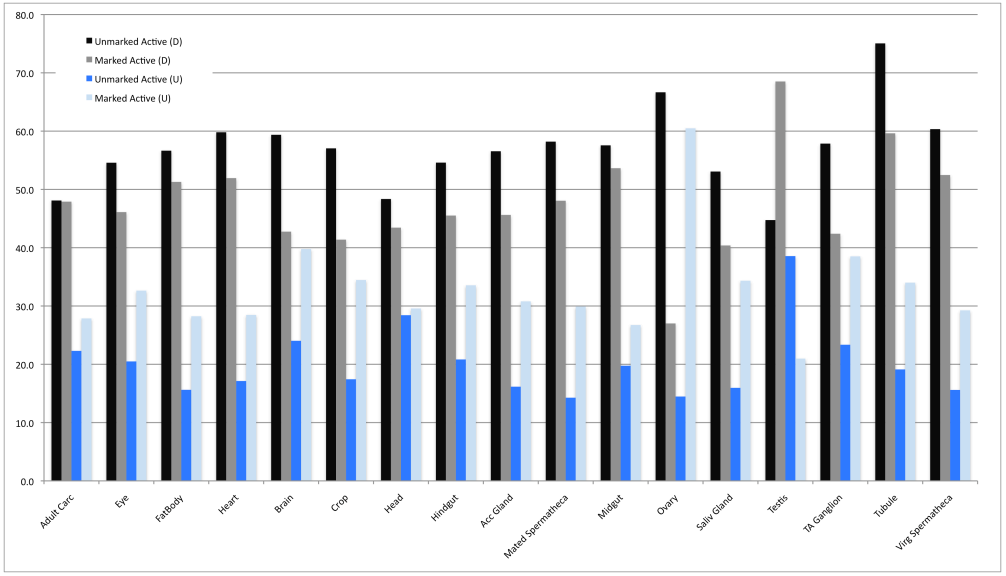


B



Supplementary Figure 9

A



B

Number of developmental stages

a

Predictability	1	2	3	4	5	6	7	8	9	10	11	12
0.25	0.87	0.80	0.79	0.80	0.67	0.66	0.56	0.46	0.44	0.30	0.20	0.04
0.5	0.00	0.00	0.08	0.08	0.08	0.09	0.17	0.18	0.12	0.14	0.16	0.02
0.75	0.00	0.11	0.05	0.03	0.07	0.12	0.09	0.12	0.15	0.21	0.19	0.05
1	0.13	0.10	0.08	0.09	0.18	0.14	0.18	0.24	0.30	0.35	0.44	0.89

b

Predictability	1	2	3	4	5	6	7	8	9	10	11	12
0.25	0.62	0.45	0.37	0.33	0.31	0.30	0.23	0.17	0.20	0.11	0.06	0.01
0.5	0.00	0.00	0.31	0.32	0.18	0.16	0.23	0.17	0.11	0.09	0.11	0.02
0.75	0.00	0.31	0.15	0.15	0.15	0.24	0.20	0.19	0.15	0.18	0.17	0.03
1	0.38	0.24	0.17	0.20	0.36	0.30	0.34	0.48	0.53	0.62	0.67	0.94

c

Predictability	1	2	3	4	5	6	7	8	9	10	11	12
0.25	0.29	0.10	0.10	0.05	0.06	0.04	0.05	0.02	0.04	0.04	0.01	0.00
0.5	0.00	0.00	0.05	0.02	0.02	0.02	0.03	0.02	0.01	0.02	0.01	0.00
0.75	0.00	0.08	0.08	0.03	0.03	0.04	0.06	0.01	0.03	0.03	0.02	0.01
1	0.71	0.81	0.77	0.90	0.89	0.91	0.86	0.95	0.92	0.91	0.95	0.98

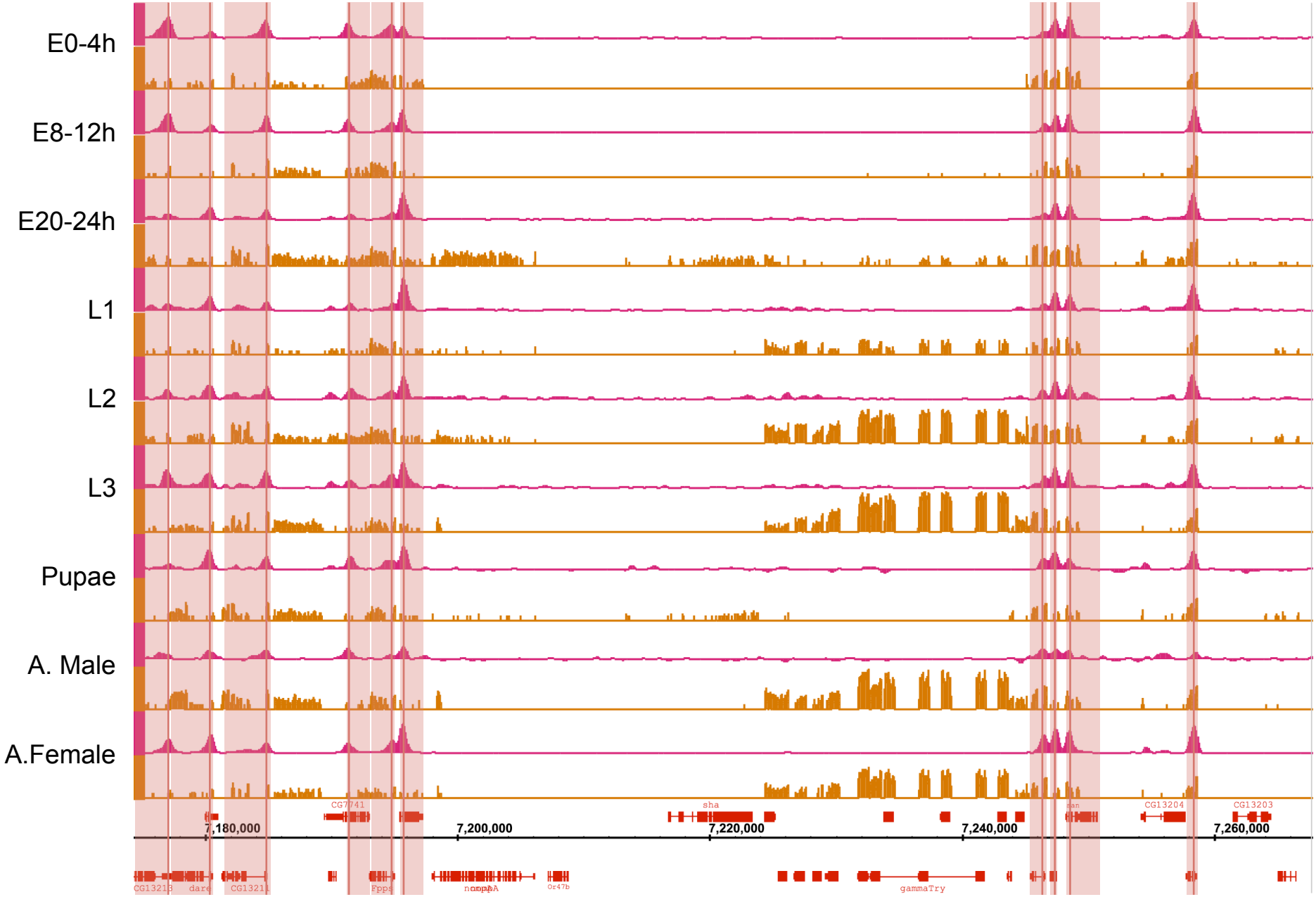
d

Predictability	1	2	3	4	5	6	7	8	9	10	11	12
0.25	0.45	0.20	0.17	0.09	0.11	0.06	0.08	0.04	0.07	0.07	0.03	0.02
0.5	0.00	0.00	0.08	0.05	0.05	0.03	0.08	0.03	0.04	0.03	0.05	0.02
0.75	0.00	0.20	0.20	0.08	0.13	0.19	0.22	0.07	0.08	0.13	0.13	0.10
1	0.55	0.60	0.56	0.78	0.72	0.71	0.61	0.86	0.81	0.77	0.79	0.87

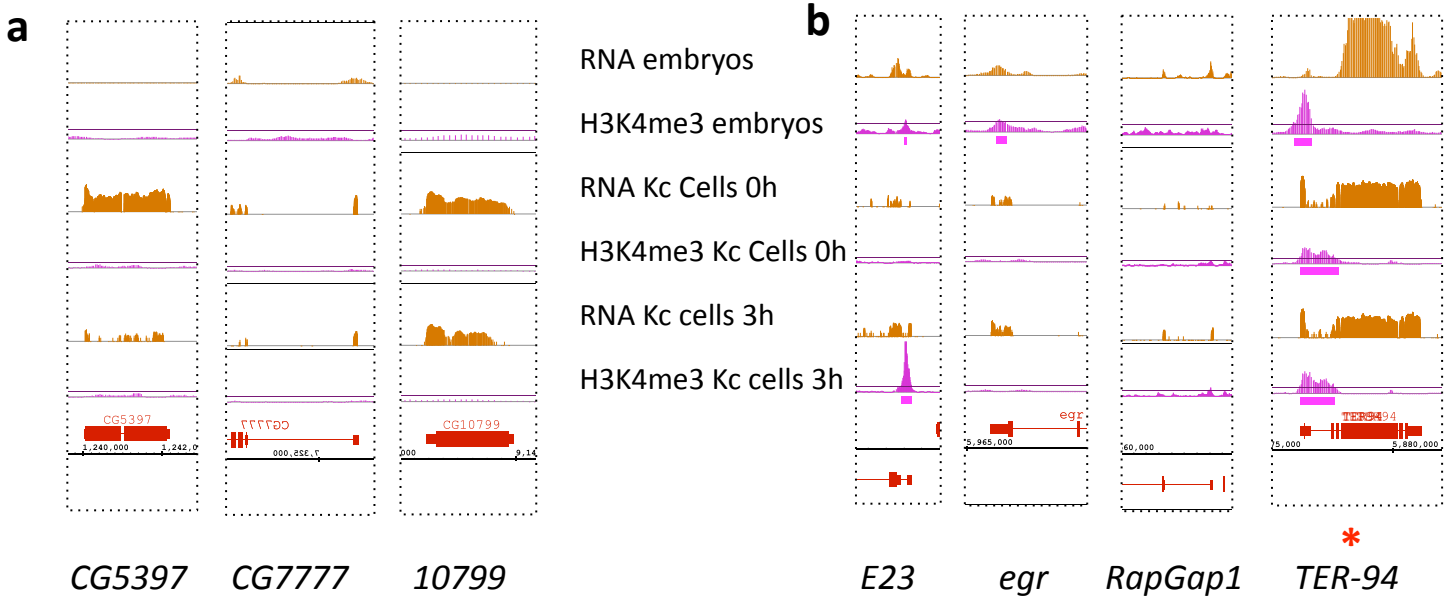
Fraction of transcripts

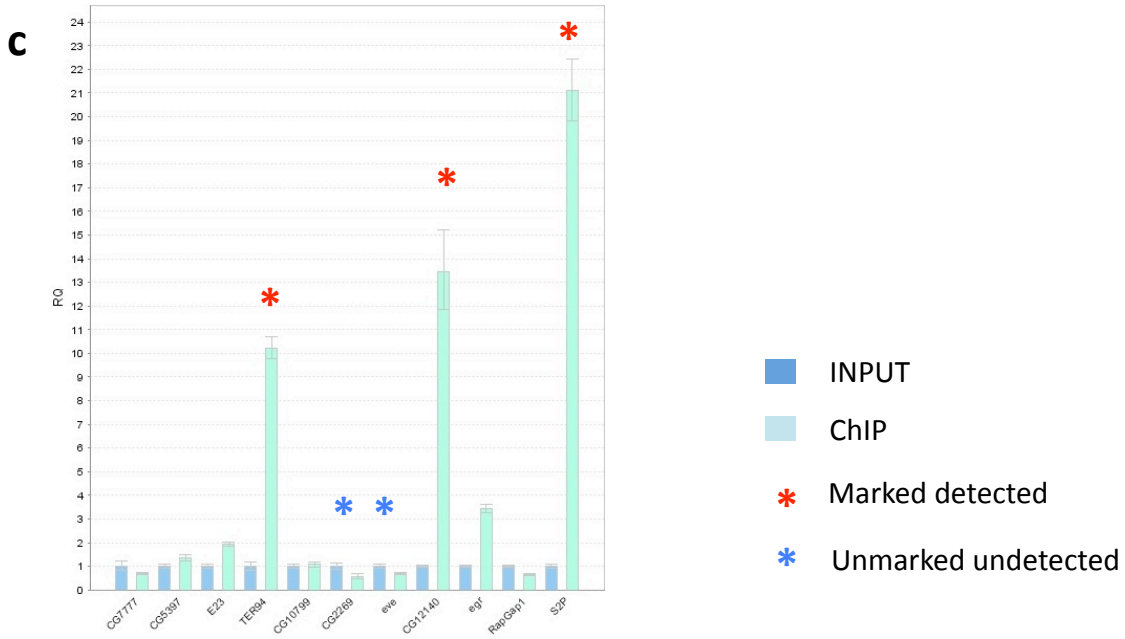
0	0.2	0.4	0.6	0.8	1

Supplementary Figure 10



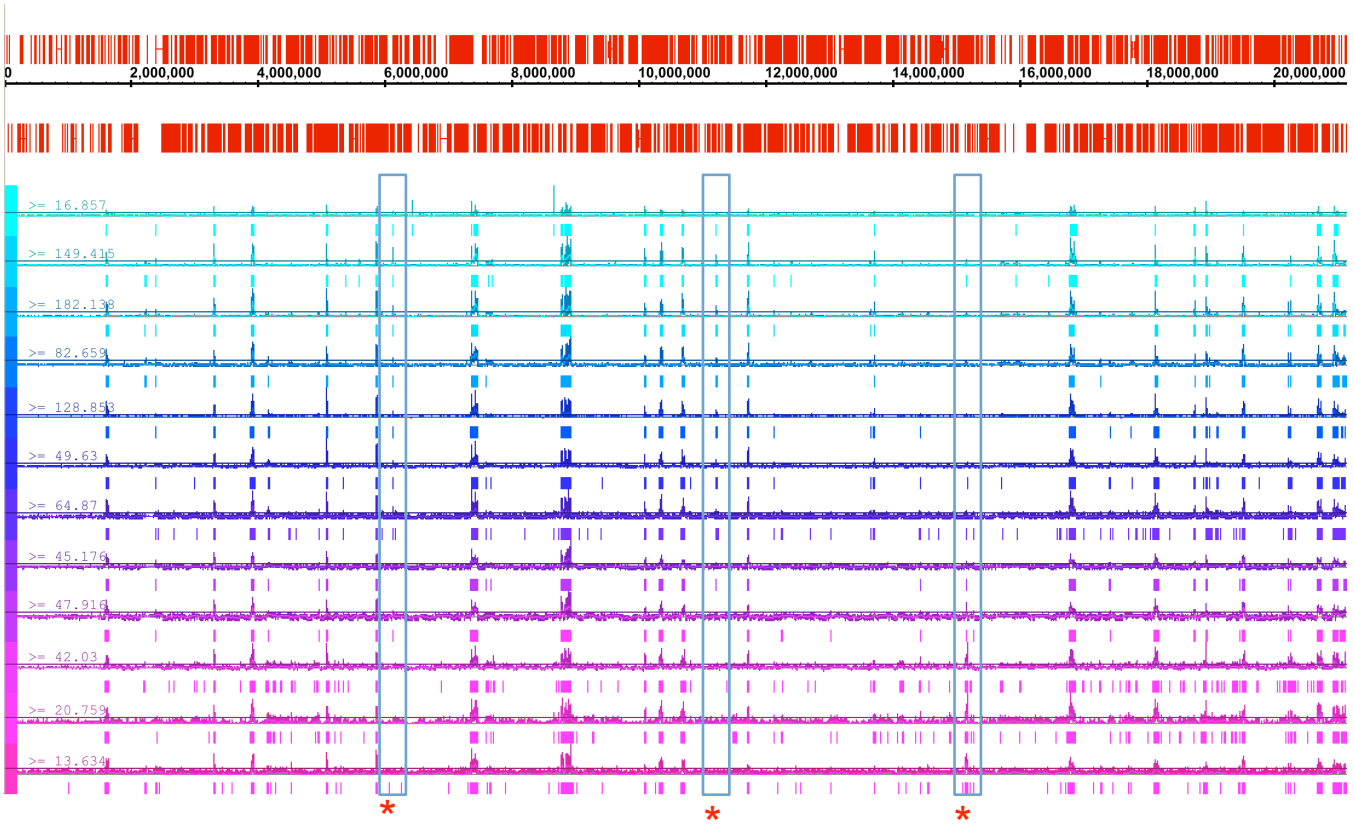
Supplementary Figure 11



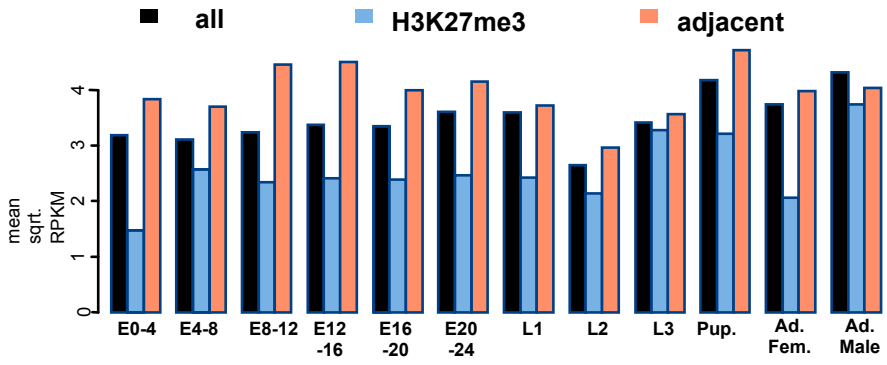


Supplementary Figure 12

a



b



Supplementary Figure 13

A

Cluster	Covered by H3K27me3 in	signif. GO cat.
85	embryo stages	transcription
89	All stages	transcription, development, segmentation
91	early embryo to pupae	endopeptidase inhibitor activity
100	late embryo to pupae	apoptosis

B **Cluster 38: Pupae-specific repression:**

Category	Term	Count	%	P-Value	FDR (BH)
GOTERM_BP_FAT	polysaccharide metabolic process	5	5.7	1.20E-02	8.40E-01
GOTERM_BP_FAT	chitin metabolic process	5	5.7	4.50E-03	8.60E-01
GOTERM_BP_FAT	aminoglycan metabolic process	5	5.7	9.70E-03	8.80E-01
GOTERM_BP_FAT	regulation of transcription	10	11.5	4.20E-02	9.90E-01
GOTERM_BP_FAT	proteolysis	9	10.3	6.30E-02	9.90E-01
GOTERM_CC_FAT	vacuolar proton-transporting V-type ATPase, V0 domain	2	2.3	9.60E-02	1.00E+00
GOTERM_MF_FAT	polysaccharide binding	6	6.9	8.70E-04	1.10E-01
GOTERM_MF_FAT	pattern binding	6	6.9	8.70E-04	1.10E-01
GOTERM_MF_FAT	carbohydrate binding	6	6.9	5.70E-03	1.50E-01
GOTERM_MF_FAT	alkaline phosphatase activity	3	3.4	3.60E-03	1.50E-01
GOTERM_MF_FAT	transcription factor activity	8	9.2	8.40E-03	1.50E-01

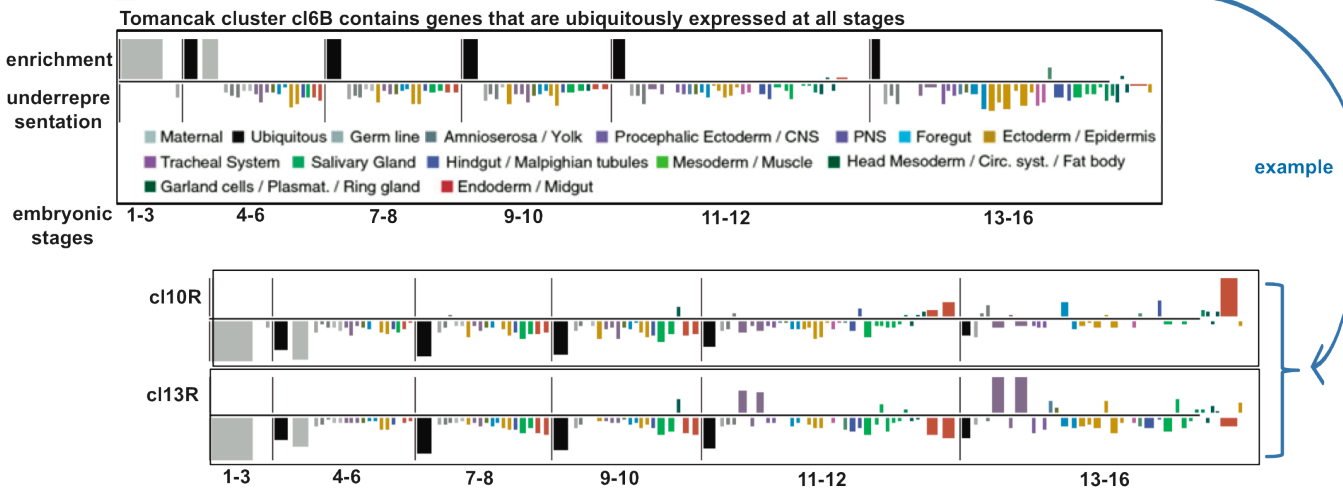
Supplementary Figure 14

A

Tomancak clusters underrepresented in Cluster 89			
Tomancak-cluster	% of Tomancak cluster genes in Tomancak-DB genes	% of Tomancak-cluster genes in Cluster 89	p-val (Fisher)
cl1B	4.6	0.79	0.030106
cl2B	6.12	0.79	0.004463
cl5B	6.47	1.57	0.016372
cl6B	12.43	1.57	1.41E-05
cl8R	3.45	0	0.022502
cl10B	18.51	2.36	3.65E-08

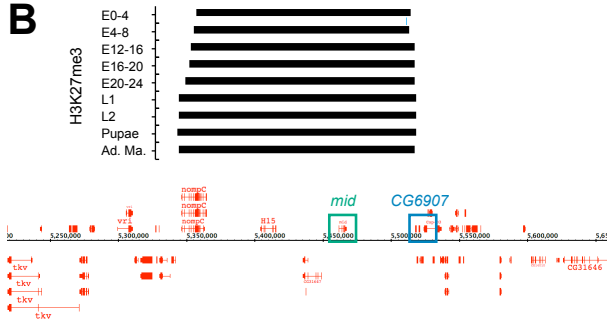
Tomancak cluster overrepresented in Cluster 89			
Tomancak-cluster	% of Tomancak cluster genes in Tomancak-DB genes	% of Tomancak-cluster genes in Cluster 89	p-val (Fisher)
cl7R	4	11.02	0.000478
cl10R	5.54	10.24	0.028077
cl12R	1.42	3.94	0.033583
cl13R	4.11	7.87	0.039649
cl15R	3.4	10.24	0.000327
cl23R	0.96	3.15	0.032052
cl24R	1.09	5.51	0.000389
cl25R	2.27	12.6	1.55E-08
cl26R	2.76	14.17	6.51E-09
cl27R	1.4	4.72	0.008215

example

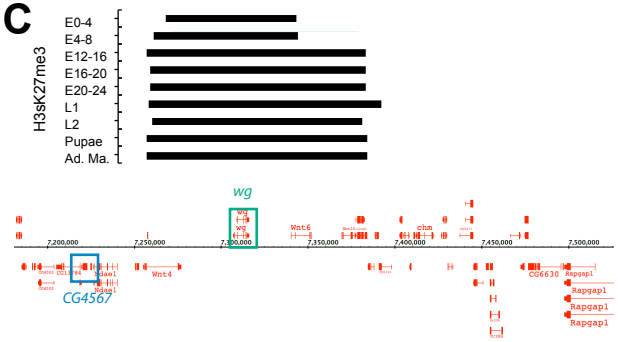


example

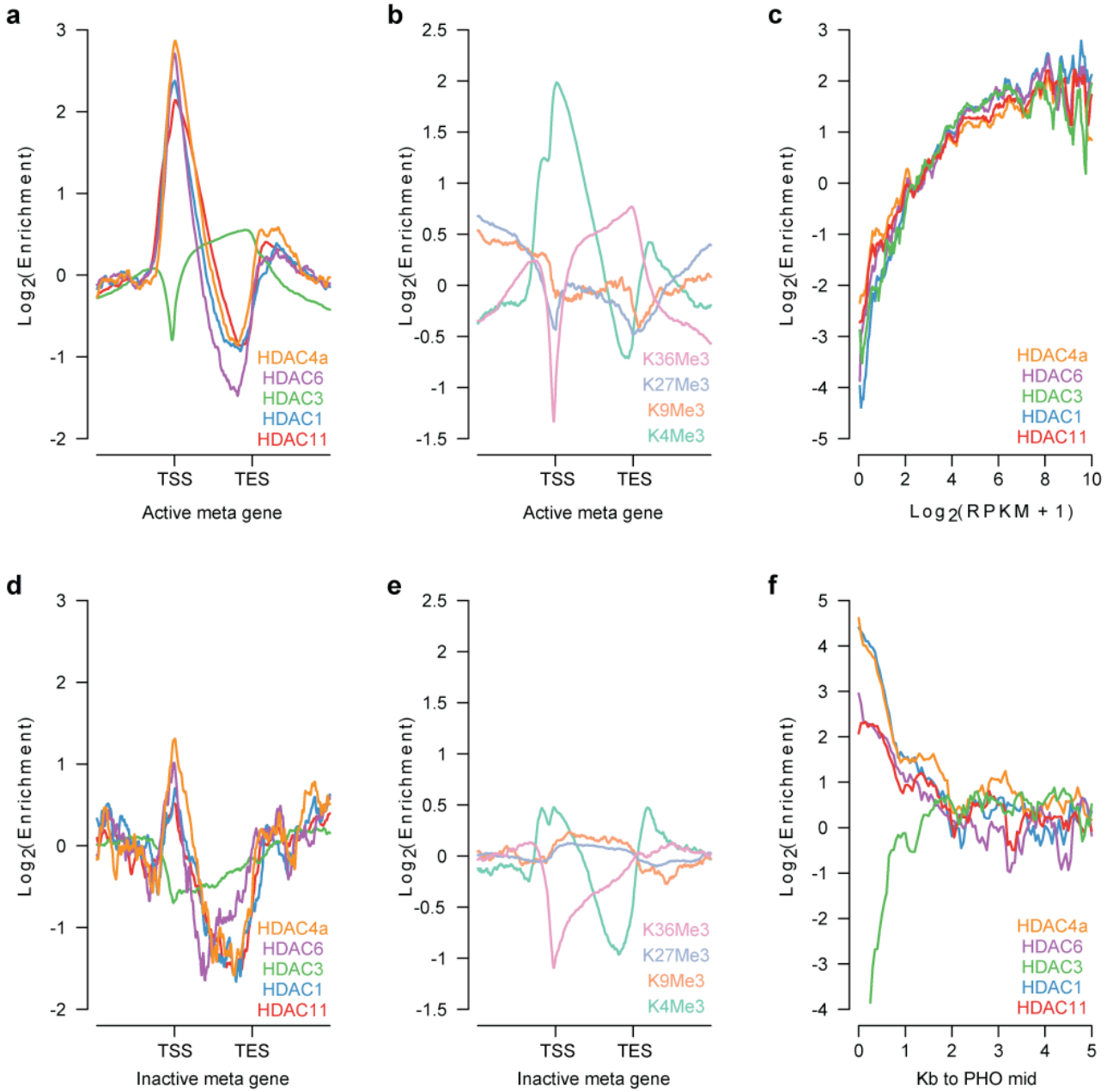
B



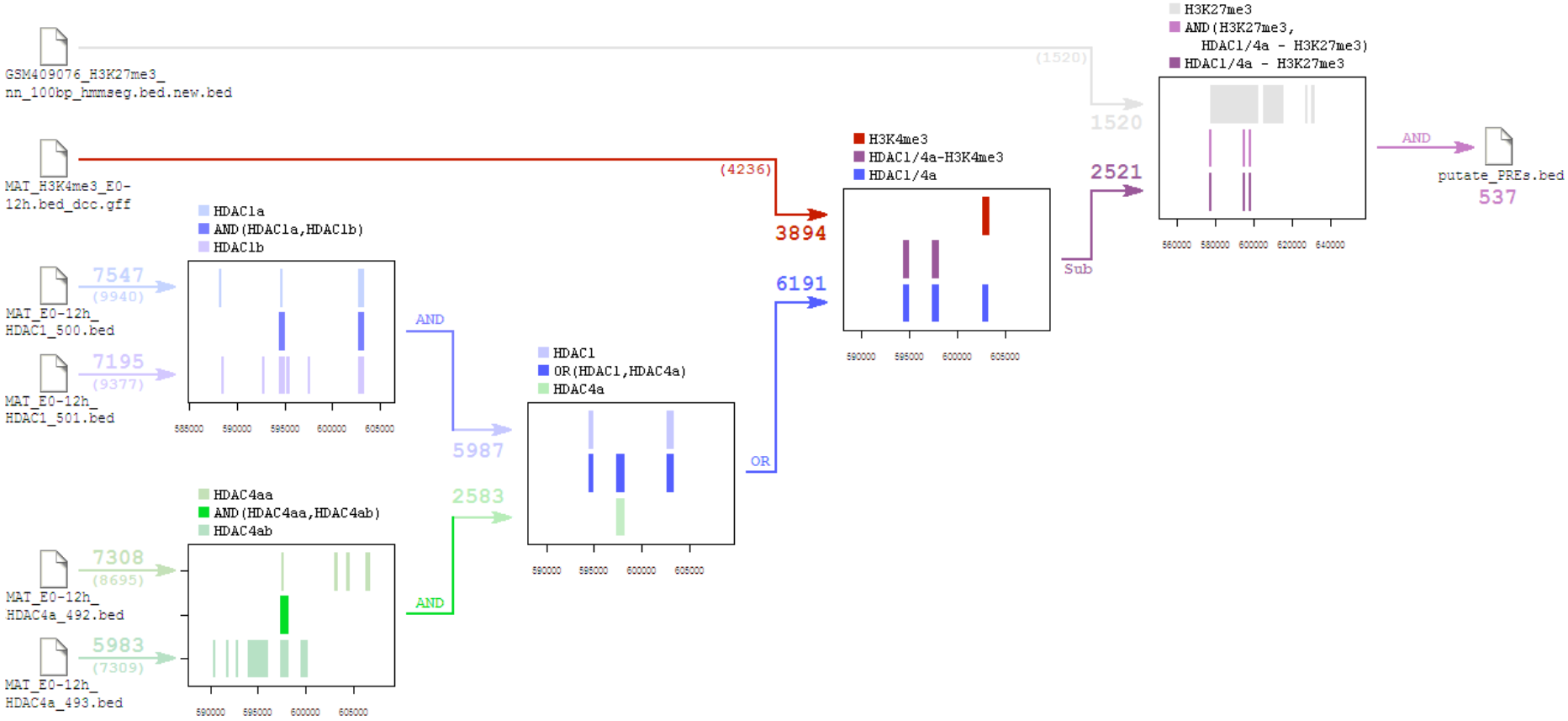
C



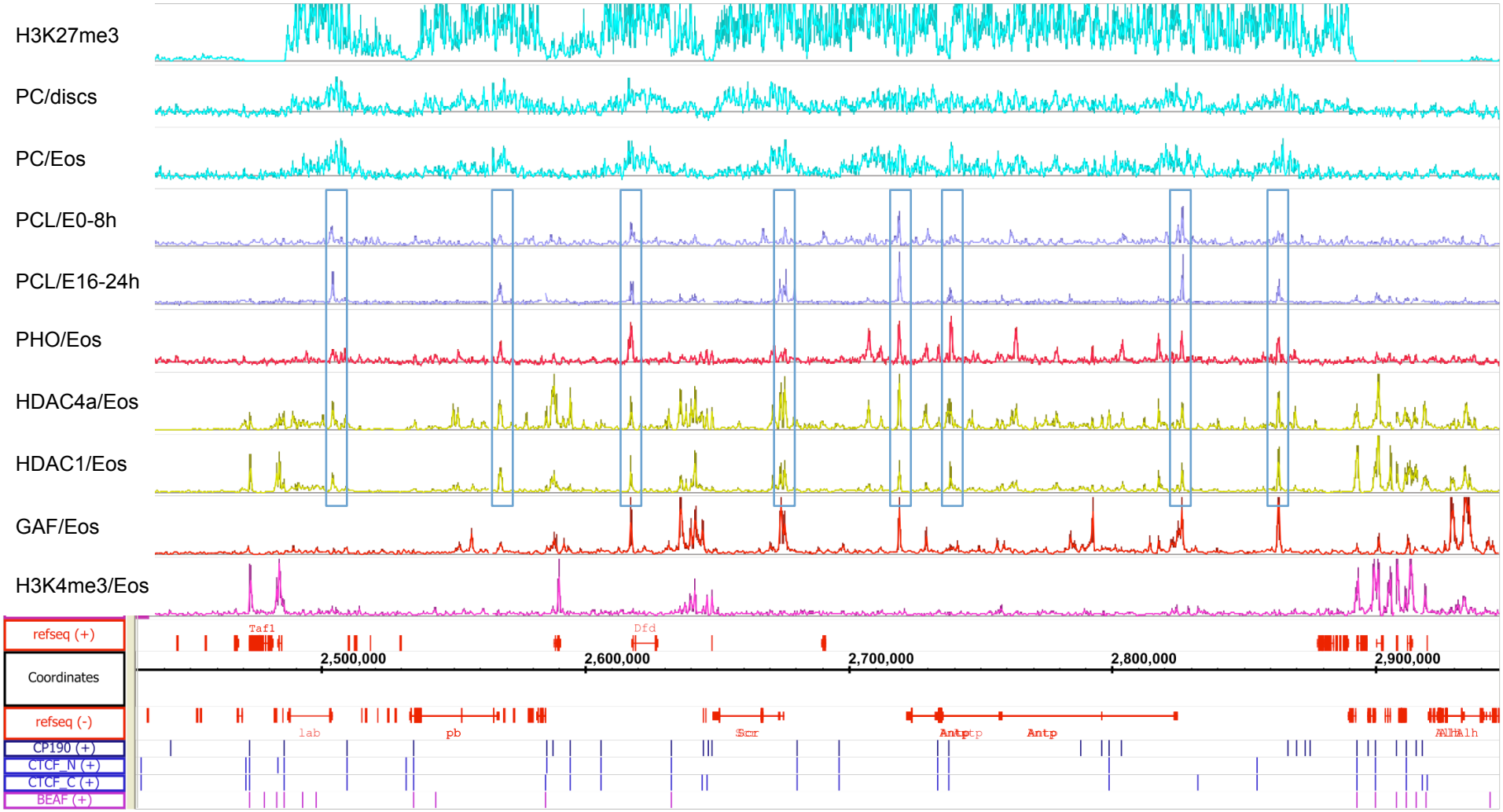
Supplementary Figure 15



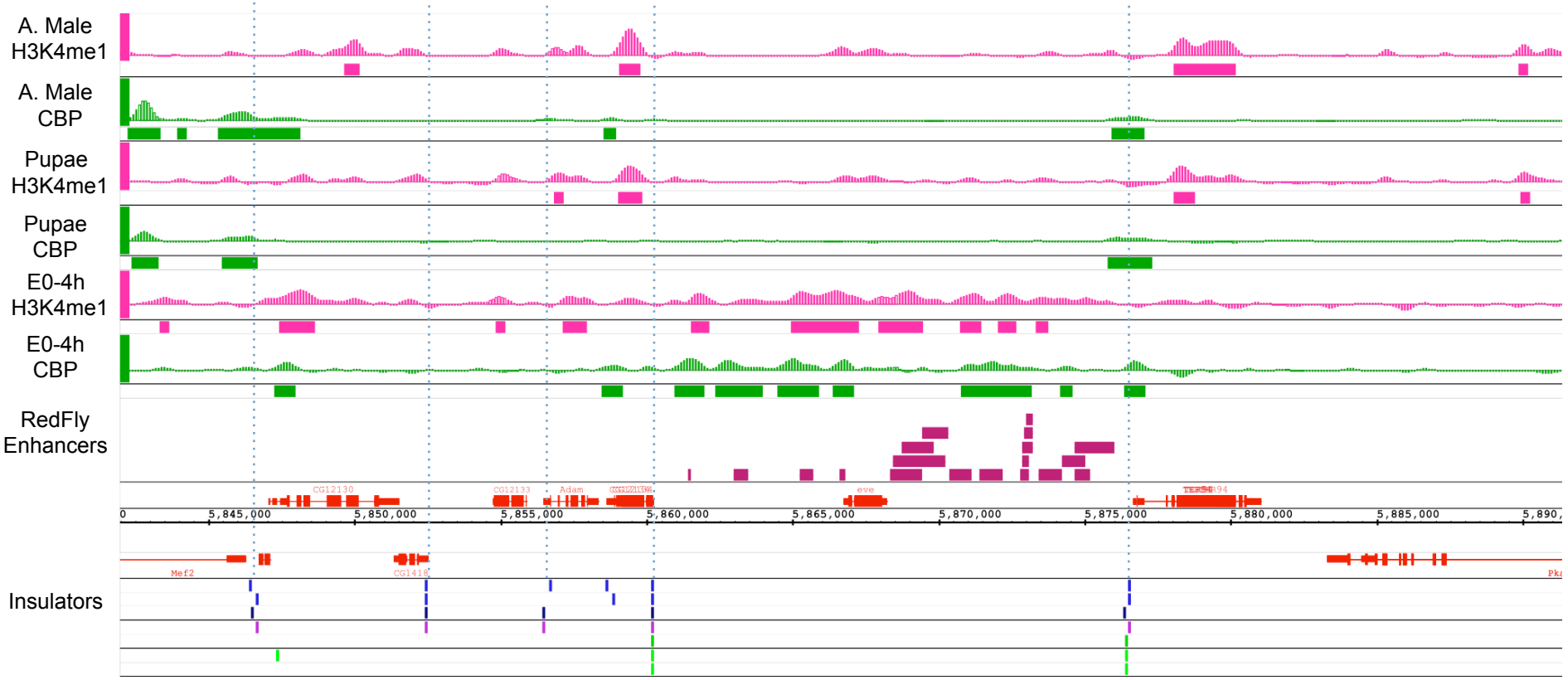
Supplementary Figure 16



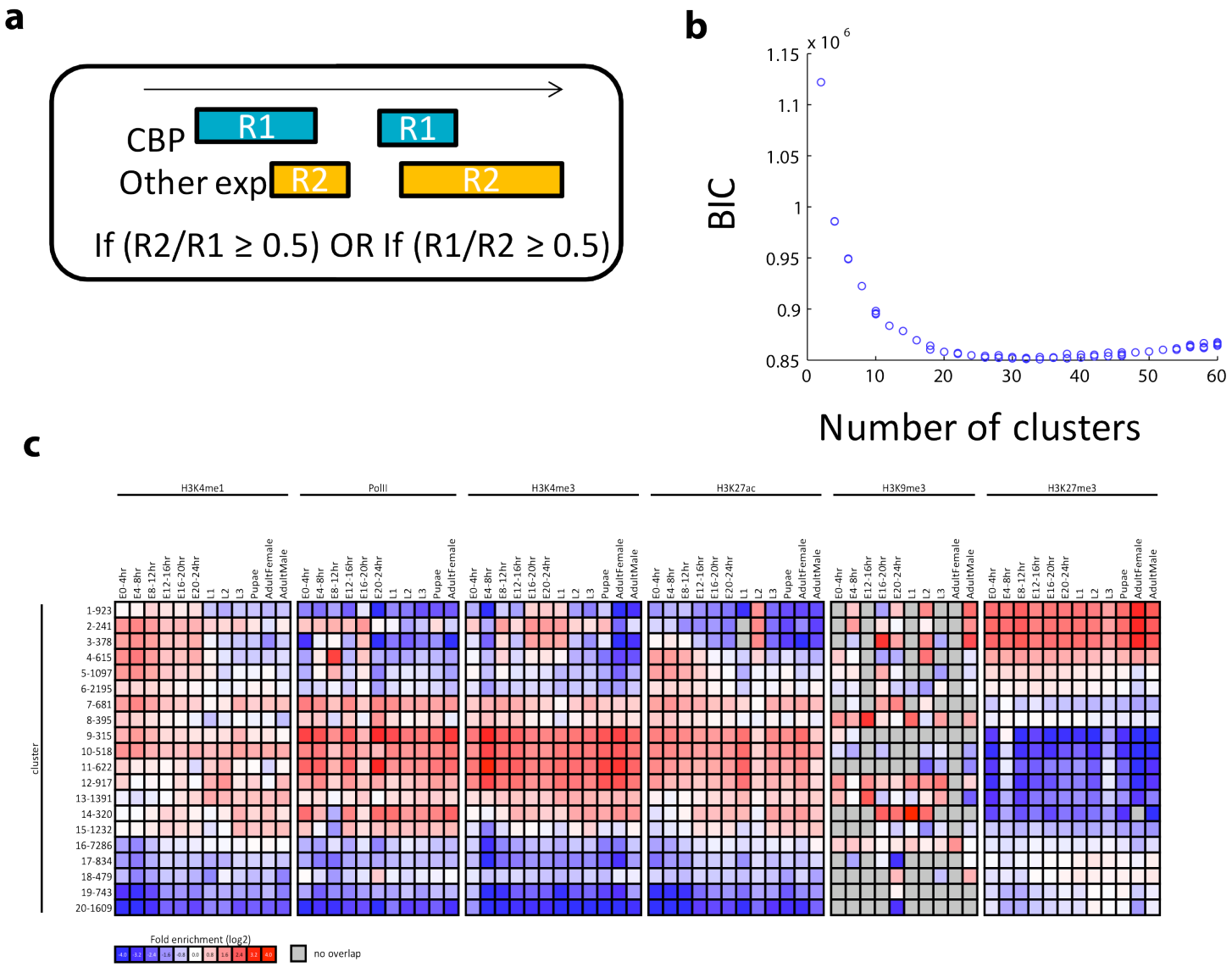
Supplementary Fig. 17



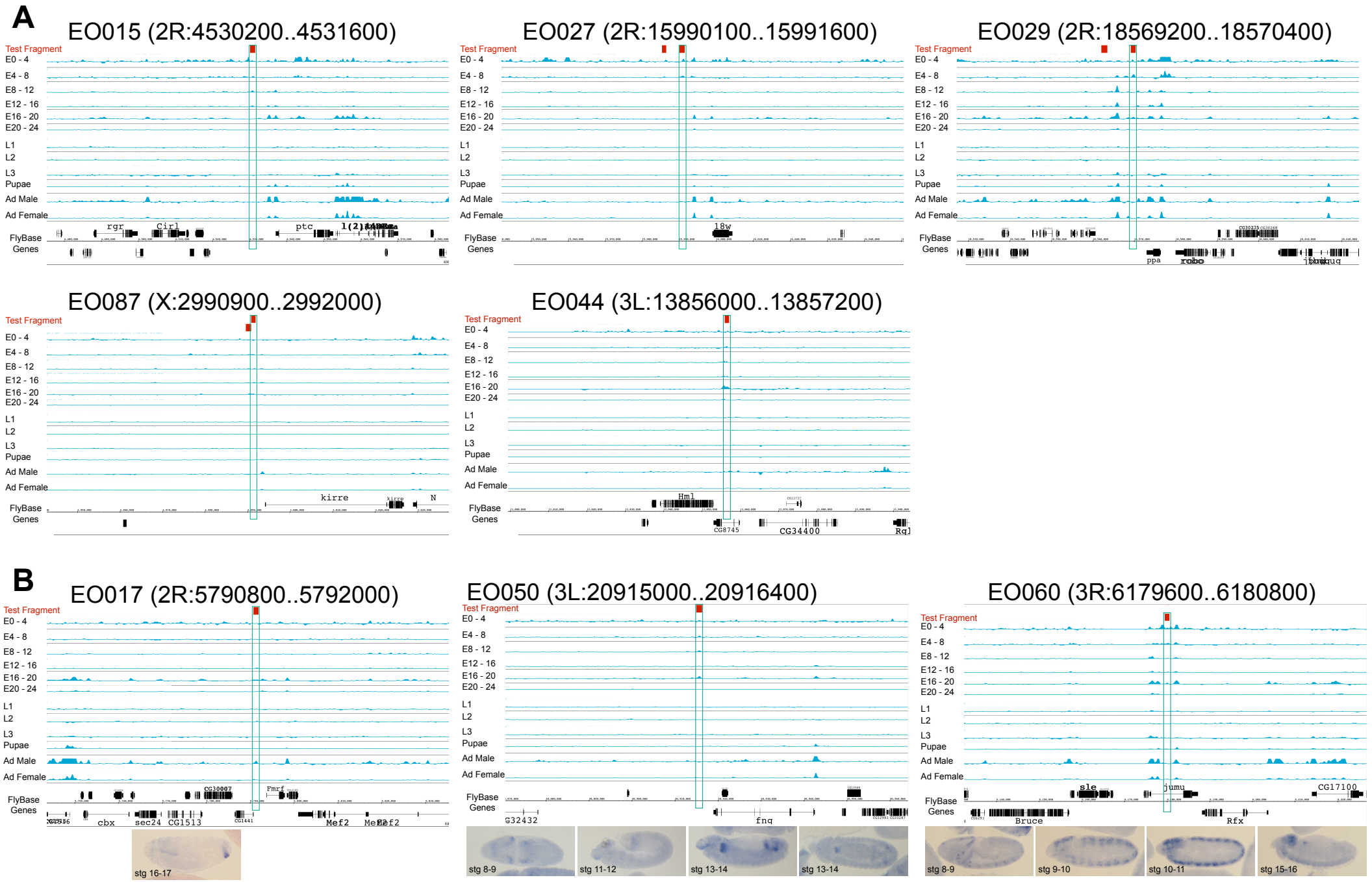
Supplementary Fig. 18



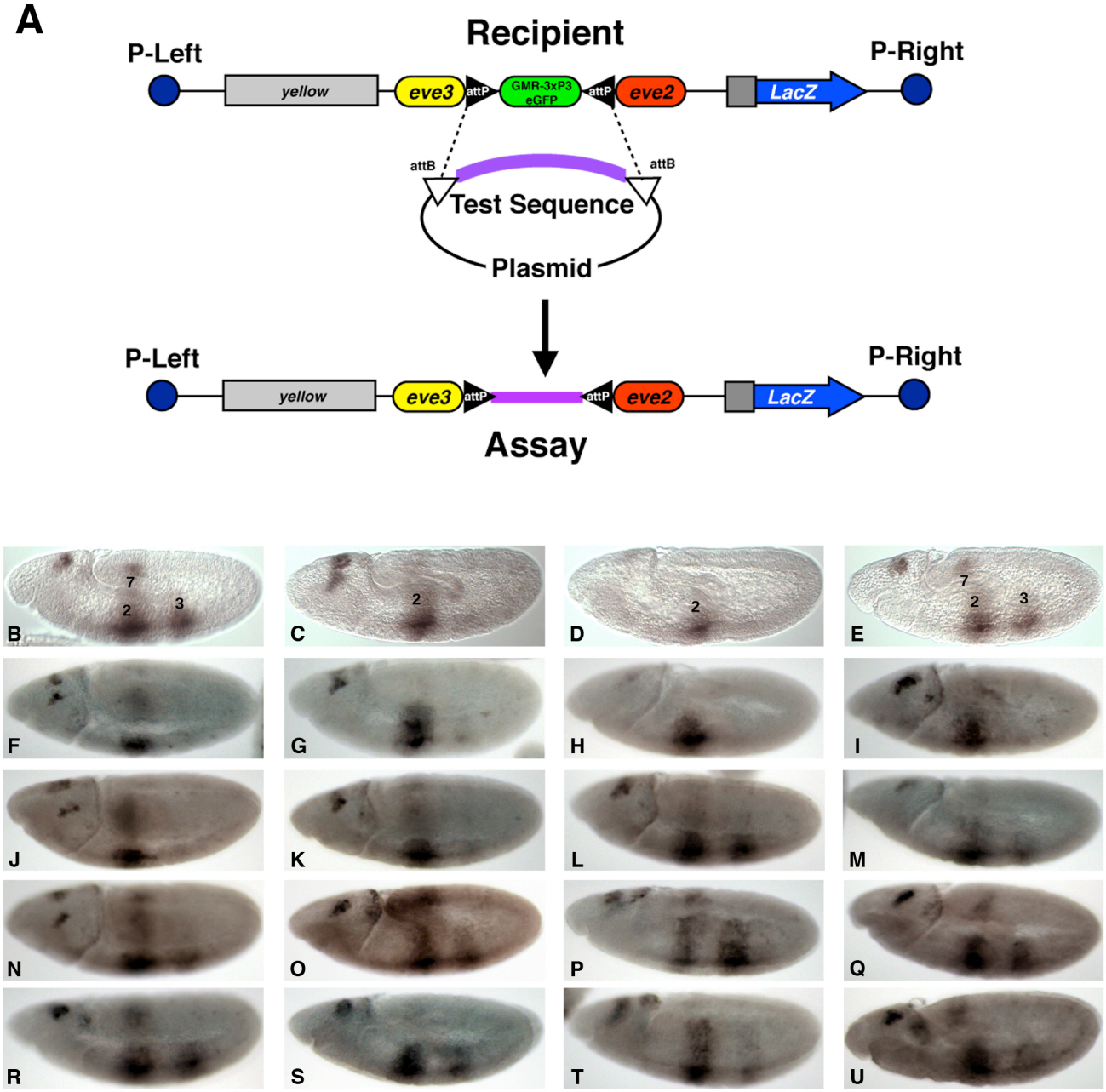
Supplementary Fig. 19



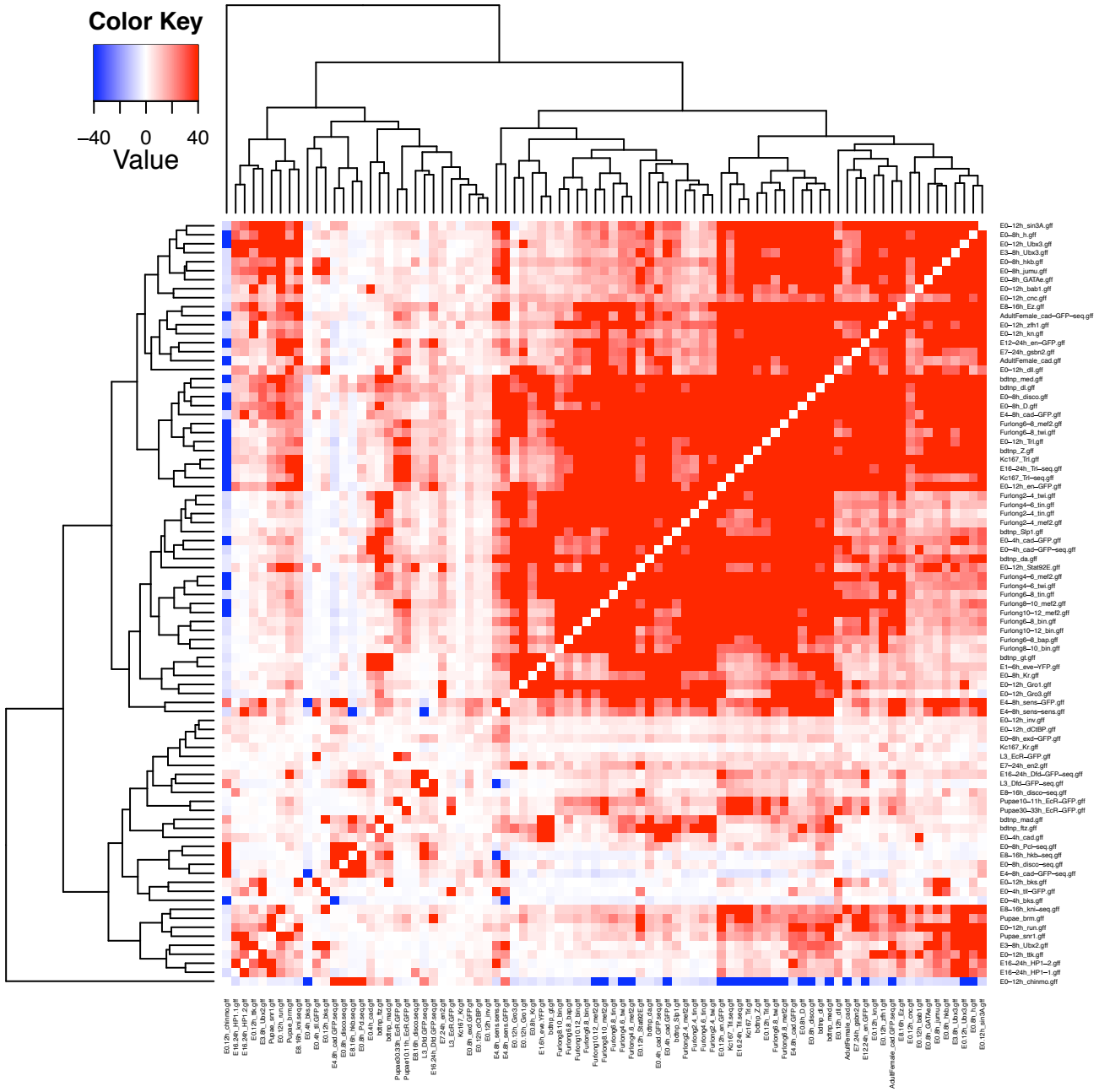
Supplementary Fig. 20



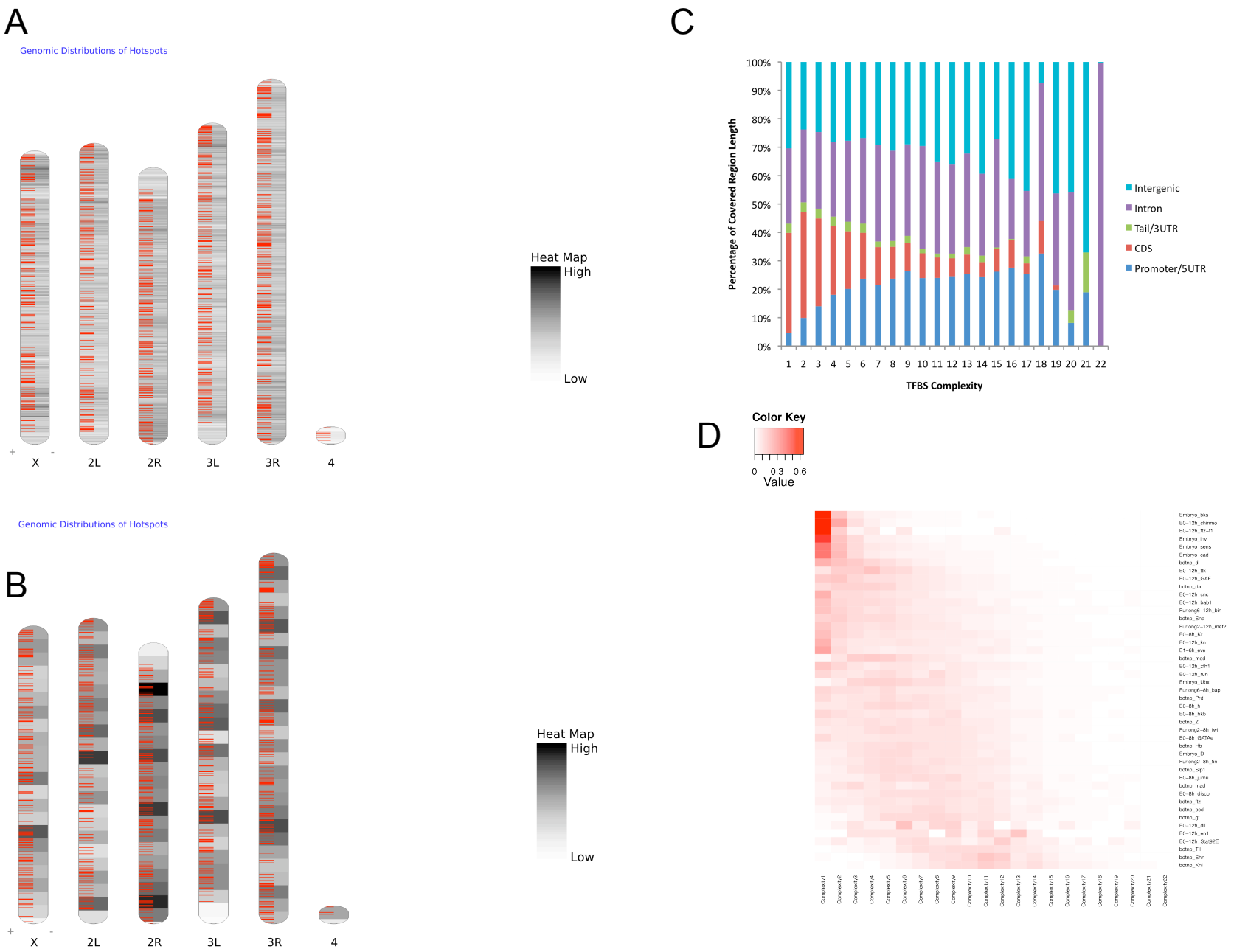
Supplementary Figure 21



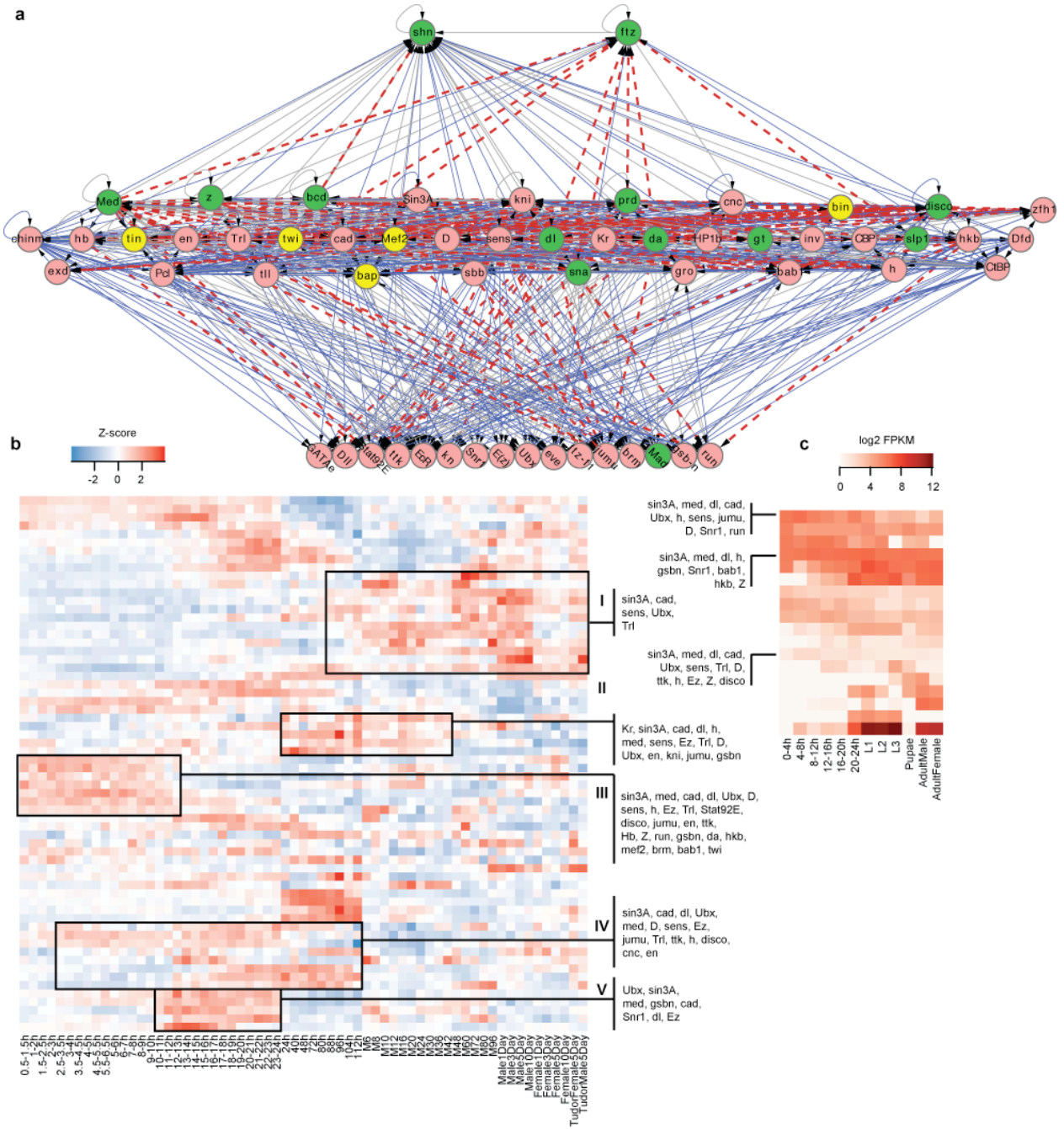
Supplementary Fig. 22



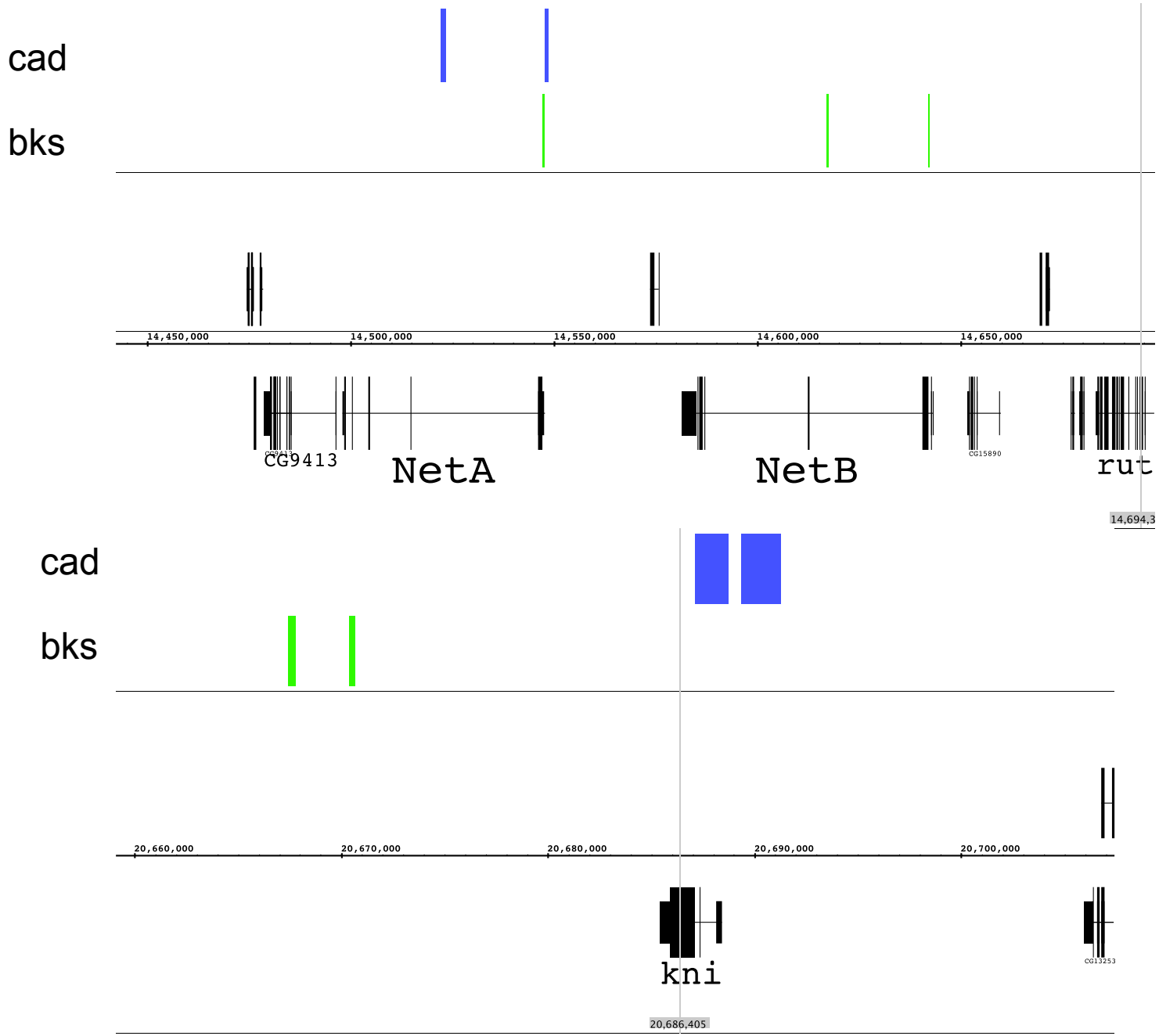
Supplementary Fig. 23



Supplementary Fig. 24

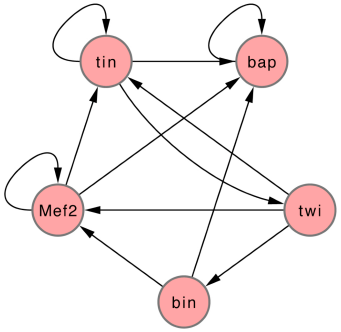


Supplementary Fig. 25

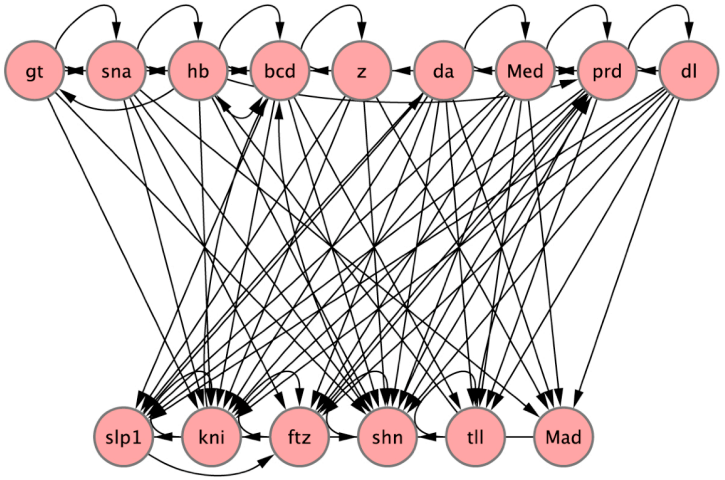


Supplementary Fig. 26

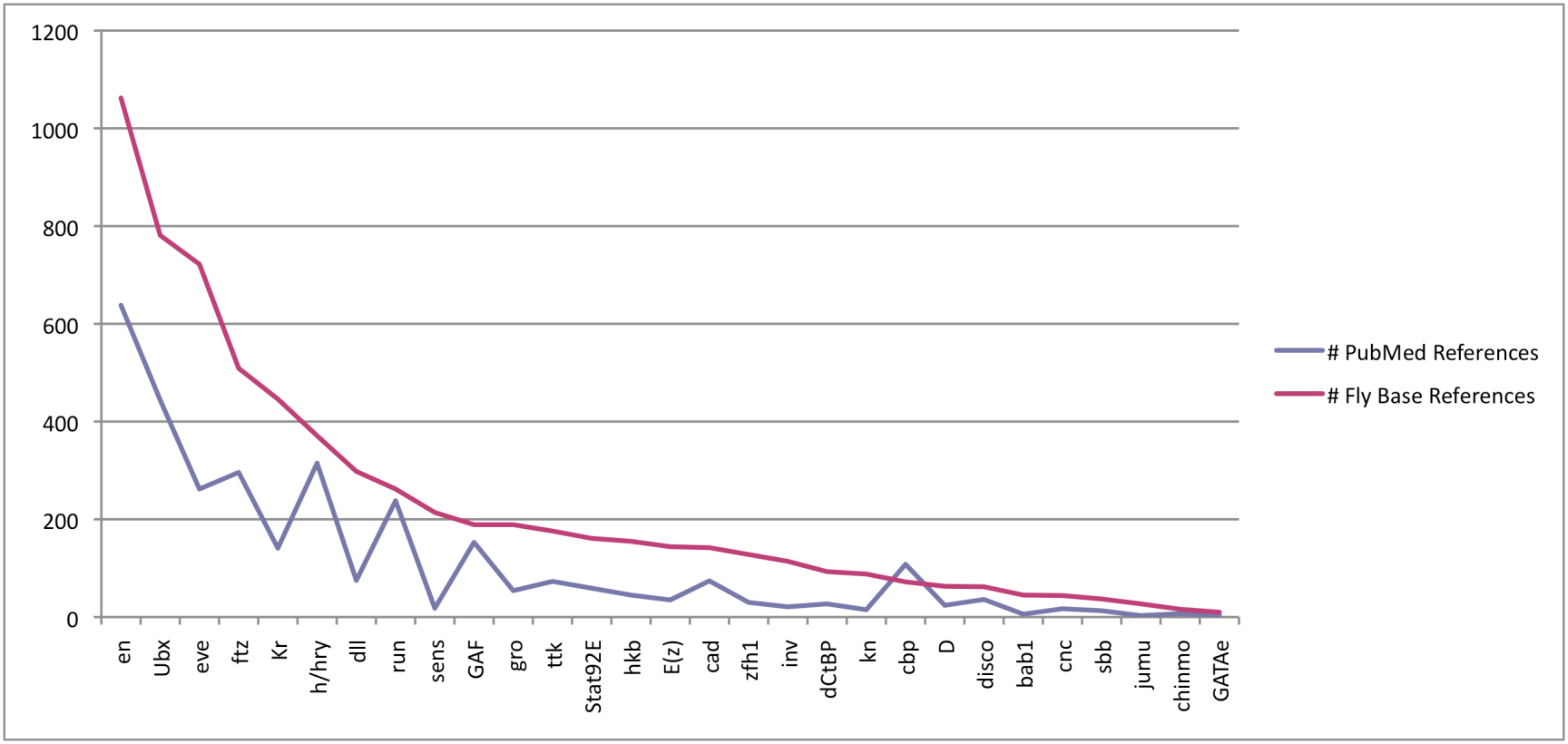
A



B



Supplementary Fig. 27



Supplementary Fig. 28