SUPPLEMENTAL METHODS (ANALYSIS)

Peak calling

During the course of the project sequencing methods evolved, the SPP peak caller went through multiple rounds of improvement and our internal pipeline changed (Kudron et al. 2018). To create a uniformly called set of peaks for the analyses presented here we processed all the ChIP-seq experiments through the ENCODE Transcription Factor and Histone ChIP-Seq processing pipeline version v1.3.6 , https://github.com/ENCODE-DCC/chip-seq-pipeline2/tree/master. (This version of the processing pipeline was used for all data submitted to the ENCODE DCC as well as data generated after October 2022.) In all cases the bwa program was used for alignment (Heng Li and Durbin 2009; Heng Li 2013). The version of bwa in that pipeline is 0.7.17 https://bio-bwa.sourceforge.net/. For very early experiments that generated only short single ended sequence reads, the alignments done with the earlier version of bwa. 0.7.8 were used. The peak caller used for all experiments is SPP version 1.15.5 (https://hbctraining.github.io/Intro-to-ChIPseq/lessons/peak_calling_spp.html) (Kharchenko, Tolstorukov, and Park 2008). Because the ENCODE DCC ceased to accept data in October, 2022, these new peaks calls are not available there but are available only through the website https://epic.gs.washington.edu/modERNresource and also through the SRA.

Clustering of transcription factor peaks into metapeaks

Exploiting the similarity between TF peaks and sequence reads aligned to the genome, we aggregated TF peaks across all the ChIP-seq experiments for each species and called metapeaks. These called metapeaks represent the bounds of the clusters of TF peaks. The TF peaks are then assigned to metapeaks based on proximity and overlap with the metapeaks.

We first aggregated all the peaks from the TF experiments for each species into a single bed file. SPP sometimes called several peaks from a single experiment that overlapped exactly with the exact same genomic span but different apices. These exactly overlapping peaks were trimmed, so that they were contiguous and non-overlapping. The endpoints of these trimmed peaks were the midpoint between apices. The apices were unchanged.

We next formed a signal track of all the aggregated trimmed TF peaks as a bedGraph file, as described at http://genome.ucsc.edu/goldenPath/help/bedgraph.html, using the program bedtools genomecov -bg (Version: v2.29.0) https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html.

The MACS2 program (version 2.2.4) https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html was selected to call the metapeaks from the aggregated TF peaks. Unlike SPP, the MACS2 program can be run without a control signal track, and for the TF peaks there is no control signal track.

The bdgpeakcall subcommand of MACS2 was used to call the metapeaks from bedGraph files. Inspection of the results showed that clusters of peaks could sometimes overlap on their edges and depending on the extent of the overlap the valley between the two putative metapeaks could be of differing depths. To split these overlapping metapeaks, we iterated the peak calling procedure using a minimum peak length of 50 bases with cutoff thresholds from 2 to 50. This created 49 output files in narrowPeak bed format https://genome.ucsc.edu/FAQ/FAQformat.html#format12. As the cutoff increases, metapeaks with an internal valley will split into 2 metapeaks when the cutoff exceeds the minimum of the valleys. Using a range of cutoff values makes it possible to split these broad metapeaks with internal valleys into separate metapeaks.

To come up with a final set of metapeaks, the multiple cutoff metapeaks were arranged into a tree structure. The root of the trees was the metapeak called with cutoff 2. The children metapeaks were the metapeaks from next higher cutoff that overlap the lower cutoff metapeak. If the higher cutoff did not go above an internal valley, there was only one child, but if the higher cutoff goes above a single internal valley, there were two children. The tree was extended to the maximum threshold. Branches extending only one generation were trimmed to avoid over splitting. For the trimmed trees, each leaf metapeak was traced up the tree to when it was first created by a split of a metapeak, that is, it has a sister, and it is added to the list of metapeaks. After all the leaf metapeaks were traced up to their origins, the genomic width of the selected metapeaks is expanded to fill the entire width of the metapeaks at root level of cutoff 2. This process is a compromise between splitting broad metapeaks with deep internal valleys and not splitting those metapeaks with very shallow internal valleys (depth of just one). The resulting metapeaks are also non-overlapping.

Each TF peak was then assigned to a metapeak as follows: If the TF peak apex was within the range of a metapeak, it was assigned to that metapeak. If the TF peak apex failed to overlap the range of a metapeak, but the peak region overlapped a single metapeak, it was assigned to that metapeak. If the TF peak region overlapped more than one metapeak range, it was assigned to the metapeak whose end was closest to the apex of the TF peak. If the TF peak did not overlap any metapeaks, then a new singleton metapeak was created for that TF peak.

Conservation scores

The *D. melanogaster* (dm6) and *D. virilis* (droVir3) assemblies were aligned using blastz (Chiaromonte, Yap, and Miller 2002; S. Schwartz et al. 2003) using blastz parameters and scoring matrices recommended on the UCSC browser (Kent et al. 2002). Similarly, the *C. elegans* (ce11) and *C. remane*i (caeRem4) genomes were aligned. The alignments were chained using axtChain (Kent et al. 2003) and processed into nets by the chainNet and netSyntenic tools (Kent et al. 2003). Conservation scores were calculated by giving matching bases a score of 2, gaps a score of 0, and substitutions the following scores A/C=T/G=A/T=C/G=-2, A/G=T/C=-1. Scores were summed across the bases in the feature of interest.

HOT site determination

We calculated a HOTness score using a kernel density estimation (KDE) approach (L. Ma and A. Victorsen, unpublished) looking for regions of maximum point pattern densities as used, for example, in defining areas of geographical regions of high crime rates. The following parameters were used: bandwidth 300, cutoff score 0.1, cutoff peak 0.00001, and local peak height 30. Thresholds were defined by stage and by chromosome and across chromosomes within each stage using 1000 iterations randomly sampling from all observed peaks. This approach labels the largest number of peaks as HOT sites (**Supplemental Figure 3A**).

In *D. melanogaster* almost all samples were embryonic, but in *C. elegans* we experimented with setting HOT site thresholds per stage as suggested by Araya et al., (Araya et al. 2014) identifying the 1% and 5% raw threshold per stage as compared to identifying the thresholds for the data set as a whole. We determined most sites were found in all of the developmental stages. Approximately 10% additional sites were identified as “HOT” when using the cutoffs set per stage rather than setting the cutoff across all stages with 92% of those being within a few percent of the cutoff. Thus, we chose to define the thresholds across the full data sets in both *C. elegans* and *D. melanogaster*, defining the 5% most highly occupied metapeaks as

HOT sites and the 1% most highly occupied metapeaks as ultra-HOT sites (**Supplemental Figure 3B, C**).

Overlap of metapeaks with chromatin features and ATAC peaks

D. melanogaster – ATAC-seq: Calderon et al., (Calderon et al. 2022) obtained 110,185 ATAC-seq regions from staged *D. melanogaster* embryos (11 overlapping time windows spanning the first 20 hours of development). As measured by an overlap of at least ten bases with ChIP metapeaks, by count 80% of ATAC regions are overlapped by a metapeak and 65% of metapeaks are overlapped by an ATAC region.

Chromatin state data - *D. melanogaster*: Kharchenko et al. (Kharchenko et al. 2011) identified 21,955 chromatin state regions (in S2-DRSC and ML-DmBG3-c2 cells using ChIP-chip), some of which were overlapping, and assigned the genome in 200 bp bins to one of nine chromatin states. We looked for overlap using just the apex of the metapeak, a region of 200 bases centered on the apex and the full metapeak, requiring an overlap of at least ten bases with the chromatin regions for the latter two measures. The three measures produced similar results, with the three states associated with actively transcribed among the most shared, and the states with heterochromatin showing less overlap (**Supplemental_File_S8**). The largest number of ChIP metapeaks and the largest number of metapeak bases overlap transcriptionally silent intergenic euchromatin, perhaps reflecting the wider range of genes active in the whole animal as opposed to cell lines.

***C. elegans* – ATAC-seq**: Janes et al. (Jänes et al. 2018b) identified 42,245 elements accessible in at least one *C. elegans* stage. As measured by an overlap of at least ten bases, 27,052 of the

ChIP metapeaks overlap 33,911 of Janes et al. elements (**Supplemental_File_S8**). In terms of overlap at the base pair level, in the embryonic data for example, 18% of bases labeled with a chromatin state overlapped an embryonic modERN metapeak and 98% of bases in embryonic metapeaks were overlapped by a region labeled with a chromatin state in the embryonic state data.

Janes et al. (Jänes et al. 2018b) further annotated their accessible regions defining 13,596 protein-coding promoters and 19,231 putative enhancers. Review of the overlap between their regions and the metapeaks reveals the largest numbers of regions overlapped by count by the metapeaks are those labeled as enhancer and coding promoter. Metapeaks that overlap ATAC regions have larger numbers of peaks (30.8 +/- 60.0) compared to metapeaks not overlapping ATAC regions (2.7+/- 4.5) and, similarly, ATAC regions that overlap a metapeak have stronger signals (18.2+/-30.0) compared to those that don't (5.0+/-5.4) , suggesting that metapeaks with low occupancy are less likely to be observed in ATAC-seq data.

To examine further the 3,756 accessible regions annotated as an enhancer and that were not overlapped by a metapeak, we assigned those regions to target genes in the same way we assigned metapeaks to target genes. Those 3,756 accessible regions were found to be more distant from the transcription start site, typically more 3' of the transcription start site, than were those accessible regions annotated by enhancer that did overlap a metapeak. Of those 2,910 genes targeted by enhancers (Jänes et al. 2018b) but not overlapped by a metapeak, 2,749 (95%) were targeted by other, non-overlapping metapeaks. Thus only 161 genes are identified as being targeted by enhancers (Jänes et al. 2018b) that are not targeted by a metapeak. Of those 161, there are only 80 that have a TPM of at least 100 in at least one cell type in the embryo (Packer et al. 2019).

***C. elegans* – eY1H comparison:** Fuxman Bass et al (Fuxman Bass et al. 2016) examined interactions between 409 of *C. elegans* TFs and the 500 bases upstream of 3,125 of *C. elegans* gene TSSs, using an enhanced yeast one-hybrid assay (eY1H) to create a gene-centered physical protein DNA interactions (PDI) network. Their results contained 26,497 PDIs of which 21,714 were defined as high quality (dataset_EV1). These high quality interactions were between 2,576 target gene promoters and 366 TFs. They report finding overlap for 20% of the 46 shared TFs detected by eY1H and by ChIP as of the 2014 release of the *C. elegans* modENCODE project. For the current modERN dataset, there are 182 TFs shared by this project and the eY1H dataset. When looking in the 500 bases upstream of the TSS for the 14,198 TF/target pairs from those 182 TFs, 1213 (8.5%) of have at least one ChIP peak overlapping that region by at least 10 bases. Of the 11,756 high quality Fuxman Bass TF/target pairs, 945 (8.0%) overlap at least one ChIP peak.

To understand better the low level of overlap between the two data sets, we examined the relationship between expression of the eY1H TFs and their targets in the single cell data sets, in the same way we had done for the ChIP-seq TFs and their targets. We compared all of the eY1H TFs against the embryo, L2 and YA scRNA-seq sets. Overall, the fraction of TFs with a cosine angle greater than 0.2 was lower in the eY1H data. For example, for the embryo, of 304 eY1H TFs with high quality targets, and more than one target, 8.8% (27/304) had a cosine angle > 0.2. (58 had only 1 target and no angle was calculated. The eY1H data set has only 123 have more than 10 targets; more than half the targets are associated with just 18 TFs). By contrast, 26.8% (33/123) of the ChIP embryo TFs had a cosine angle of > 0.2. Some notable TFs have low angles or very few targets in the eY1H set, e.g. HLH-1 has only one target. Of the eY1H that have a cosine angle of > 0.2, about half also have a high cosine angle in the ChIP data.

Interestingly, those TFs have more than a third of TF-target pairs overlapping between eY1H and ChIP, with PHA-4 showing 54% overlap (20/37). For others, the eY1H data may be more informative data than the ChIP-seq data, e.g., ZTF-6 has a high angle (0.21) in the eY1H data and little overlap with the ChIP data (2/95) and a weak cosine angle (0.07) in the ChIP data.

TF Pearson correlations for the worm and fly stages based on co-occurrence in same metapeak

Pearson correlations were used to evaluate how often TFs occur together in the set of metapeaks. Pairs of TFs that occur often in the same metapeaks will have higher correlation values. The initial set of metapeaks used in this analysis ranged in size of 2 peaks to 84 peaks, corresponding to the upper limit of non-HOT sites in the worm. A binary matrix was constructed for each life stage in the worm and fly. These matrices were metapeaks by TFs, with a value of zero indicating the TF is not in the metapeak and value one indicating there was at least one peak of that TF in the metapeak. For each lifestage in each species, if the number of TFs in the metapeak was less than two, that metapeak was not included in the calculations. For each of the binary matrices, a Pearson correlation was then calculated for each pair of TFs across the metapeaks. The correlation matrices were reordered by hierarchical clustering and displayed in heatmap form.

Motif methods

For known motifs, we collected fly and worm motifs determined by *in vitro* experiments, including SELEX, PBM, and B1H, from the Cis-BP database (Weirauch et al. 2014) with direct and inferred evidence. For the motifs of a TF, we concatenated all the unique position-weighted matrices (PWMs) into a single file in the MEME motif format. In total, 374 fly experiments and

227 worm experiments have known *in vitro* motifs. Locations of the motif occurrence in the genome were identified by FIMO (Grant, Bailey, and Noble 2011). We intersected the motif occurrence regions and the peaks or metapeaks regions by Bedtools (Quinlan and Hall 2010).

For motif inference, we first converted the ChIP-seq regions to fasta files using Bedtools (with genome versions dm6 and ce11 for fly and worm respectively). We then used STREME (v5.4.1) (Bailey 2021) with default parameters to infer TF binding motifs from those ChIP-seq regions. We compared the inferred motifs to known motifs by TOMTOM (Gupta et al. 2007). For each experiment, if any one of the best three inferred motifs matched any one of the known motifs for that TF, we counted it as an experiment with successful inference.

For each ChIP-seq experiment, we used each of the following criteria to create several different input subsets: (i) keeping all peaks with no filters; (ii) keeping top 20% peaks sorted by SPP score; (iii) removing peaks that fall in metapeaks with size larger than 277 or 85, respectively for fly and worm; (iv) removing peaks that fall in metapeaks with sizes larger than 53 or 31, respectively for fly and worm. Note that STREME will report an error if there were too few or zero peaks after filtering, and such subsets of that experiment will be disregarded. To make fair comparisons, for each of the subset types (ii), (iii) and (iv), we randomly sampled the same number of peaks from all peaks as well. We repeated the sampling three times and reported the average. All inferred motifs from all groups are available on https://github.com/modERNresource.

For the 288 and 290 experiments in fly and worm respectively where the TF did not have *in vitro* motifs, we compared their inferred motifs from group (i) to those from group (iv), described above. If any one of the top three inferred motifs from group (i) was significantly similar (TOMTOM q-value<0.05) to one of the top three inferred motifs from group (iv), we considered this experiment as one with consistent inference results.

Peak and metapeak target assignment

We relied on the following assumptions to guide the peak and metapeak assignments to target genes. TFs operate at transcription start sites (TSS) and the closer the peak or metapeak is to the TSS, the more likely it is operating to influence gene expression of that target gene. The peak or metapeak can be either upstream or downstream to the TSS. It is not likely that a TF will influence gene expression of a target gene, if there is an intervening TSS of a different gene. It is ambiguous when a peak or metapeak lies between two different genes' TSSs, and the distance to each of these two TSSs is not dramatically different. In such cases, more than one target gene assignment for a single peak or metapeak was made.

The following describes the algorithm used to assign peaks and metapeaks to target genes. Using the single base apex of the peak or metapeak as the location of a peak or metapeak, we found the closest TSS to the apex in either the positive or negative genomic direction and assigned the gene with the closest TSS as the primary target. Other information recorded about the assignment included the distance and direction to the TSS, whether the apex was within an exon or intron, or whether the apex was outside the gene, on either the 5' or 3' side of the gene. After assignment of a primary target, an alternate target was chosen if the gene with the next nearest TSS to the apex in the opposite genomic direction from the primary target was a different gene than the primary. The result was that an apex located between two genes was assigned both a primary and alternate target gene. The distance and other information recorded can be used to assess how confident the assignment is to the alternate target gene. If the distance is large compared to the distance to the primary target, the alternate target assignment may be less likely to be operational. In addition, the location of peaks and metapeaks relative to the features of the target genes and their transcripts were recorded, again using proximity of the peak to the feature. The recorded relationships included being extragenic (either 5' or 3') or intragenic (either an internal TSS's, an exon or an intron). For genes with multiple TSS's both the transcript and gene positions were recorded.

Remapping fly single cell data

As the initial analysis of the single cell RNA-seq fly data set (Calderon et al. 2022) for TF-target relationships yielded disappointing results, we attempted to refine the annotation. To find additional cell types, we removed the very early cells (0-6 hours) along with yolk nuclei, subsetting their clusters 1, 5, 8, 9, 11, 12, 15, 16, 18, 21, 22, 23 to create a new cell data set (cds) (238,445 cells out of 547,805 starting cells). Reducing the dimensions of this subset using first 300 PCAs and then 2 dimensions with UMAP in monocle3 (https://cole-trapnell-lab.github.io/monocle3/) with default parameters produced a more highly structured map (**Supplemental Figure 9A**). After reclustering and grouping some clusters, dividing others and incorporating the prior neural annotations (Calderon et al. 2022), we produced the anonymous annotation (**Supplemental Figure 9B**). The cells had relatively low counts, and highly expressed genes that, based on other evidence, should have been very cell specific, had a low-level background throughout the maps. These factors may have limited resolution and contributed to the need to use high expression thresholds in the random forest models. Calculating the TPMs for each gene in each cell type and weighting by the differential expression yielded a heuristic score for potential markers of each cell type. These markers were then compared to the fly in situ data set (https://insitu.fruitfly.org/cgi-bin/ex/insitu.pl) of the Berkeley Drosophila Genome Project and other literature to assign each cell type anatomical names (**Supplemental Figure 9C, D**). After two additional rounds of refinement, we identified a total of 83 cell types in addition to the cell types removed from the cell subset (maternal, yolk nuclei, early germline, etc.) (**Supplemental_File_S9**). To obtain a more robust estimate of gene expression, we used a bootstrap calculation (1000 iterations with replacement) similar to that used by Packer et al. to obtain the TPM values of each gene in each cell type (Packer et al. 2019).

TF versus target expression

To measure the complexity of expression we calculated the entropy (reference https://en.wikipedia.org/wiki/Entropy_(information_theory)) of each across the different cell types with their time bins defined in the worm embryo data set (Packer et al., 2019) and the reanalyzed fly embryo data set from Calderon et al., 2022. Entropy was calculated in R using the function:

```
entropy <- function(x){
  s <- sum(x)
  g <- 0
  for (i in 1:length(x)){
    if (x[i]!=0){
      p <- x[i]/s
      g <- g - p*log2(p)
  }
 }
 return(g)
  }
```

Where x is the tpm of the gene in each cell type. High values of entropy indicate a uniformity of expression across cell types and low values indicate very restricted expression in just one or a few cell types.

After exploring the impact of various methods of filtering the input peaks and targets, we settled on removing all targets of HOT sites and all singleton metapeaks. Removing targets of peaks with low relative signal strength did improve the scores of some TF-target pairs, but severely reduced the number of targets for some TFs. Expression plots were created in R using ggplot, assigning each cell type to a broad cell class and ordering the plots by the cell class and then the cell type.

Random forest model of cell type expression

Regression models were trained to predict single cell gene expression in individual cell types from ChIP-seq TF binding sites. The goal of building these models was to determine which TFs were most important in determining expression in each individual cell type. A random forest machine learning method (https://urldefense.com/v3/__https://www.randomforestsrc.org/__;!!K-Hz7m0Vt54!iUhkkDeBFzaT4erzyUP5WycP36QDAPYmgmzrL9D0w_nFOd2BH9xMRu8_P0dHwHJJM6qnAvdWl-uPHQ$)

was selected for this application because of the non-linear relationship expected between TF binding strength and gene expression. Also, the random forest model can be interrogated, after it is trained, for the most important TFs in predicting the gene expression in the cell type.

The independent variable in the predictor matrix is the TF binding signal strength. The binding signal is part of the output for each called peak by the SPP peak calling program. The dependent or response variable is each cell type's gene expression profile. In order to relate the response variable to the predictor variables, each metapeak must be associated with one or more target genes whose single cell expression has been measured in the cell type. After modeling all the cell types in the embryonic, larval, and adult stages, a matrix of cell types by TF will be produced for each of those life stages. The values in this matrix reflect the relative importance of the TF in determining the expression of the target genes in the cell type.

Relating TF peaks to target genes is done on the basis of genomic proximity of the containing metapeak to the TSS. All of the TF binding sites have been grouped into metapeaks as described above. These metapeaks are assigned a primary target gene, which is the gene with a TSS closest to the apex of the metapeak. An alternative target gene may also be assigned to the metapeak, which is a gene in the opposite genomic direction from the primary target that has the closest TSS. Each binding site in the metapeak is assigned to the target genes assigned to the metapeak.

Multiple different models were built for all the cell types, depending on which metapeaks, TFs, and targets are selected for modeling, and any transformation performed on the signal binding strength of each binding site. Many of the ChIP-seq binding sites were determined in a single life stage, while some were done at multiple stages. When modeling the expression in a cell type, only binding sites measured in a close life stage were used in the modeling. Thus, embryonic ChIP-seq experiments were used to model embryonic single cell expression, while L1, L2, and L3 stage ChIP-seq experiments were used to model L2 larval expression. and adult and L4 ChIP-seq experiments were used to model adult expression for the worm. Only embryonic ChIP-seq experiments in the fly were used as this was the only stage for which single cell expression data was available.

The SPP peak calling program outputs a signal strength for each peak called. In order for these signal strengths to be used in the modeling, they need to be normalized so that each experiment is comparable. To do the normalization, the peaks in each experiment are sorted by their signal strength. They are then assigned a normalized signal strength between zero and one based on their rank in the experiment. The peak with the strongest signal in the experiment will be assigned a normalized signal of one, and the peak with the lowest signal strength in the experiment will get a zero normalized signal strength.

The predictor matrix is constructed with the independent variables in the columns, all the TFs measured in the experiments done in the life stage of the cell type being modeled. The rows of the predictor matrix represent the data points, an association of a cluster to a target gene. The values in the matrix in a given row are a measure of the binding site for each TF in the cluster. If a cluster is associated with more than one target, there will be more than one row in the matrix corresponding to that cluster. Genes that have no associated cluster are not represented in the model. Some genes will have multiple associated metapeaks, and there will be a row for the gene in the predictor matrix for each associated metapeak. The response vector is the single cell expression in the cell type being modeled for the target genes in the rows of the predictor matrix. The number of rows in the predictor matrix will equal the number of entries in the response vector.

Forty different predictor matrices and response vectors were constructed for each worm cell type, depending on which TFs were selected (feature selection), which metapeaks were selected, which targets were selected, and which function of normalized signal strength was selected.

Transcription factor selection:

1) only factors with nonzero measured single cell expression in the cell type

2) only factors with expression in the cell type of at least 5% of the maximum expression of all the cell types in the life stage.

Metapeak selection:

1) no HOT site and no singleton clusters, only those clusters with 2 to 84 peaks (of all life stages) in the cluster

2) clusters with 2 to 30 peaks

3) clusters with 2 to 84 peaks, but within 2KB of the target gene's TSS or in the first intron or first exon of the target gene

4) clusters with 2 to 30 peaks, but within 2KB of the target gene's TSS or in the first intron or first exon of the target gene

5) clusters with target genes that do not have a HOT site associated with them

Metapeak target selection:

1) Use primary targets only

2) In addition to primary targets, use alternative targets that are less than twice the distance from the cluster as the primary target

Signal Strength:

1) use the normalized signal rank without modification

2) divide the normalized signal rank by the log of the distance between the cluster and its target

Each cell type in each life stage was modeled with all forty different predictor models. When training a random forest, we used a procedure referred to as bagging, which means each decision tree is built with a random selection of the data points. The result is an out of bag ensemble of data points, not used in the training, that can be used to assess the accuracy of each of the models. This out of bag ensemble also makes it possible to assess the importance of each TF in the accuracy of the prediction model by seeing how much the model error changes when a given TF is permuted in the trained model. Those TFs that cause the greatest increase in the model error when permuted, are the most important for the prediction.

The models were ranked by the root mean square error for each cell type. The model ranked best for the most cell types was selected as the best modeling approach for each life stage. For the worm, the model that used TFs with at least 5% of the maximum expression, no HOT sites and no singleton clusters, close alternative targets, and distance modified signal strength was judged to be the best.

For the fly embryonic stage, a different model performed best. This model used TFs where the expression was greater than the mean expression plus one standard deviation. Metapeaks more than 84 peaks were not included in the fly model, consistent with the worm models.

REFERENCES

Araya, Carlos L., Trupti Kawli, Anshul Kundaje, Lixia Jiang, Beijing Wu, Dionne Vafeados, Robert Terrell, et al. 2014. “Regulatory Analysis of the *C. elegans* Genome with Spatiotemporal Resolution.” *Nature* 512 (7515): 400–405.

Bailey, Timothy L. 2021. “STREME: Accurate and Versatile Sequence Motif Discovery.” *Bioinformatics (Oxford, England)* 37 (18): 2834–40.

Chiaromonte, F., V. B. Yap, and W. Miller. 2002. “Scoring Pairwise Genomic Sequence Alignments.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 115–26.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” *Bioinformatics (Oxford, England)* 27 (7): 1017–18.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. 2007. “Quantifying Similarity between Motifs.” *Genome Biology* 8 (2): R24.

Kent, W. James, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. 2003. “Evolution’s Cauldron: Duplication, Deletion, and Rearrangement in the Mouse and Human Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (20): 11484–89.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6): 996–1006.

Kharchenko, Peter V., Artyom A. Alekseyenko, Yuri B. Schwartz, Aki Minoda, Nicole C. Riddle, Jason Ernst, Peter J. Sabo, et al. 2011. “Comprehensive Analysis of the Chromatin Landscape in Drosophila Melanogaster.” *Nature* 471 (7339): 480–85.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv [q-Bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics (Oxford, England)* 26 (6): 841–42.

Schwartz, Scott, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller. 2003. “Human-Mouse Alignments with BLASTZ.” *Genome Research* 13 (1): 103–7.