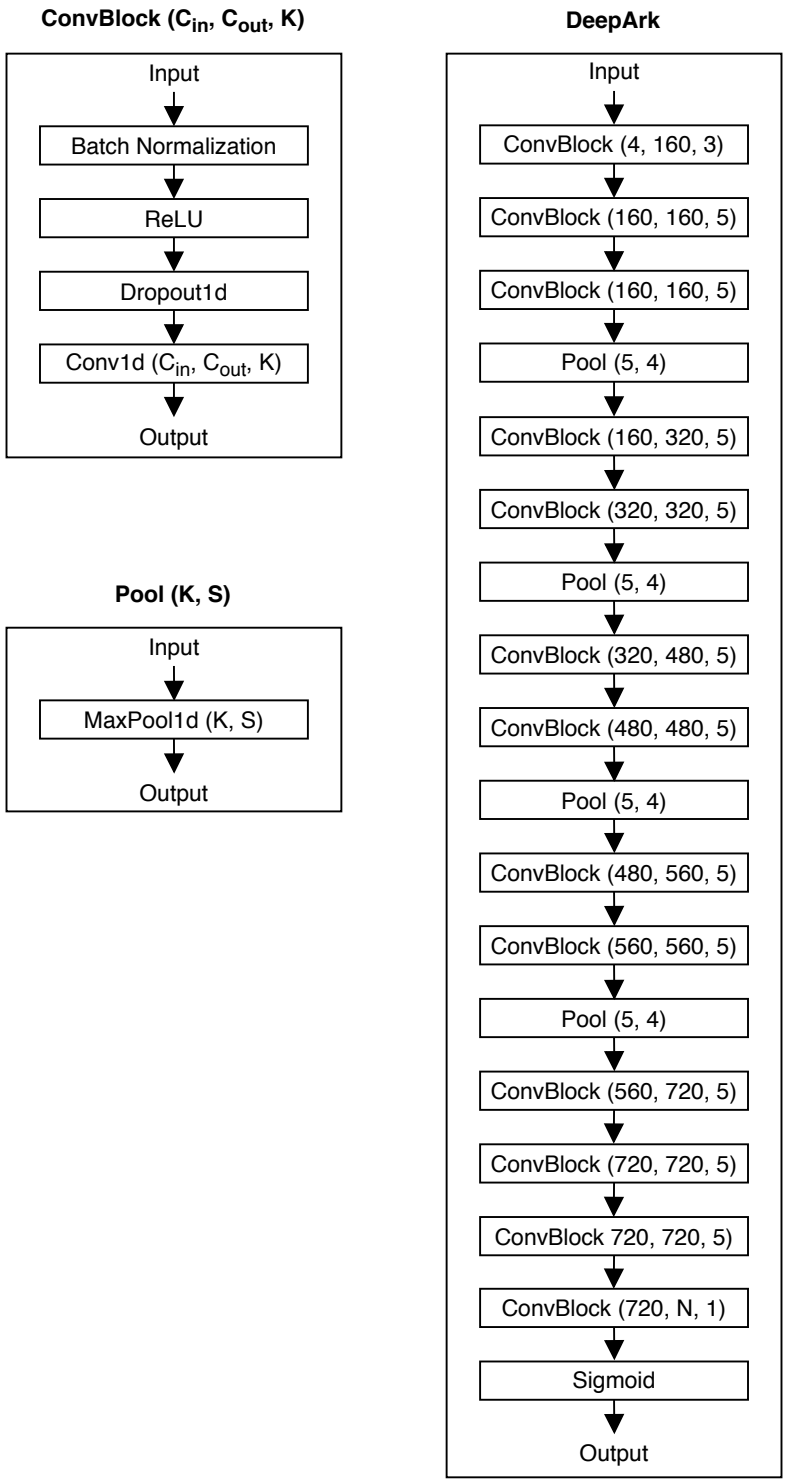
Supplemental Material for

Modeling transcriptional regulation of model species with deep learning

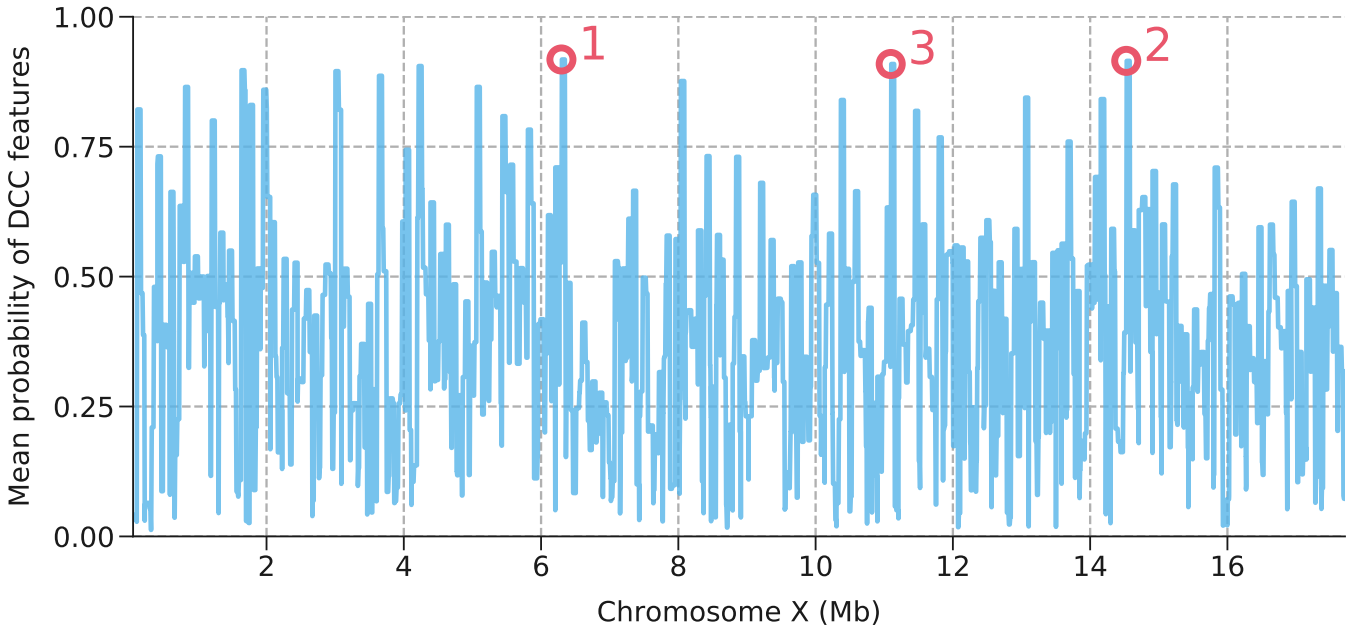
Evan M. Cofer, João Raimundo, Alicja Tadych, Yuji Yamazaki, Aaron K. Wong, Chandra L. Theesfeld, Michael S. Levine, and Olga G. Troyanskaya

List of Supplemental Figures:

List of Supplemental Tables:

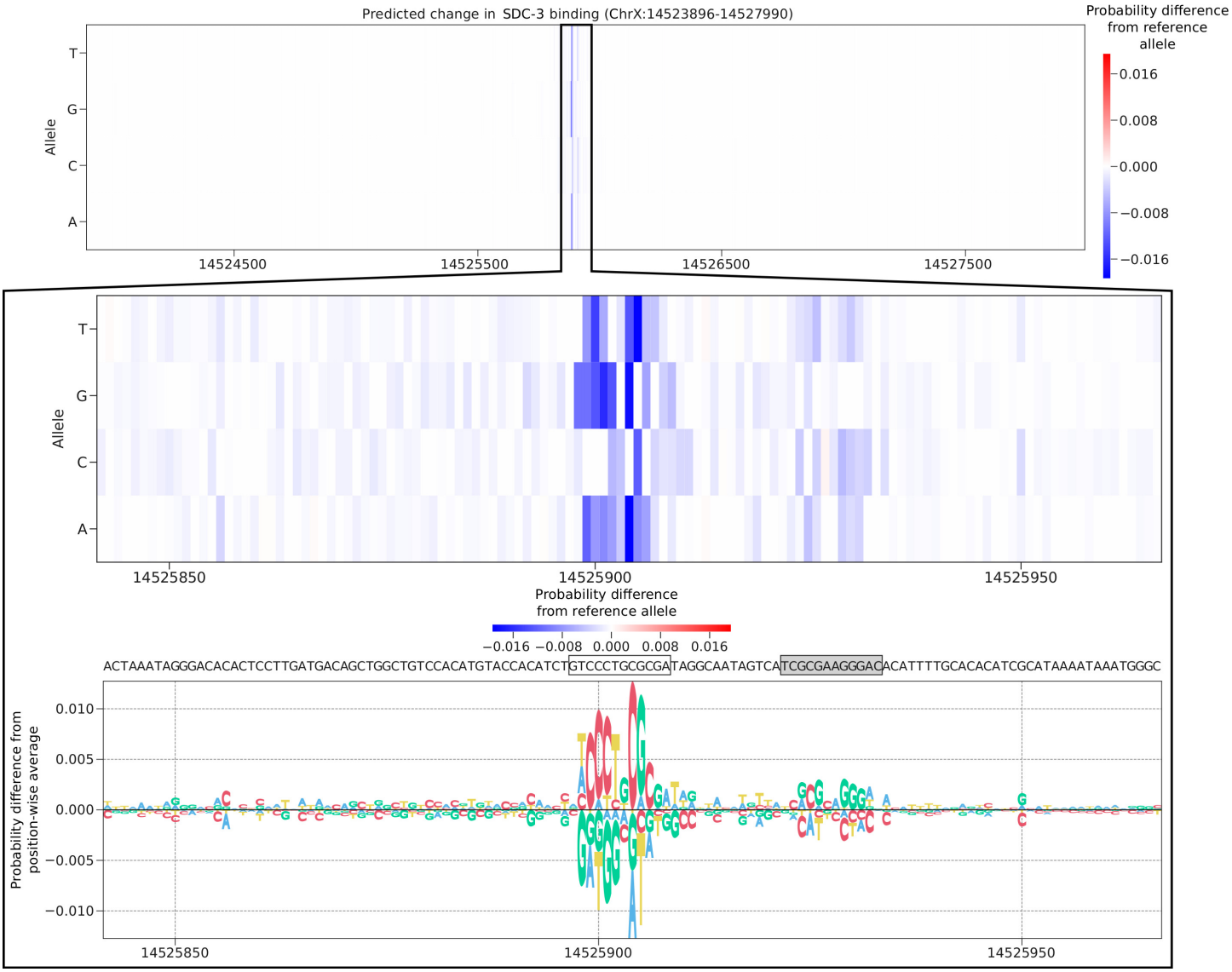


Supplemental Figure S1: Overview of the DeepArk model architecture. Convolutional blocks have C_{in} input channels, C_{out} output channels, and a kernel size of K. Pooling blocks have a kernel size of K and a stride of S. The dropout rate used by the spatial dropout layers during training varied according to species (**Supplemental Table S7**).

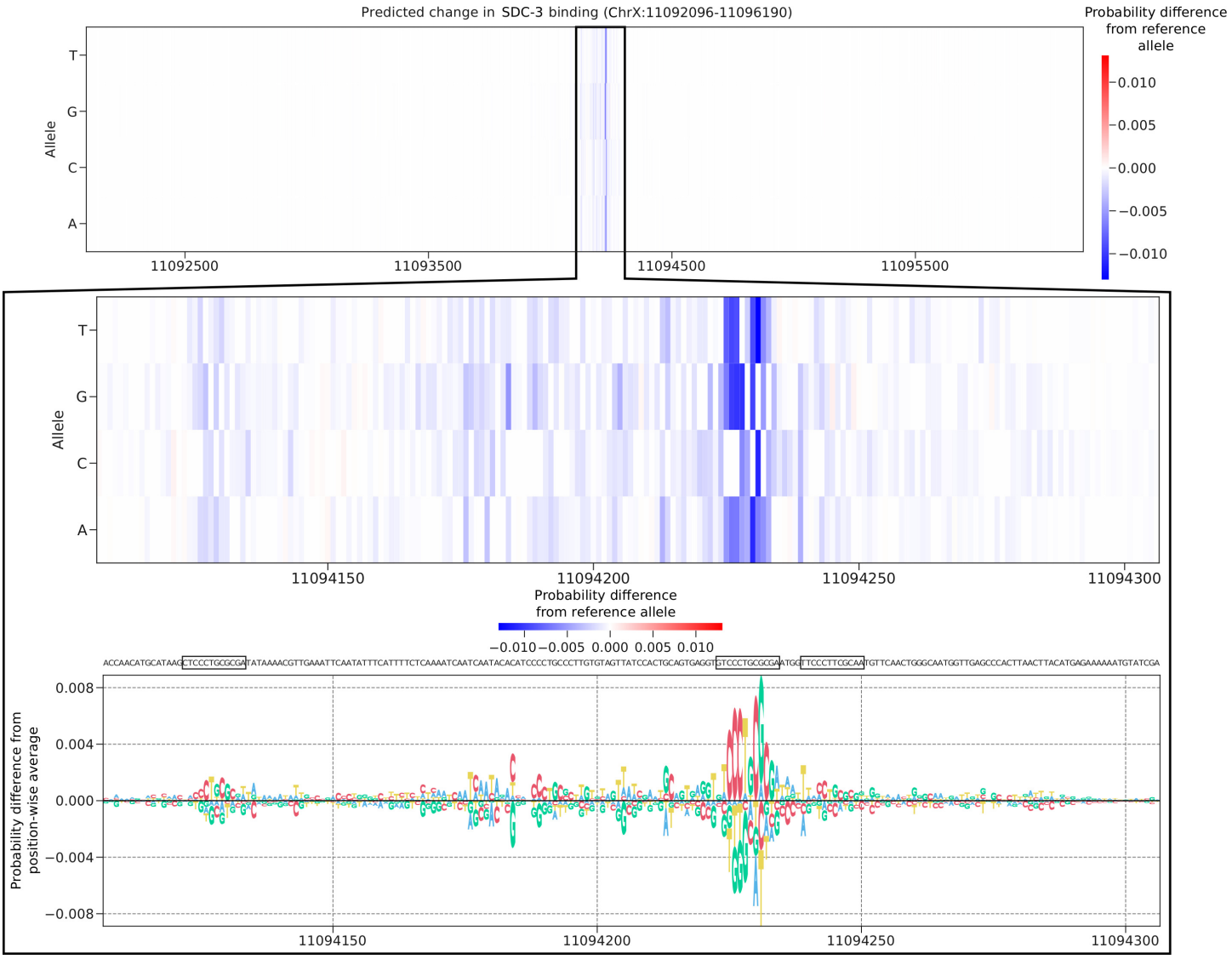


Supplemental Figure S2: DCC-bound regions of the *C. elegans* X Chromosome as predicted by DeepArk. The mean probability of DCC features (**Supplemental Table S3**) throughout the X Chromosome of *C. elegans*. To enhance readability, we plot the maximum value per 50 Kb bin of the chromosome. The positions with the largest (ce11:ChrX:6294496-6298590), second largest (ce11:ChrX:14523896-14527990), and third largest (ce11:ChrX:11092096-11096190) predicted probabilities are marked with red circles and labeled with their rank.

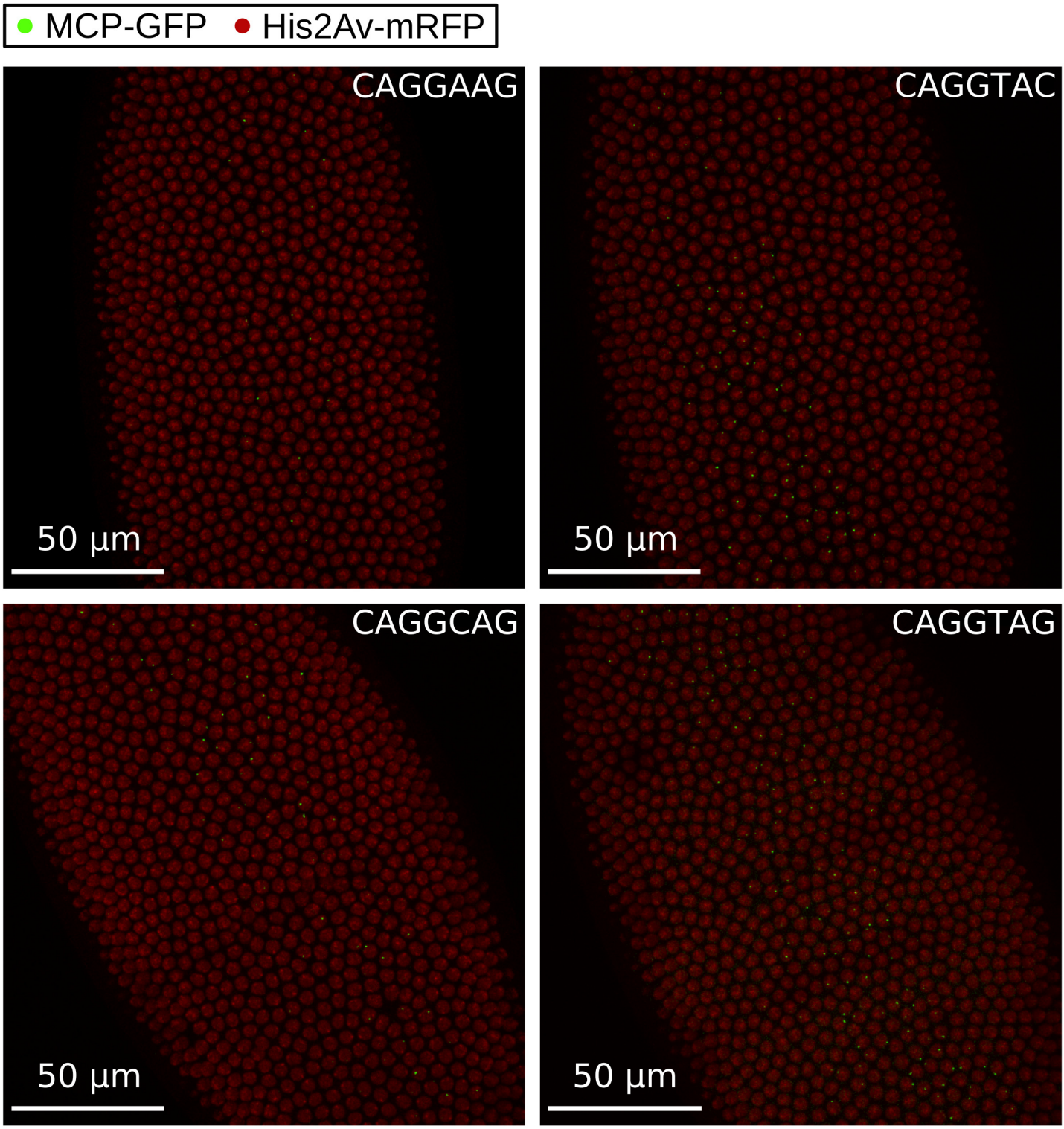
Supplemental Figure S3: *In silico* saturated mutagenesis applied to the first high-confidence DCC binding site in *C. elegans*. In the top panel, we show the predicted effect of every possible SNP in the first high-confidence DCC binding site (ce11:ChrX:6294496-6298590) on the probability of SDC-3 binding (accession no. SRX2228883) relative to the prediction for the reference sequence for all positions in the 4095 bp input sequence. The middle panel is the same as the top panel, except zoomed in on the most critical 100 bp at the center of the sequence. The bottom panel is also zoomed in on the most critical 100 bp at the center of the sequence, but the score for a particular variant in this panel is visualized as the difference between the predicted probability for the sequence containing that variant and the mean predicted probability of all alleles at the same position. For clarity, the reference sequence is shown along the top of the bottom panel. The positions in the reference sequence that contain significant matches for the "recruitment elements on X" or "*rex*" motif (Jans et al. 2009) are outlined. If a significant match occurs on the forward strand, the box has a light grey fill, otherwise it has a transparent fill. Note that the right-most motif hit contains a near-perfect match (TCGCGCAGGGAA) to the *rex* consensus sequence, and appears highly predictive of SDC-3 binding. Mutations at this site also appear to greatly diminish the predicted probability of SDC-3 binding. The other *rex* motif hits (AAGCGAAGGGAC on the left, and TGGCGCAGGGGG in the middle) deviate more from the canonical *rex* motif, and appear less critical to SDC-3 binding than the right-most one.



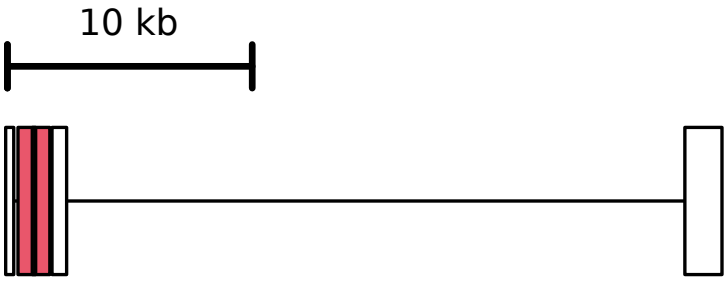
Supplemental Figure S4: *In silico* saturated mutagenesis applied to the second high-confidence DCC binding site in *C. elegans*. In the top panel, we show the predicted effect of every possible SNP in the second high-confidence DCC binding site (ce11:ChrX:14523896-14527990) on the probability of SDC-3 binding (accession no. SRX2228883) relative to the prediction for the reference sequence for all positions in the 4095 bp input sequence. The middle panel is the same as the top panel, except zoomed in on the most critical region of 125 bp near the center of the sequence. The bottom panel is also zoomed in on the most critical 125 bp near the center of the sequence, but the score for a particular variant in this panel is visualized as the difference between the predicted probability for the sequence containing that variant and the mean predicted probability of all alleles at the same position. Additionally, the reference sequence is shown along the top of the bottom panel. The positions in the reference sequence that contain significant hits for the "recruitment elements on X" or "*rex*" motif (Jans et al. 2009) are outlined. If a significant hit occurs on the forward strand, the box has a light grey fill, otherwise it has a transparent fill. The left (TCGCGCAGGGAC on the reverse strand) and right (TCGCGAAGGGAC) *rex* motif hits appear important to SDC-3 binding.



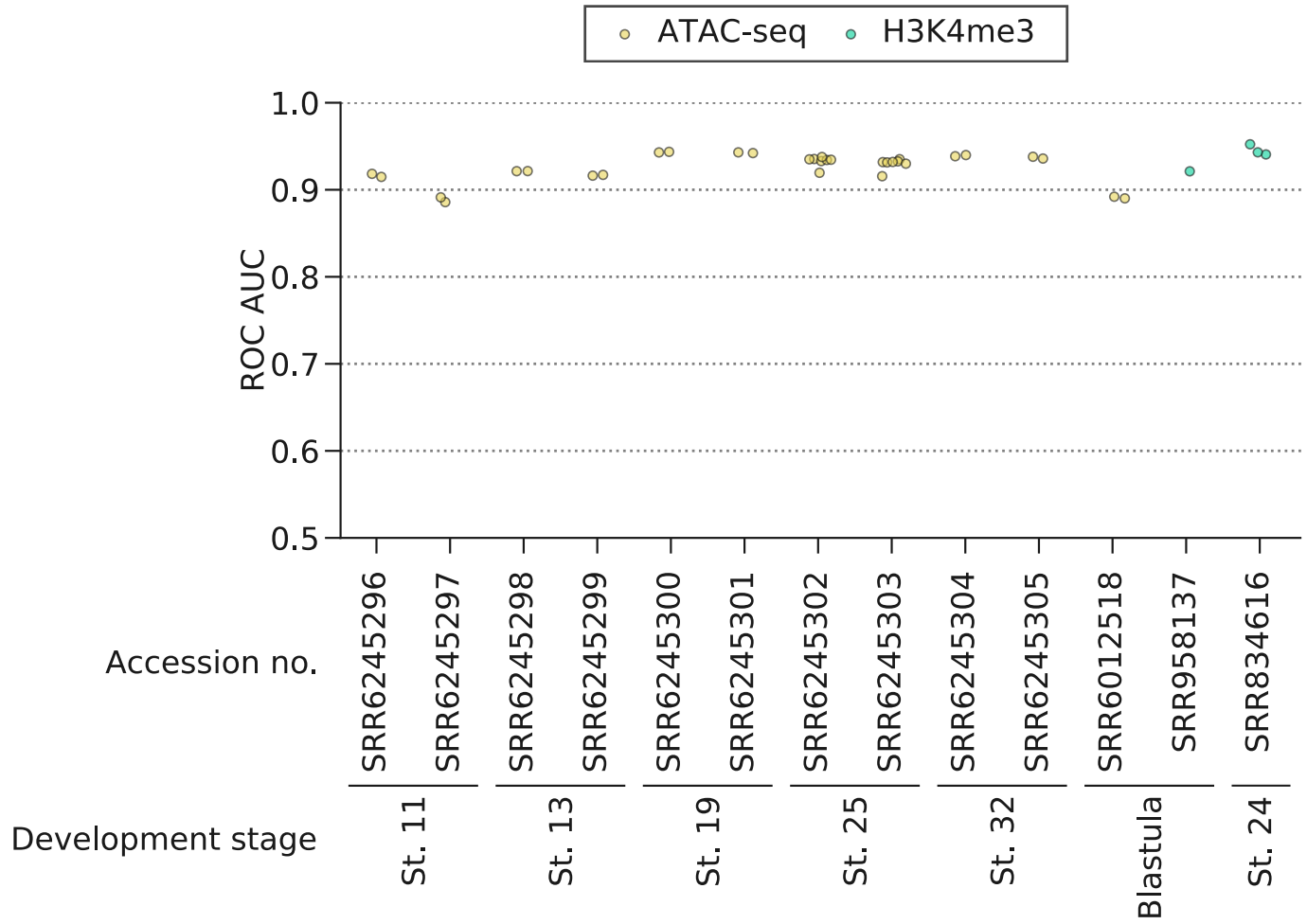
Supplemental Figure S5: *In silico* saturated mutagenesis applied to the third high-confidence DCC binding site in *C. elegans*. In the top panel, we show the predicted effect of every possible SNP in the third high-confidence DCC binding site (ce11:ChrX:11092096-11096190) on the probability of SDC-3 binding (accession no. SRX2228883) relative to the prediction for the reference sequence for all positions in the 4095 bp input sequence. The middle panel is the same as the top panel, except zoomed in on the most critical region of 200 bp near the center of the sequence. The bottom panel is also zoomed in on the most critical 200 bp near the center of the sequence, but the score for a particular variant in this panel is visualized as the difference between the predicted probability for the sequence containing that variant and the mean predicted probability of all alleles at the same position. For clarity, the reference sequence is shown along the top of the bottom panel. The positions in the reference sequence that contain significant hits for the "recruitment elements on X" or "*rex*" motif (Jans et al. 2009) are outlined. If a significant hit occurs on the forward strand, the box has a light grey fill, otherwise it has a transparent fill. Reading from left to right, the first *rex* motif hit (TCGCGCAGGGAG on the reverse strand) is a perfect match to the consensus sequence. The second *rex* motif hit (TCGCGCAGGGAC on the reverse strand) is a near perfect match, while the third *rex* motif hit (TTGCGAAGGGAA on the reverse strand) is more degenerate. Nevertheless, the clustered second and third hits appear highly predictive of SDC-3 binding, which is consistent with existing literature showing that closely spaced *rex* sites result in improved DCC localization (McDonel et al. 2006).



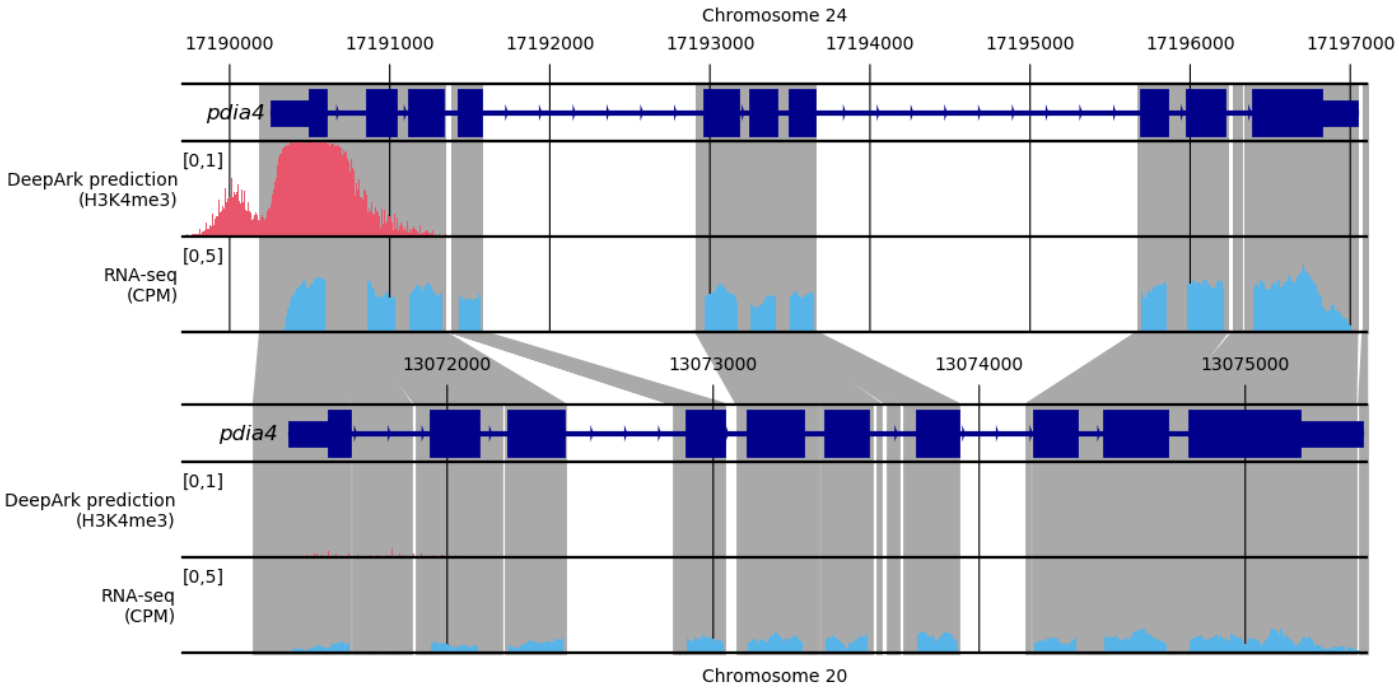
Supplemental Figure S6: Live imaging of the *T48* mesodermal enhancer. Fluorescence microscopy of *Drosophila* embryos during minute 20 of nuclear cycle 14 showing active transcription of the *T48* mesodermal enhancer, and illustrates the different levels of transcriptional activation driven by the different alleles. These are the raw versions of the false colored images presented in the main text (**Figure 3C**).



Supplemental Figure S7: Overview of the *T48* gene. The *T48* gene (dm6:Chr3R:26881734-26910997) is shown, and the mesodermal enhancer (dm6:Chr3R:26882237-26883537) is drawn as a red box in the first intron. The location of the variants (dm6:Chr3R:26882886-26882889) is visualized as a black vertical line occurring near the center of the enhancer.



Supplemental Figure S8: The DeepArk model for *D. rerio* accurately predicts regulatory features in the *O. latipes* genome. The plot shows DeepArk's test set performance for *D. rerio*-to-*O. latipes* interspecies regulatory feature prediction, as quantified by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve (**Supplemental Table S6**). The horizontal axis indicates the target regulatory feature in *O. latipes* that we sought to predict with the DeepArk model for *D. rerio*. There are multiple scatter points for each target regulatory feature in *O. latipes* because there is generally more than one *D. rerio* regulatory feature that is comparable to a given *O. latipes* target regulatory feature. For instance, each ATAC-seq experiment from *O. latipes* during Stage 13 (accession numbers SRR6245298 and SRR6245299) would be predicted by each of the two DeepArk regulatory features corresponding to ATAC-seq from *D. rerio* at 13 hpf (accession numbers DCD003087SQ and DCD003090SQ).



Supplemental Figure S9: Interspecies predictions with DeepArk indicate diminished *cis*-regulatory activity in *O. latipes* relative to *D. rerio*. For the *pdia4* gene, which is highly conserved in both *D. rerio* (top) and *O. latipes* (bottom), DeepArk's predictor for H3K4me3 at 6 hpf (accession no. DCD000648SQ) predicts a loss of H3K4me3 at *pdia4*'s promoter in the *O. latipes* genome, which would be associated with diminished or loss of expression at 13 hpf in *O. latipes* (Tena et al. 2014; Marlètaz et al. 2018). Accordingly, normalized coverage counts in RNA-seq from *D. rerio* at 6 hpf and *O. latipes* at 13 hpf show diminished expression of *pdia4* in *O. latipes* relative to *D. rerio*. CPM, counts per million mapped reads.

Supplemental Table S7: The hyperparameters used by each DeepArk model.

Species	Initial learning rate	Dropout probability	Batch size	Weight decay	Momentum
Caenorhabditis elegans	0.1	0.15	128	3×10^{-6}	0.9
Danio rerio	0.1	0.2	128	1×10^{-6}	0.9
Drosophila melanogaster	0.1	0.2	128	3×10^{-6}	0.9
Mus musculus	0.3	0.15	128	1×10^{-6}	0.9

Supplemental Table S9: Thresholds used to filter datasets for each species.

Species	Assay type	Target type	Minimum peaks	Minimum mapped reads
Caenorhabditis elegans	DNase-seq	chromatin	500	5000000
Caenorhabditis elegans	ChIP-seq	transcription factor	500	2000000
Caenorhabditis elegans	ChIP-seq	histone mark – narrow	500	2000000
Caenorhabditis elegans	ChIP-seq	histone mark – broad or enriched in repetitive regions	500	5000000
Danio rerio	ATAC-seq	chromatin	2500	25000000
Danio rerio	ChIP-seq	transcription factor	2500	10000000
Danio rerio	ChIP-seq	histone mark – narrow	2500	10000000
Danio rerio	ChIP-seq	histone mark – broad or enriched in repetitive regions	2500	25000000
Drosophila melanogaster	DNase-seq	chromatin	500	5000000
Drosophila melanogaster	ChIP-seq	transcription factor	500	2000000
Drosophila melanogaster	ChIP-seq	histone mark - narrow	500	2000000
Drosophila melanogaster	ChIP-seq	histone mark – broad or enriched in repetitive regions	500	5000000
Mus musculus	DNase-seq	chromatin	5000	50000000
Mus musculus	ChIP-seq	transcription factor	5000	20000000
Mus musculus	ChIP-seq	histone mark – narrow	5000	20000000
Mus musculus	ChIP-seq	histone mark – broad or enriched in repetitive regions	5000	50000000

Supplemental Table S13: Primer sequences used for *T48* enhancer mutants.

Name	Primer sequence
T48_CAGGAAG_fw	TCGCACGCAGAACTTCCTGCCTCTGGCCATCCC
T48_CAGGAAG_rv	GGGATGGCCAGAGGCAGGAAGTTCTGCGTGCGA
T48_CAGGTAC_fw	CGCACGCAGAAGTACCTGCCTCTGGCCATCCC
T48_CAGGTAC_rv	GGGATGGCCAGAGGCAGGTACTTCTGCGTGCG
T48_CAGGCAG_fw	TCGCACGCAGAACTGCCTGCCTCTGGCCATCCC
T48_CAGGCAG_rv	GGGATGGCCAGAGGCAGGCAGTTCTGCGTGCGA
T48_CAGGTAG_fw	CGCACGCAGAACTACCTGCCTCTGGCCATCCCGCTTGCAC
T48_CAGGTAG_rv	GTGCAAGCGGGATGGCCAGAGGCAGGTAGTTCTGCGTGCG