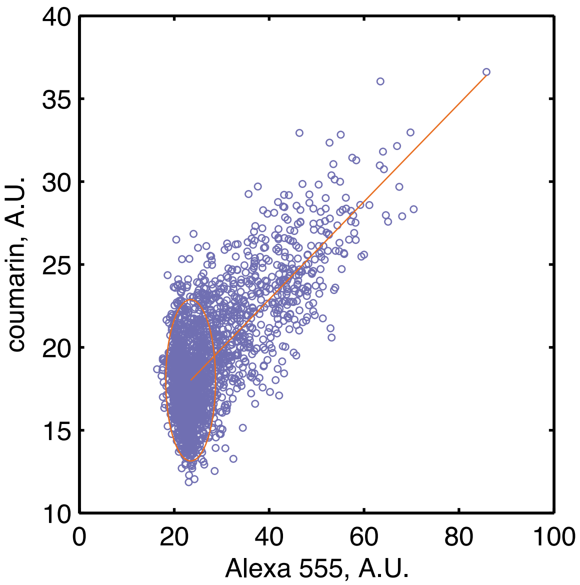
# Supplementary Information

*Table of Contents*

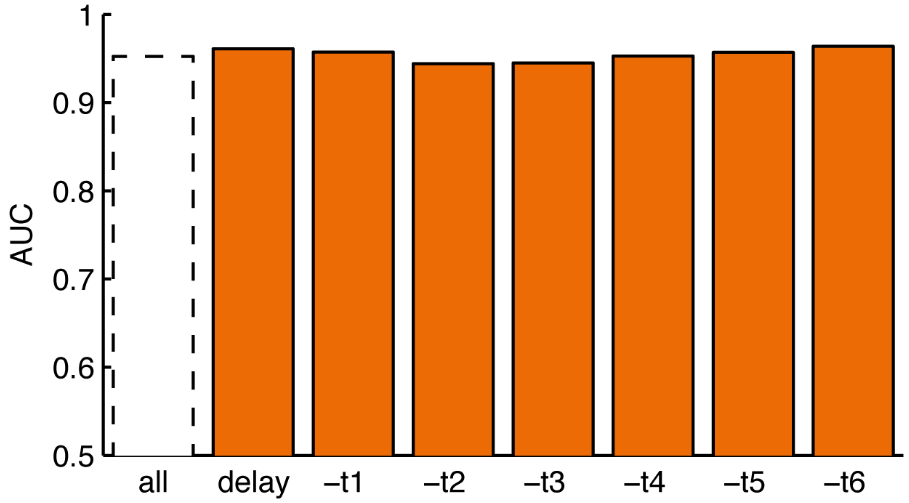
## Supplementary Figure 1. Linearity of tyramide amplification protocol



In order to improve the signal to noise ratio in the mRNA *in situ* hybridizations, we use a tyramide amplification reaction that amplifies the signal of the antibodies used to detect the RNA probes (Luengo Hendriks et al., 2006). To compare tyramide amplified signal to directly labeled antibodies in single embryos and ensure the linearity of this amplification step, we hybridized a lacZ-DNP labeled probe to one of the transgenic reporter lines. We then detected the DNP probes using an Alexa 555-labeled anti-DNP-HRP antibody. We conducted the tyramide amplification using coumarin. We could thus compare the Alexa 555 and coumarin signals directly. We stained nuclei using SYTOX Green and imaged the embryos using our standard protocol. We then processed the image stacks into PointClouds to segment the nuclei are segmented and un-mix the channels. Because the Alexa 555 signal becomes increasingly difficult to detect with embryo depth, we compared the coumarin and Alexa 555 signals in the top third of each embryo. We identified the “on” cells—those expressing the lacZ reporter—by thresholding using the same mode + 1 standard deviation metric described in Equation (7) of the main text. We then fit a line through these “on” cells.

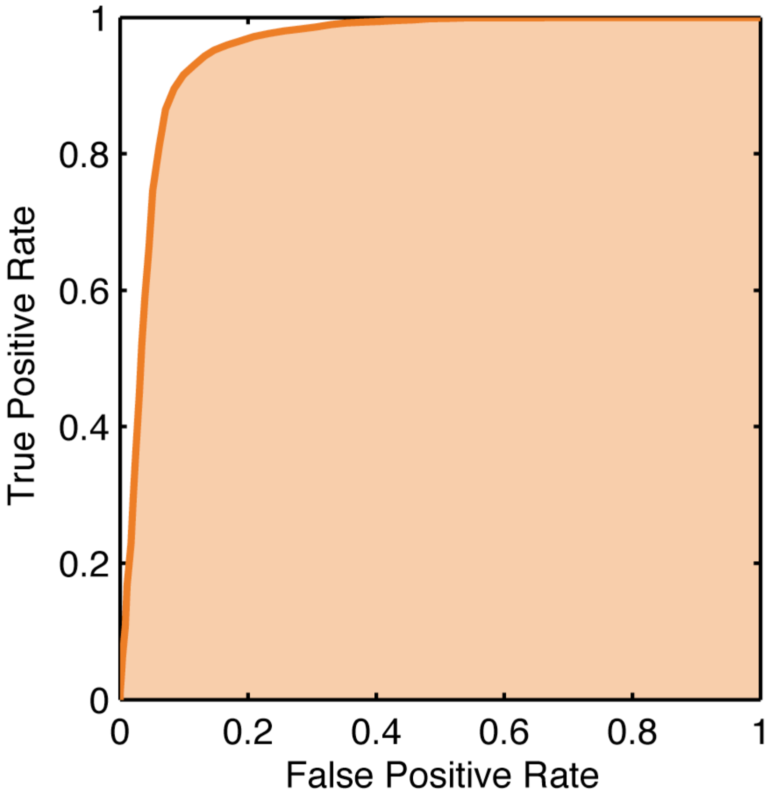
Though this experiment is not a perfect test of the linearity of the signal to RNA concentration, since we have not measured the linearity of the anti-DNP-HRP antibody, it demonstrates that a saturation of the amplification reaction is unlikely. Based on the inspection of plots for 10 embryos, we found no evidence of a bend in the plot that would be expected if the amplification reaction was reaching saturation. A plot for a typical embryo is shown above. The circle encloses the points corresponding to the “off” cells, and the line is the linear fit through the “on” cells. In addition, as can be seen in Figure S1 of (Fowlkes et al., 2011), when histograms are plotted to show the distribution of intensity values found in the gene expression atlas, there is no peak at the high end of the spectrum, as would be expected if the amplification reaction was progressing until saturation.

**Supplementary Figure 2. Fitting subsections of the *dmel* endogenous *hb* pattern**



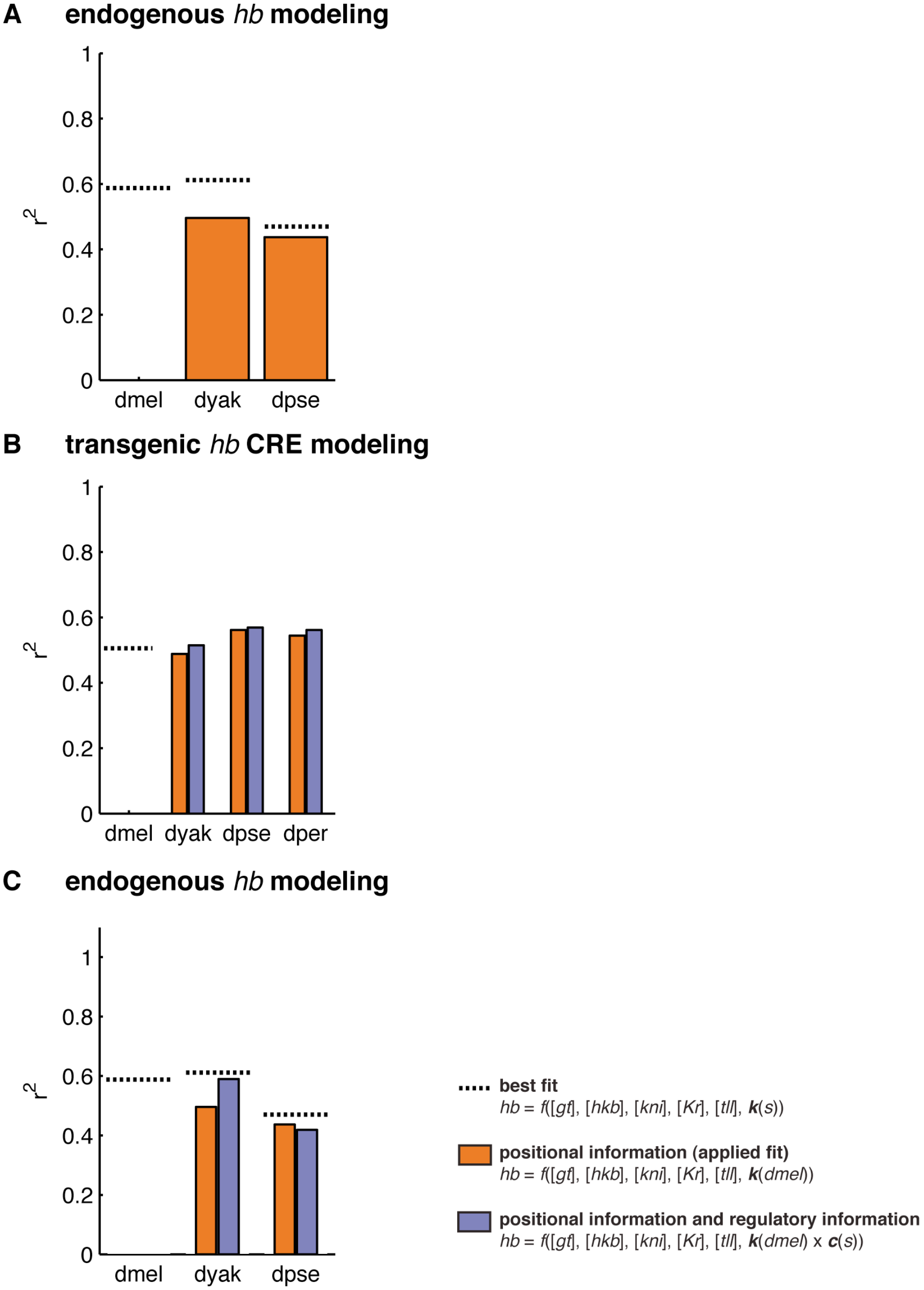
Here we show the results of fitting a linear model to the posterior 36% of the *dmel* endogenous *hb* expression data. The dotted bar gives the AUC resulting from a model that includes all the data points; in this model the inputs of each time point are used to fit the output from the same time point. The “delay” bar gives the AUC resulting from using the inputs from one time point to fit the output of the subsequent time point. The remaining bars show the results of excluding the data points from one time point at a time. The exclusion of some time points, which correspond to a similar reduction in the size of the dataset as the “delay” model, give similar or better results than the time delay.

**Supplementary Figure 3. Calculation of the ROC AUC**



Here we show a receiver operating characteristic (ROC) curve for the *dmel* best fit shown in Figure 3A. The shaded area below the curve corresponds to the area under the curve (AUC). The ROC curve shows the tradeoff between the true positive rate, the fraction of experimentally-determined “on” cells that are correctly predicted, and the false positive rate, the fraction of experimentally-determined “off” cells that are incorrectly predicted. As the cutoff that separates “on” and “off” predictions is loosened, both the true and false positive rates increase. The AUC summarizes the entire curve as a single number. A random classifier has an AUC of 0.5, and a perfect classifier has an AUC of 1. This ROC curve has an AUC of 0.95, indicating that a high true positive rate can be achieved simultaneously with a small false positive rate.

## Supplementary Figure 4. Model results using $r^2$



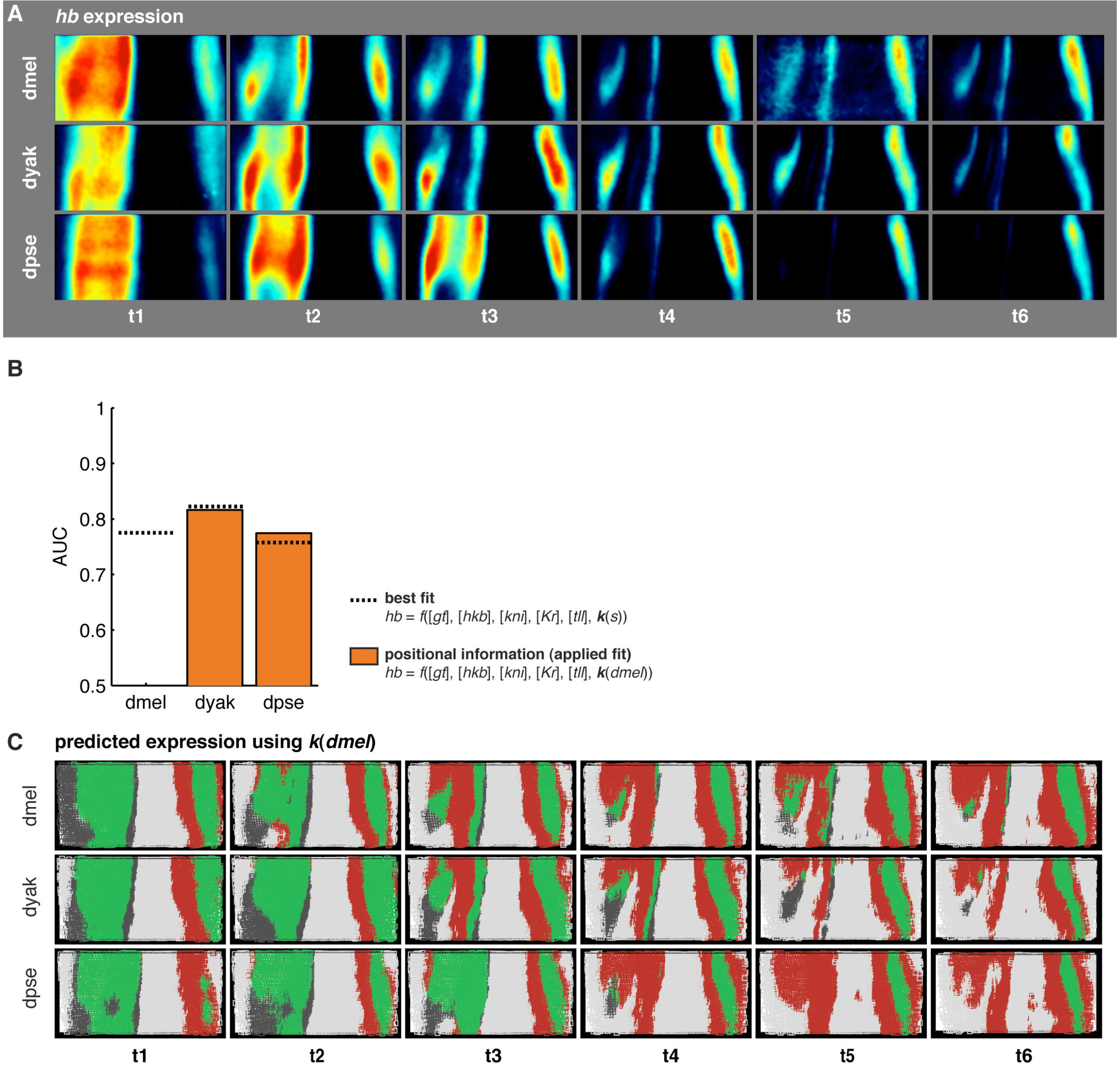
A endogenous *hb* modeling

B transgenic *hb* CRE modeling

C endogenous *hb* modeling

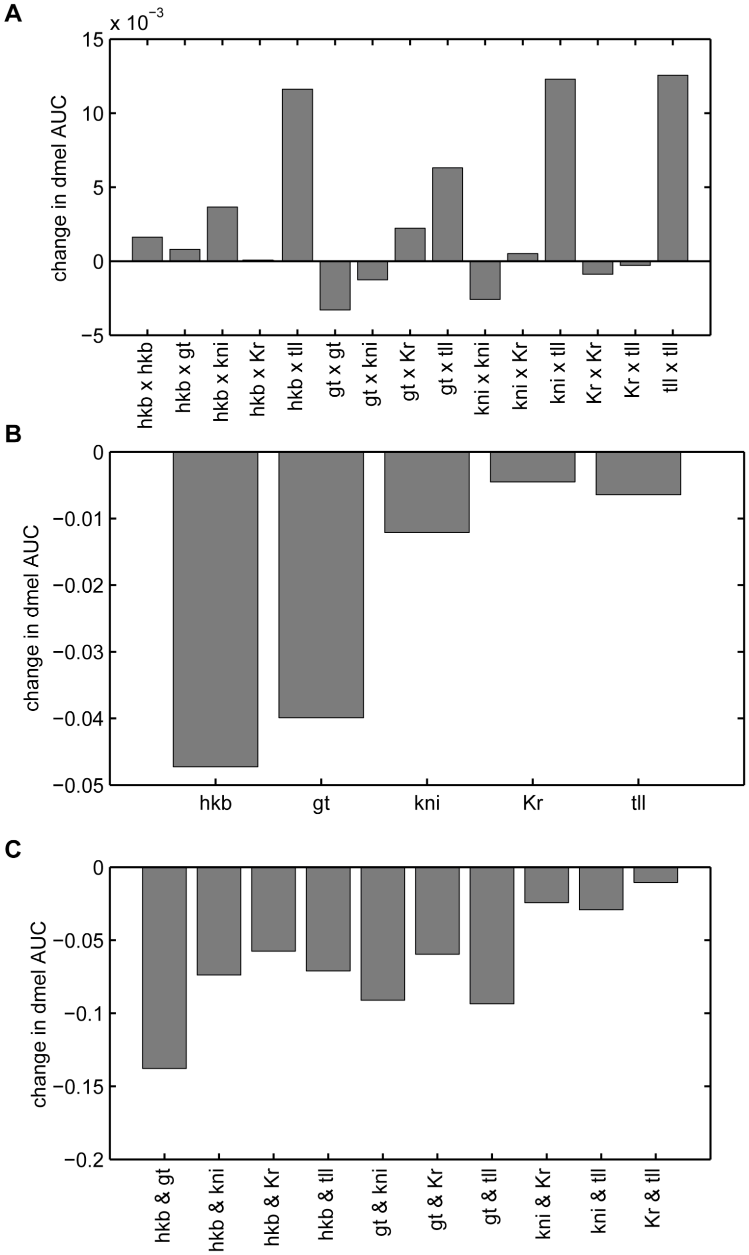
Here we show the modeling results using an $r^2$ score, instead of the ROC AUC. (A) corresponds to Figure 3A, endogenous *hb* modeling, (B) is Figure 5A, transgenic *lacZ* modeling, and (C) is Figure 6, endogenous *hb* modeling. The dotted line corresponds to using a species-specific ***k***(*s*) parameter vector, the orange bars show the results using ***k***(*dmel*) (positional information), and the purple bars show the results using ***k***(*dmel*) and a sequence weight (regulatory logic).

## Supplementary Figure 5. The whole embryo endogenous *hb* pattern



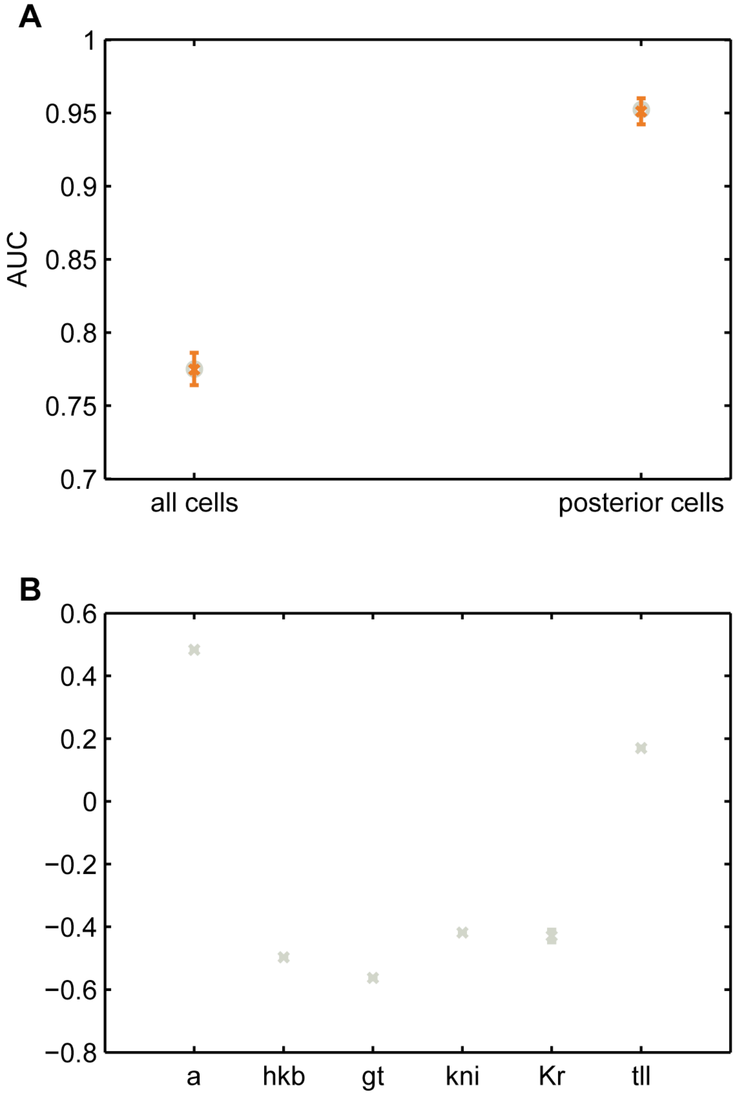
In panel (A), we show the endogenous expression pattern of *hb,* as in Figure 2, for the whole embryo, and in panel (B), we show the results of applying a multiple linear regression to the whole embryo. The performance of the model is substantially worse on the whole embryo, which is not surprising, since the anterior pattern is thought to be controlled by a different CRE using different regulators than the posterior stripe. In panel (C), we show the detailed cell-by-cell performance of the linear model using ***k***(*dmel*), using the same color code as in Figure 3.

## Supplementary Figure 6. Modifications to the model do not greatly improve performance in the native context



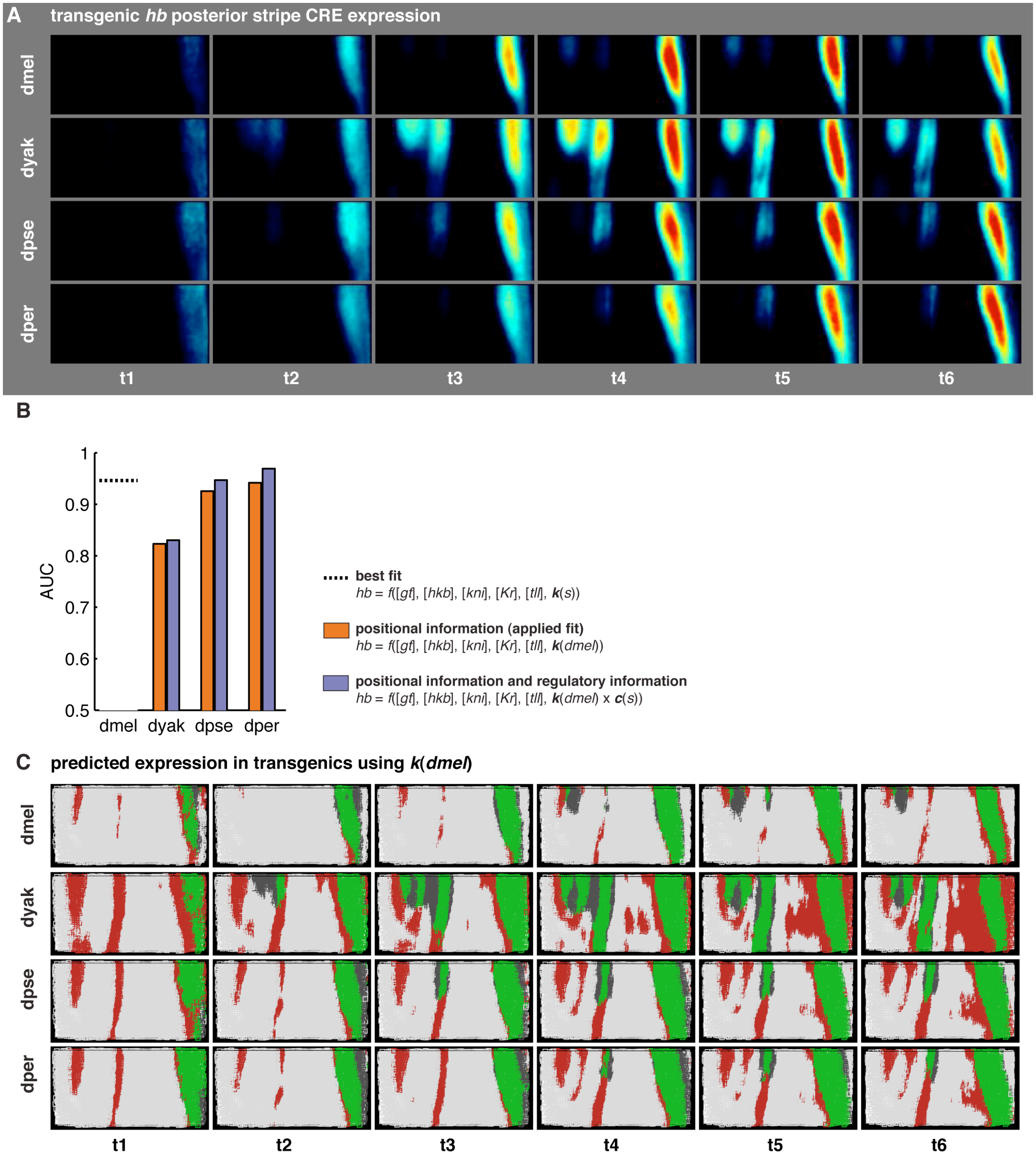
In panel (A), we show the change in the performance of the model on the endogenous *dmel* data set when a cross-term is added, as compared to the model without any cross-terms. The cross-terms are modestly affect the performance of the model. In panels (B) and (C), we show the change in the performance of the model on the *dmel* data set when a regulator or two regulators are dropped from the model.

**Supplementary Figure 7. The linear regression model for endogenous expression is robust to cross-validation**



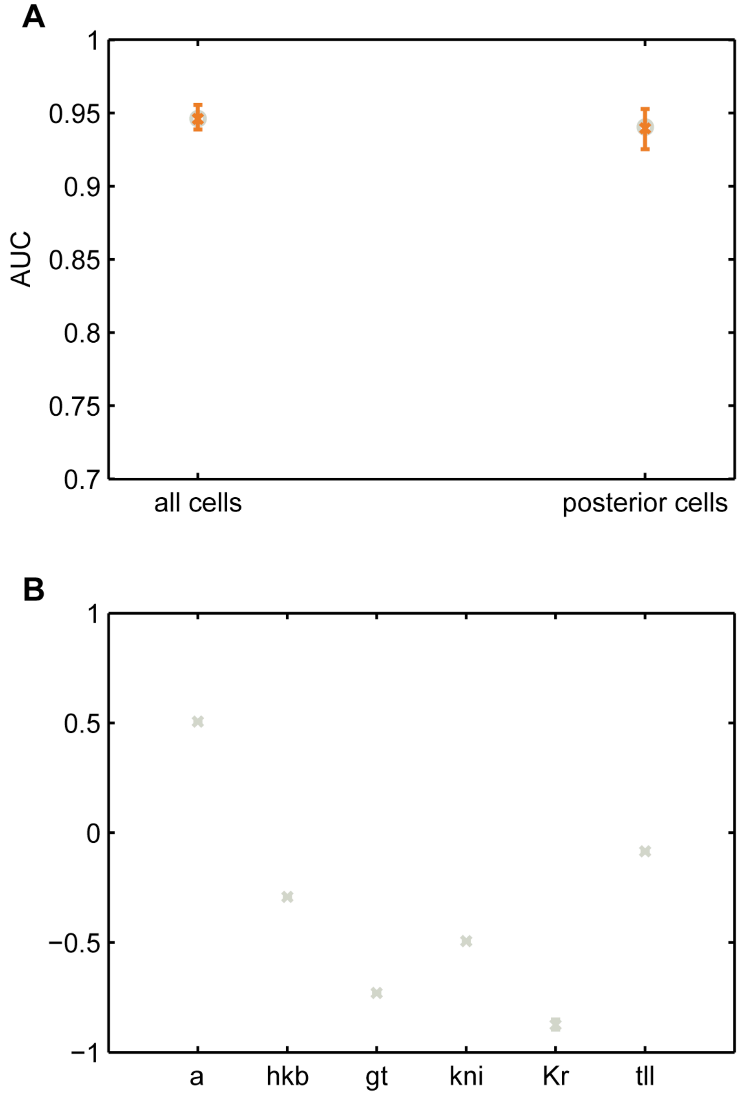
In panel (A), we show the results of a 10-fold cross validation exercise, in which 9/10 of the *dmel* endogenous data is used to train the multiple linear regression model and 1/10 of the data set is used to evaluate the model. We plot the AUC for all the *dmel* cells and just the posterior cells in grey, and the cross-validation results in orange. The orange 'x' shows the mean AUC, and the error bars show the minimum and maximum AUCs for the 10 folds of the data. In panel (B), we show the mean, minimum and maximum values of the coefficients *a* and ***k*** resulting from the cross-validation. Both the AUC and the values of these coefficients vary very little in the cross-validation, indicating that the model is not over-fit.

## Supplementary Figure 8. The whole embryo transgenic *hb* posterior stripe CRE pattern



In panel (A), we show the expression pattern of *lacZ* reporter, as in Figure 4, for the whole embryo. In panel (B), we show the results of applying a multiple linear regression to the whole embryo. The performance of the model is worse for the *dyak* whole embryo pattern, but the addition of the sequence weights is once again useful in the transgenic case for prediction purposes. In panel (C), we show the detailed performance of the linear model, without any sequence weights, using the same color code as in Figure 3.

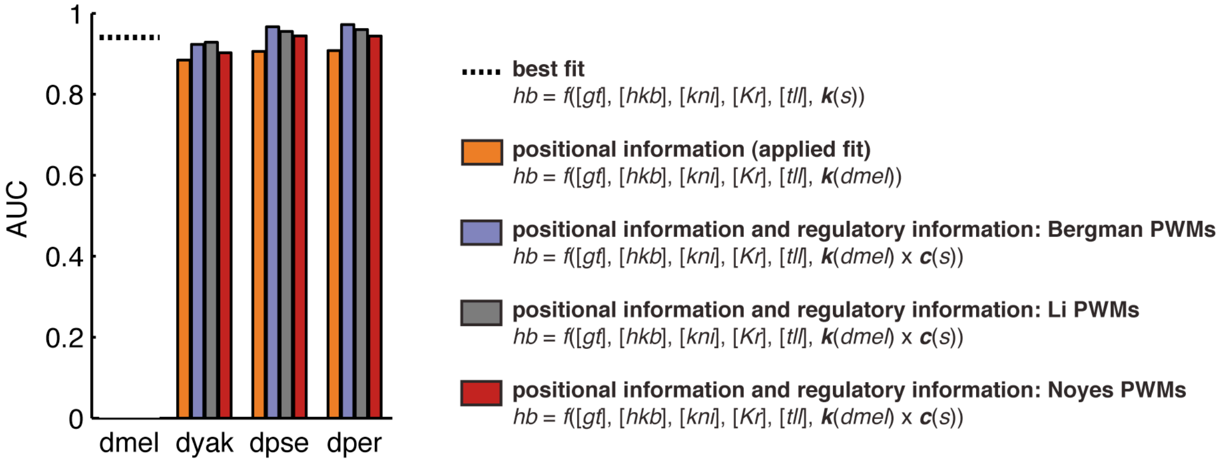
**Supplementary Figure 9. The linear regression model for transgenic expression is robust to cross-validation**



In panel (A), we show the results of a 10-fold cross validation exercise, in which 9/10 of the *dmel* transgenic data is used to train the multiple linear regression model and 1/10 of the data set is used to evaluate the model. We plot the AUC for all the *dmel* cells and just the posterior cells in grey, and the cross-validation results in orange. The orange 'x' shows the mean AUC, and the error bars show the minimum and maximum AUCs for the 10 folds of the data. In panel (B), we show the mean, minimum and maximum values of the coefficients *a* and ***k*** resulting from the cross-validation. Both the AUC and the values of these coefficients vary very little in the cross-validation, indicating that the model is not over-fit.

**Supplementary Figure 10. The sequence weight improves the model predictions for the transgenic lines, independent of PWM selection**

**transgenic *hb* CRE modeling**



To compare the performance of different position weight matrices (PWMs), we calculated the sequence weight using the PWMs from (Bergman et al., 2005), (Li et al., 2011) and (Noyes et al., 2008). We added a pseudocount of 0.01 to the PWMs from (Li et al., 2011) and a pseudocount of 1 to the PWMs from (Noyes et al., 2008). There was no PWM for *hkb* in (Li et al., 2011), so we used the one from (Bergman et al., 2005) instead. Here we plot the resulting AUCs from using each set of PWMs to predict the expression patterns in the transgenic lines carrying the *dyak, dpse,* and *dper hb* posterior stripe CREs. All three sets of PWMs yield sequence weights that improve the performance of the model as compared to the performance without the sequence weights (orange bars). We chose to use the PWMs from (Bergman et al., 2005), as they gave the best performance on average.

## The Effect of Measurement Non-linearities on the Model

Gathering high-quality expression data in intact animals is still a technical challenge. As we show in Supplementary Figure 1, the amplification step we use to increase our signal-to-noise ratio shows no evidence of saturation. However, since we have not directly measured the linearity of the anti-DNP-HRP antibody itself, this experiment is not a definitive demonstration of the linear relationship between measured signal and RNA level. Non-linearities would likely underestimate the amount of RNA in the most highly expressing cells.

One of the advantages of using the ROC AUC measurement is that it is relatively insensitive to non-linearities in the input and output measurements of the circuit under study. A non-linearity in the output *hb* measurement would cause us to underestimate the amount of mRNA in the most highly expressing cells. However, when we evaluate the predictions, we threshold the measured *hb* levels into “on” and “off” cells using the mode + 1 standard deviation (Equation (7), Materials and Methods). Therefore, so long as this measurement still places these cells in the “on” category, the AUC would remain unchanged. In the case where the non-linearity is in the inputs, we consider activators and repressors separately. In cells where a repressor is highly expressed, even with an underestimation of the true level, the predictions likely already put the cell in the “off” category. In our circuit, there is only one weak activator, *tll*. An underestimation of its peak levels may cause a handful of cells to be predicted to be “off” instead of “on.” However, given its small contribution (Table I) and the fact that the false negative cells do not overlap *tll*’s expression domain (dark grey cells in Figure 3B), it is unlikely that this effect is at play in our system.

## Estimating the Effect of Measurement Error on the AUC

In order to understand the magnitude of a significant change in the AUC, we calculate how error would propagate to the *hb* predictions using the formula:

$$\sigma_{pred}^2(s)=\left(k_s^{gt}\right)^2\cdot\sigma_{gt}^2(s)+\left(k_s^{hkb}\right)^2\cdot\sigma_{hkb}^2(s)+\left(k_s^{kni}\right)^2\cdot\sigma_{kni}^2(s)+\left(k_s^{Kr}\right)^2\cdot\sigma_{Kr}^2(s)+\left(k_s^{tll}\right)^2\cdot\sigma_{tll}^2(s)$$

Here, $\sigma_{pred}^2(s)$ is the estimated error in the *hb* predictions for species *s*, $\sigma_{gt}(s)$ is the standard deviation for *gt* in Table S2 of (Fowlkes et al., 2011), and $k_s^{gt}$ is the coefficient of the linear model as reported in Table I of the main text. Since expression values for each gene are measured in separate experiments, we assume the covariation in measurement error between genes is zero. This formula yielded estimated average variance of the predicted *hb* expression level of 0.0091, 0.0134, and 0.0064 for *dmel*, *dyak*, and *dpse*, respectively. This indicates that *dpse* predictions have the smallest amount of uncertainty, but that in all species, the level of uncertainty is small compared to the level of *hb* predicted in “on” cells.

To estimate the effect of these errors on the AUCs, we added a normally distributed error term (mean = 0, variance as estimated above) to our *hb* predictions and calculated the resulting AUC. We repeated this calculation 1000 times, and we found the standard deviation of the AUCs was 0.0021, 0.0022, and 0.0026, for *dmel*, *dyak*, and *dpse*, respectively, values that are smaller than the differences on which we base our conclusions.

**Calculation of Sequence Weight**

We use an approximation based on statistical mechanics in order to calculate the sequence weight. The probability of a particular site *k* being bound by one transcription factor is

$$p(bound) = \frac{e^{-\Delta G_k}}{Z}$$

Here $\Delta G_k$ is the binding energy of the site, in units of thermal energy (kT), and *Z* is the partition function. We assume that this transcription factor is either bound to this particular site or another site in the genome, therefore

$$Z = e^{-\Delta G_k} + \sum_{i \neq k} e^{-\Delta G_i}$$

we then write

$$\begin{aligned} p(bound) &= \frac{e^{-\Delta G_k}}{e^{-\Delta G_k} + \sum_{i \neq k} e^{-\Delta G_i}} \bullet \frac{e^{\Delta G_{ref}}}{e^{\Delta G_{ref}}} \\ &= \frac{e^{-\left(\Delta G_k - \Delta G_{ref}\right)}}{e^{-\left(\Delta G_k - \Delta G_{ref}\right)} + \sum_{i \neq k} e^{-\left(\Delta G_i - \Delta G_{ref}\right)}} \\ &\approx \frac{e^{-\left(\Delta G_k - \Delta G_{ref}\right)}}{N} \end{aligned}$$

As in (Fields et al., 1997), we define $\left(\Delta G_k - \Delta G_{ref}\right)$ as the specific binding energy of the TF to a sequence such that $\sum_{i \neq k} e^{-\left(\Delta G_i - \Delta G_{ref}\right)} = N$ , where *N* is the genome size. We assume $N \gg e^{-\left(\Delta G_k - \Delta G_{ref}\right)}$. To estimate $\Delta G_k - \Delta G_{ref} = \Delta\Delta G_k$ using a position weight matrix, we assume that each base pair of the site contributes to this energy independently and therefore

$$-\Delta\Delta G_k \propto \sum_{i=1}^{w} \ln \frac{p_i\left(b(i)\right)}{q\left(b(i)\right)},$$

where $p_i\left(b(i)\right)$ is the frequency of observing the base *b*(*i*) at position *i* of the binding site, $q\left(b(i)\right)$ is the background frequency of base *b*(*i*), and *w* is the length of the binding site.

$$p(bound) \propto \frac{e^{\sum_{i=1}^{w} \ln \frac{p_i(b(i))}{q(b(i))}}}{N} = \frac{\prod_{i=1}^{w} e^{\ln \frac{p_i(b(i))}{q(b(i))}}}{N} = \frac{\prod_{i=1}^{w} \frac{p_i\left(b(i)\right)}{q\left(b(i)\right)}}{N}$$

To get an estimate of the total binding capacity of a stretch of sequence of length *l*, we assume the genome sizes of the different species *s* are approximately equal and calculate the sequence weight

$$c(s) = \frac{\sum_{i=1}^{l-w+1} \prod_{j=1}^{w} \frac{p_i(b(i))}{q(b(i))}}{c(dmel)}$$

In our modeling efforts, we multiply the sequence weight for a particular species and TF by the relative concentration of the TF in each cell and the fitted parameter *k*(*dmel*).

To calculate the likelihood a particular binding arrangement of TFs requires a more detailed calculation which takes into account the number of TF molecules in the system and effects like the physical occlusion of TFs binding overlapping sites and cooperative TF binding. Previous studies have implemented these types of models (Segal et al., 2008; He et al., 2010), and these models require fitting a number of parameters to describe the TF molecule numbers and effects mentioned above.

Another effect we neglect in the sequence weight is the saturation of sites: beyond a certain concentration of a TF, the site is bound with very high probability, and therefore our multiplication of the sequence weight with the relative TF concentration overemphasizes the binding capacity of a strong site. Since we do not know the absolute concentrations of TFs or the absolute binding energies of binding sites, we cannot find saturated sites analytically. Instead, to look for sites likely to be saturated, we searched for binding sites that accounted for 50% or more of the total value of the sequence weight and found a single *hkb* site in the *dyak* posterior stripe CRE that fit this criteria. We thresholded its value to its 99th percentile value and found that the thresholded sequence weight performed better than the un-thresholded sequence weight.

## References

Bergman C, Carlson JW, Celniker SE (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* **21:** 1747-1749

Fields DS, He Y, Al-Uzri AY, Stormo GD (1997) Quantitative specificity of the Mnt repressor. *J Mol Biol* **271:** 178-194

Fowlkes CC, Eckenrode KB, Bragdon MD, Meyer M, Wunderlich Z, Simirenko L, Luengo Hendriks CL, Keranen SV, Henriquez C, Knowles DW, Biggin MD, Eisen MB, Depace AH (2011) A conserved developmental patterning network produces quantitatively different output in multiple species of Drosophila. *PLoS Genet* **7:** e1002346

He X, Samee MA, Blatti C, Sinha S (2010) Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol* **6:** e1000935

Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol* **12:** R34

Luengo Hendriks CL, Keränen Fowlkes Simirenko L, Weber DePace Henriquez Kaszuba Hamann Eisen MB, Malik Sudar Biggin MD, Knowles DW (2006) Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. *Genome Biol* **7:** R123

Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic acids research* **36:** 2547-2560

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451:** 535-540