nature genetics

Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster even skipped* gene

Hilde Janssens[1], Shuling Hou[2], Johannes Jaeger[1], Ah-Ram Kim[1], Ekaterina Myasnikova[3], David Sharp[4] & John Reinitz[1]

Here we present a quantitative and predictive model of the transcriptional readout of the proximal 1.7 kb of the control region of the *Drosophila melanogaster* gene *even skipped* (*eve*). The model is based on the positions and sequence of individual binding sites on the DNA and quantitative, time-resolved expression data at cellular resolution. These data demonstrated new expression features, first reported here. The model correctly predicts the expression patterns of mutations in *trans*, as well as point mutations, insertions and deletions in *cis*. It also shows that the nonclassical expression of stripe 7 driven by this fragment is activated by the protein Caudal (Cad), and repressed by the proteins Tailless (Tll) and Giant (Gt).

The fundamental principles of transcriptional control elucidated in *Escherichia coli* by Jacob and Monod are based on the idea that one or a small number of binding sites for regulatory proteins has a distinct biological function. In many metazoan genes, control regions ('promoters') contain many kilobases (kb) of DNA, and the nearly one-to-one relationship between binding site and function is lost. This is a fundamental problem because binding sites can be found directly from sequence if enough examples from chemical experiments are known, whereas difficult *in vivo* experiments are required to assay function. What is needed to solve this problem are methods of determining the transcriptional readout of large segments of DNA containing many binding sites, no single one of which controls a phenotypic function.

A central organizing idea in metazoan molecular genetics is the '*cis*-regulatory module', or CRM (also known as an enhancer)[1]. In certain cases, it has been shown that complex expression patterns can be decomposed into simple components, each acting independently and each controlled by a short, contiguous segment of DNA[2] that contains clustered sites for transcription factors[3–7]. This constitutes a classic, but incomplete, picture of CRM function. There is no inherent reason that CRMs must lie on a contiguous segment of DNA or that they must contain clusters of sites. These two properties are instead artifacts of the experimental and informatic methods used to identify CRMs, and there is experimental information that cannot be understood on the basis of this conventional picture of CRM structure. A more fundamental understanding of CRM structure and function is of great importance, because it would permit the prediction of gene expression directly from sequence.

Gaining this new level of understanding requires a method of determining the physiological consequences of a ligand bound to a particular binding site in an environment of other bound factors. The desired method must embody the fundamental principles from which CRMs are constructed, particularly the rule or rules responsible for the functional independence of CRMs. A proposal that the functional independence of CRMs is a consequence of short-range repression is of central importance. This proposal is based on the fact that activators of transcription are operative at large distances from the basal promoter in the absence of intervening insulating elements, but that this activation can be suppressed ('quenched') by certain repressors bound within 100–200 bp of the activator site[8]. This rule is compatible with the existence of clusters, as a cluster would correspond to a group of binding sites for activators and associated quenchers, with the latter able to repress activators bound within the given cluster but not distant ones. It is evident that such a rule can determine expression patterns controlled from sites outside a cluster as well. Here we consider a situation where gene expression is driven from outside classical CRMs and demonstrate that a predictive understanding of gene expression controlled by these sequences can be obtained by a quantitative model based on the rule stated above.

The control region of the gene *eve* of *D. melanogaster* provides excellent examples of experimental data that can and cannot be understood in terms of the classic picture of CRMs. Two segments of DNA that drive transcription in the early embryo have been identified on the 5′ side of the *eve* transcription unit[9,10]. One of these segments is the smallest contiguous segment of DNA that drives expression of *eve* stripe 2, which is known as minimal stripe element 2 (MSE2)[11]. A second segment was originally identified as the smallest contiguous element that drives *eve* stripe 3, and hence was called minimal stripe element 3 (MSE3)[12]. Because this latter DNA segment

[1]Department of Applied Mathematics and Statistics, and Center for Developmental Genetics, Stony Brook University, Stony Brook, New York 11794-3600, USA. [2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [3]Department of Computational Biology, Center of Advanced Studies, St. Petersburg State Polytechnic University, St. Petersburg 195251, Russia. [4]Chief Science Office, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. Correspondence should be addressed to J.R. (reinitz@odd.bio.sunysb.edu).

Received 28 April; accepted 17 August; published online 17 September 2006; doi:10.1038/ng1886

also expresses *eve* stripe 7 at the same level as stripe 3, it is now usually referred to as the 3/7 enhancer[13]. In parallel with these experimental studies, informatics investigations of DNA sequence have revealed a correlation between the presence of a CRM and the identification of clusters of binding sites. Both MSE2 and MSE3 show strong statistical evidence of such clusters[3,6].

Other properties of MSE2 and MSE3 are not consistent with this picture of CRMs. Stripe 7 is not fully under the control of MSE3, because 1.7 kb of *eve* upstream regulatory sequences can drive stripe 7 expression at low levels[10,11] and short deletions in the 5′ *eve* control region can eliminate stripe 3, but not stripe 7 (ref. 9). Moreover, in a construct containing the native *eve* gene without MSE2, there is still residual expression of stripe 2 (ref. 14). These results demonstrate that the formation of native *eve* stripes 2 and 7 cannot be understood in terms of MSE2 and MSE3 alone and that stripe 7 does not have a minimal stripe element. It is reasonable to suppose that expression of stripe 7 controlled by the native gene is in part driven by these extra-MSE3 sequences.

Here we apply a previously described model of CRM function[15] to the problem of *eve* stripe 7 expression driven by sequences outside of the 3/7 enhancer. We first present quantitative expression data on the expression of p1.7*eve-lacZ*, a construct in which the proximal 1.7 kb of *eve* control DNA drives *lacZ* expression. These quantitative data reveal several novel features of the expression pattern that were overlooked in previous qualitative work. We next apply our model to the quantitative data, and demonstrate that an internally self-consistent model of p1.7*eve-lacZ* expression can be constructed based on expression data and ligand-binding sites predicted from sequence with positional weight matrices (PWMs). Using the specific model we have constructed, we demonstrate that stripe 7 expression in p1.7*eve-lacZ* is a consequence of widespread activation and localized repression. The model correctly predicts the expression pattern of separated fragments of a CRM, and can treat the situation where two pieces of DNA, each of which cannot drive any expression separately, can drive strong spatially localized expression when fused. The model also confirms the classical picture of MSE2 regulation[11,16]. Analysis of the model shows that activation is supplied by Cad protein, whereas the limits of stripe 7 expression from p1.7*eve-lacZ* are set by Tll on the posterior and Gt on the anterior.

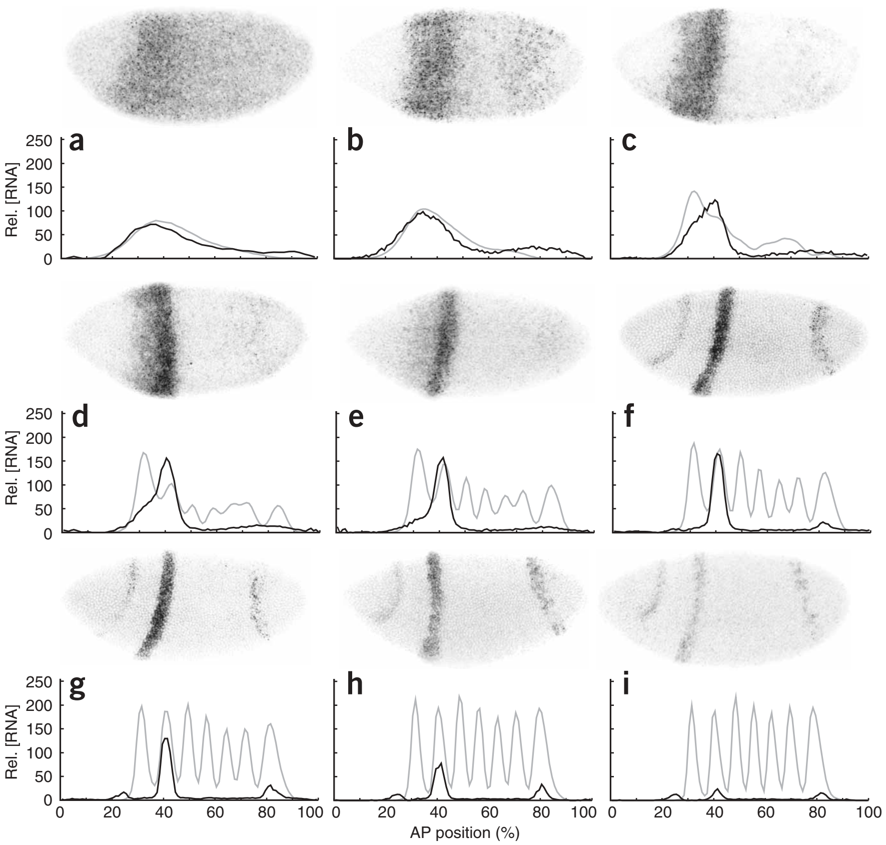

Figure 1 Dynamic quantitative expression of a *lacZ* reporter construct. *lacZ* (black) driven by 1.7 kb of *eve* upstream regulatory sequences at nine different times: (**a**) C13, (**b**) T1, (**c**) T2, (**d**) T3, (**e**) T4, (**f**) T5, (**g**) T6, (**h**) T7 and (**i**) T8. The corresponding endogenous Eve protein pattern (gray) is shown for comparison. Expression levels are in relative units based on fluorescence, and AP position is in units of percentage increasing toward the posterior. For each time point, one embryo stained for *lacZ* mRNA is shown above the corresponding averaged expression profile at that specific time. All embryos presented are oriented with anterior to the left and dorsal up. The expression seen at around 25% AP from T5 on is due to vector sequences in the P-transposon[11].

RESULTS

Quantitative gene expression data

We quantitatively monitored the expression of RNA in the blastoderm from a *lacZ* reporter gene under the control of 1.7 kb of DNA 5′ to the transcription start site of the *D. melanogaster eve* gene.

Our data reveal several new quantitative features of the p1.7*eve-lacZ* reporter construct (**Fig. 1**). First, the early expression pattern closely approximates the early expression profile of the entire *eve* gene. Second, stripe 2 expression from this construct peaks during time class T5 in cycle 14A (C14A) and then declines. Third, this construct also has a broad posterior expression domain that forms at T1 and gradually restricts to the position of *eve* stripe 7 by T5. Fourth, stripe 7 of the reporter is most strongly expressed at T6-T7, with a maximum expression level that corresponds to approximately 20% of the maximum of stripe 2 (**Fig. 1**). Fifth, stripe 2 expression driven by MSE2 with no proximal sequences reaches a maximum at T6 and does not decline (data not shown).

Model results

Our model of transcriptional regulation has been described in detail elsewhere[15]. An overview of the model Equations is given in **Box 1**, and a brief description of the basic modeling ideas is given below.

We imagine that transcription initiation is an enzymatic process catalyzed by adapter factors that lower an activation energy barrier by an amount proportional to the number of activators present. We describe the effects of activation energy by an Arrhenius rate law, which will cause an exponential increase in transcription rate (up to a predefined maximum) as more adapters are recruited. Adapter factors are recruited by bound activators, which can act at long range. We imagine that $1/C_a$ activators are required to recruit an adapter, so that

BOX 1 THE MODEL EQUATIONS

Fractional occupancy of DNA-binding factors

$$f_{i[m_i,n_i;a]} = \frac{K_{i[m_i,n_i;a]}\nu^a}{1+K_{i[m_i,n_i;a]}\nu^a} \quad [1]$$

$$f_{i[m_i,n_i;a]} = \frac{K_{i[m_i,n_i;a]}\nu^a}{1+K_{i[m_i,n_i;a]}\nu^a + K_{j[m_j,n_j;b]}\nu^b} \quad [2]$$

Quenching

$$F^A_{i[m_i,n_i;a]} = f^A_{i[m_i,n_i;a]}\prod_k (1-q(d_k)E_b f^Q_{k[m_k,n_k;b]}) \quad [3]$$

Recruitment

$$N = \sum_a C_a \sum_i F^A_{i[m_i,n_i;a]} \quad [4] \longrightarrow M = f^{AF}N \quad [5]$$

Transcription rate

$$[mRNA] \propto R = \begin{cases} R_0\exp[-(\Theta - QM)] & \text{iff } QM < \Theta \\ R_0 & \text{otherwise} \end{cases} \quad [6]$$

ν^a is the concentration of ligand a. The fractional occupancy $f_{i[m_i,n_i;a]}$ of site i between bases m_i and n_i, binding ligand a, is given by [1] without and [2] with the influence of overlapping binding sites, and is corrected by the quenching factor $1 - q(d_k)E_b f^Q_{k[m_k,n_k;b]}$ [3], in which d_k is the distance in bases from site i to site k and $q(d) = 1$ if $d < 100$ and $q(d) = 0$ if $d > 150$, with linear interpolation in between. E_b is a measure of quenching efficiency. $F^A_{i[m_i,n_i;a]}$ is the fractional occupancy of activators corrected for quenching, N is the number of binding sites for recruiting adapters, f^{AF} the fractional occupancy of these sites, M is the number actually bound, Θ is the height of the activation energy barrier ΔA in the absence of activation, Q is the amount that ΔA is reduced by each bound adaptor, and R_0 is the maximum transcription rate. C and K are explained in the text.

C_a is a measure of the activator's efficiency (**Box 1**). An activator can be prevented from acting in two ways: by competitive binding to an overlapping site; or by quenching, which takes place if quenchers are bound within range (~150 bp) of the activators. Each activator interacts with about six quenchers (**Supplementary Fig. 1** online). The equations for the action of quenchers are written in such a way that each bound quencher within range multiplies the amount of activator bound by a factor less than one, so that one quencher bound to its site does little, but many quenchers multiply the activator's activity by many factors, each less than one, reducing the activator's activity to near zero.

Binding of ligands, recruitment of adaptor factors by activators and quenching are the fundamental molecular interactions represented in the model. Concentrations of ligands (**Supplementary Fig. 2** online) are taken as input and the model gives the concentration of *lacZ* mRNA as output. It is evident from a comparison of **Figure 1b** and **Figure 1c** that *lacZ* mRNA has a half life of less than 6 min, which is short compared with the time scale of changes in gene expression, and hence the concentration of mRNA will be proportional to the transcription rate. The important consequences of this fact are, first, that the model parameters can be obtained by fitting to gene expression data and, second, that model predictions can be tested against *in situ* hybridization data. Specifically, once the model is fit to a particular DNA construct or constructs, it can be used to predict the expression pattern of any construct derived from those used in the fit by site-directed mutagenesis, deletion and even insertion if the binding sites on the inserted segment are known. For each ligand a (except for Gt; see Methods), we take the binding affinity $K_{i[m_i,n_i;a]}$ of the strongest site among all those that bind that ligand as a free parameter and constrain the other K values by treating the PWM score as a free energy of binding[17]. Thus, in this work each ligand a has two independent parameters associated with it: K_a and either C_a or E_a depending on whether it is an activator or quencher. Θ and R_0 are also free parameters, so a model with L ligands will have $2L + 2$ free parameters.

To use the model, a set of binding sites must be specified. Initial runs were done using 17 binding sites found by footprint experiments[11,16,18,19] (**Fig. 2**). These sites were not sufficient to produce the correct pattern (**Supplementary Figs. 3** and **4** online). There were two notable patterning defects. One was a lack of activation in the posterior part of the embryo from cleavage cycle 13 (C13) to T3 (**Fig. 2a**), and another was a displacement of stripe 7 toward the posterior (**Fig. 2b**).

In subsequent fits, additional binding sites and ligands were added (**Fig. 3**, **Supplementary Figs. 3** and **4** online and **Supplementary Table 1** online). These new sites were computationally predicted and were selected by a stepwise lowering of the significance threshold for PWM scores. We considered the quality of a solution to be improved if the root mean square score was decreased (**Supplementary Table 2** online) and if at least one of the following features of the data was reproduced more accurately: (i) activation in the posterior part of the

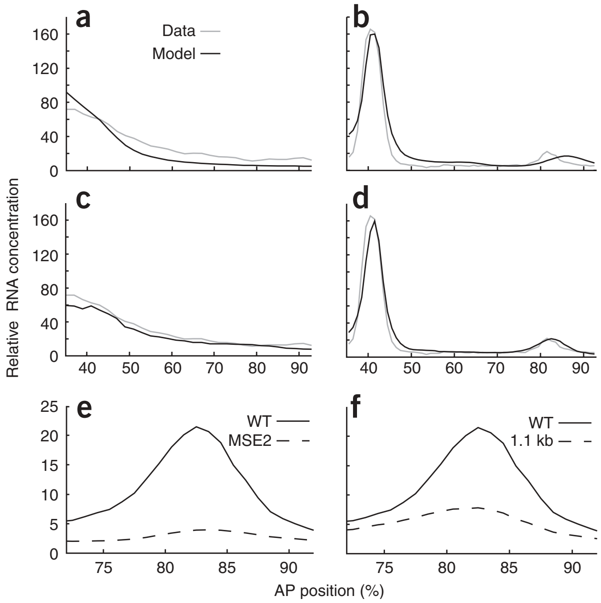

Figure 2 Summary of model output. Model output compared to data for the 17-site model (**a**,**b**) and the 34-site model (**c**–**f**), at C13 (**a**,**c**) and T5 (**b**,**d**–**f**). (**e**,**f**) Prediction of stripe 7 formation (dashed line), driven by MSE2 alone (**e**) or by the 1.1 kb of *eve* upstream regulatory sequences without MSE2 (**f**) compared with the 34-site model (WT). Axes are as in **Figure 1**. Models shown are p1.7_17_11 (17 sites) and p1.7_17_51 (34 sites).

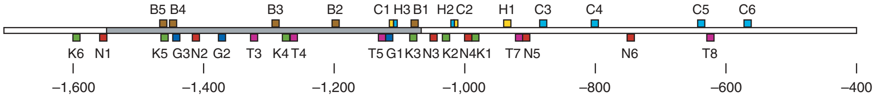

Figure 3 Schematic view of 1.7 kb of *eve* 5′ regulatory sequences. The binding sites shown are the sites used in the 34-site models p1.7_17_51-55. Binding sites for activators are shown above, sites for repressors below the sequence. The MSE2 region is indicated by a gray box. Names of each site with binding site coordinates can be found in **Supplementary Table 1** online. The 5′ side of the sequence is to the left.

embryo from C13 to T3; (ii) position of the posterior border of stripe 7; (iii) level of activation of stripe 7; (iv) position and steepness of the posterior border of stripe 2; and (v) lack of additional expression around 64% anterior-posterior position (AP; 0% at anterior pole) in T3 to T6.

Adding Tll as a repressor to the model defines the posterior border of stripe 7 more accurately (**Fig. 4** and **Supplementary Fig. 3**), but the intensity of stripe 7 remains lower than in the data (**Fig. 4a,b**). In the 17-site model and the 22-site model including Tll, the main activating site for stripe 7 is Hunchback3 (Hb3) (**Fig. 4d,g,j**). *In silico* deletion of the Hb3 site decreases stripe 7 expression significantly (**Fig. 4j**). Quenching by Tll of the Hb3 site restricts the expression to the exact position of stripe 7 (**Fig. 4b**,**e**,**h**,**k**). This is confirmed by *in silico* deletion of the Tll5 site, which gives rise to more posterior expression (**Fig. 4k**). However, Hb cannot provide sufficient stripe 7 activation even if more binding sites are added (**Supplementary Fig. 3**). Sufficient activation requires Cad (**Fig. 4c**,**f**,**i**,**l** and **Supplementary Fig. 3**). The major contribution comes from the Cad1 and Cad6 sites (**Fig. 4f**), where quenching of the latter by the Tll8 site demarcates the posterior border (**Fig. 4i**).

Cad is also required for correct expression in the posterior part of the embryo from C13 to T3 (**Fig. 2a** and **Supplementary Fig. 5** online). However, in the 28-site model that includes Cad, Cad gives rise to additional expression around 64% AP, which is absent from the data (**Supplementary Figs. 3** and **6** online). This expression was decreased by adding Knirps (Kni) as a repressor to the model (**Supplementary Figs. 3** and **6**). The expression at 64% AP in the 17-site model is due to Bicoid (Bcd), mainly through the Bcd2 and Bcd5 sites, and is enhanced by Cad activation through the Cad1 and Cad6 sites. In the 34-site model, Kni quenches the Cad1 site by binding at the Kni3 and Kni4 sites.

In summary, the 34-site model includes Kruppel (Kr), Gt, Kni and Tll as repressors and Bcd, Hb and Cad as activators (**Fig. 3**). Adding more than 34 binding sites did not increase the quality of the patterns significantly (**Supplementary Fig. 3**). In the 34-site model, the regulatory mechanism for stripe 2 formation is in complete agreement with the classic experimental analysis[11] (**Supplementary Fig. 7** online): stripe 2 expression is mainly activated by Bcd and Hb. The anterior border is set by Gt and the posterior border is set by Kr.

The model shows predictive ability. The role of Kni in repressing p1.7*eve-lacZ* implies that this construct should show ectopic expression at about 64% AP in *kni* mutants (**Supplementary Fig. 2**). This prediction is correct (**Fig. 5**), and we are not aware of any previous observations of the expression pattern of p1.7*eve-lacZ* in this genotype. The correctness of the model can also be assessed by comparing its behavior with published data that were not used in fits. For example, the model predicts that the subset of 17 sites present in MSE2 gives rise to stripe 2, but not to stripe 7 (**Fig. 2e**), behavior in agreement with experimental data (unpublished data). Furthermore, the model also predicts that the remaining 17 sites present in the region between MSE2 and the basal promoter give rise to little stripe 7 expression (**Fig. 2f**). Previous experimental work, not used to fit this model, includes an extensive study in which various binding sites on MSE2 were mutagenized to inactivity, and in some cases new sites were added to compensate for the deleted sites[20]. For example, mutating the Hb3 site greatly reduces expression levels, which can then be somewhat restored by adding a new Bcd site and more

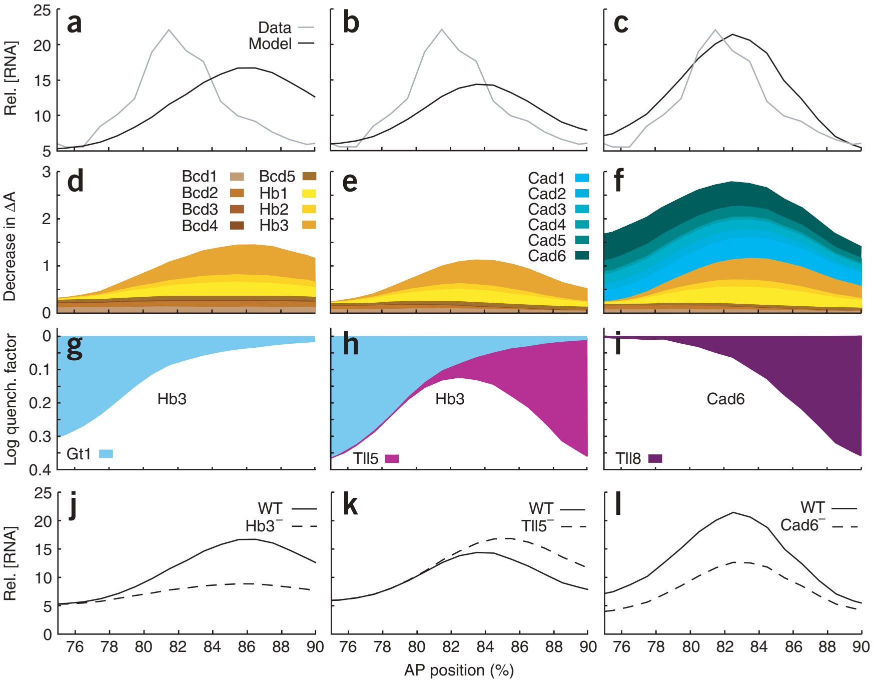

Figure 4 Regulatory analysis of stripe 7 expression. Analysis of the 17-site model (**a**,**d**,**g**,**j**), the 22-site model (**b**,**e**,**h**,**k**) and the 34-site model (**c**,**f**,**i**,**l**) at T5. (**a**–**c**) Model output in comparison with data. (**d**–**f**) Sum of activating contributions. Colored areas represent activating contributions by individual Bcd, Hb and Cad binding sites. The height of each colored area is given by $C_a F^A_{i[m_i,n_i;a]}$ ([4] in **Box 1**). Sum of quenching contributions to the Hb3 site in the 17-site model (**g**), the 22-site model (**h**) and the Cad6 site in the 34-site model (**i**) are shown. Colored areas represent quenching contributions by individual nearby binding sites. The height of each colored area is given by $\log(1 - q(d_k)E_b f^Q_{k[m_k,n_k;b]})$ ([3] in **Box 1**). (**j**–**l**) Prediction of stripe 7 formation when the Hb3 site is deleted from the 17-site model (**j**), the Tll5 site is deleted from the 22-site model (**k**) and the Cad6 site is deleted from the 34-site model (**l**) compared with nonmutated (WT) models. The *x* axis is as in **Fig. 1**. Models shown are p1.7_17_11 (17 sites), p1.7_17_66 (22 sites) and p1.7_17_51 (34 sites).

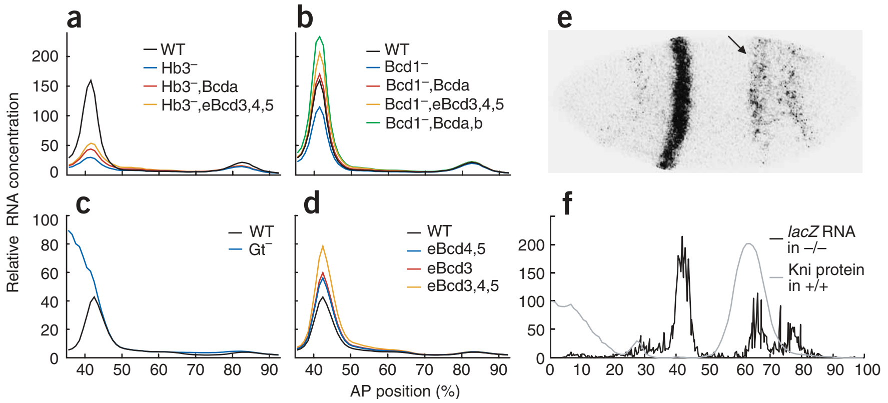

Figure 5 Model predictions. Relative RNA concentrations driven by the 34-site model (**a**,**b**) and by the sites present in MSE2 of the 34-site model (**c**,**d**) at T5. Wild-type model output in black. (**a**) Deletion of the Hb3 site (blue) decreases transcription. Addition of the Bcda site (red) or enhancement of the Bcd3, Bcd4 and Bcd5 sites (yellow) causes an increase in expression. (**b**) Deletion of the Bcd1 site (blue) decreases stripe 2 intensity, which is restored to wild-type levels by adding the Bcda site (red). Expression exceeds wild-type levels when sites Bcd3, Bcd4 and Bcd5 are enhanced (yellow) or when both sites Bcda and Bcdb (green) are added to a Bcd1 deficient CRM. (**c**) Deletion of all three Gt sites (blue) causes expansion of stripe 2 anteriorly. (**d**) Enhancing the Bcd3 site (red) causes a similar rise in stripe 2 intensity as enhancing the Bcd4 and Bcd5 sites (blue), while the enhancement of sites Bcd3, Bcd4 and Bcd5 (yellow) causes even stronger stripe 2 expression. Axes are as in **Figure 1**. Model shown is p1.7_17_51. (**e**) *lacZ* expression in a *kni* mutant embryo at T6; the arrow indicates ectopic expression. All 15 mutant embryos scanned expressed the ectopic stripe. (**f**) Quantitative *lacZ* expression profile of the embryo shown in **e**. The average wild-type Kni protein pattern (gray) is shown for comparison. Axes and embryo presentation are as in **Figure 1**.

strongly restored by enhancing the binding affinity of three existing Bcd sites. This and ten similar experiments were modeled (**Fig. 5a–d**). In each case where a particular mutation increased, decreased or spatially extended the domain, the model predicts the experiment correctly. It also predicts the rank order of expression intensity of pairs of compensating mutations correctly, although in one case the relative expression level of one such pair and wild type (Bcd1⁻ and eBcd3, eBcd4, eBcd5; Bcd1⁻ and Bcda) was incorrect.

DISCUSSION

We believe the results presented here constitute a qualitative and quantitative advance in understanding transcription. Although we took advantage of specific features of the *D. melanogaster* blastoderm system to perform this study, the ideas and methods used are in no way specific to that organism. The future of molecular genetics depends on going beyond the isolation of CRMs and clusters of binding sites to a point where gene expression can be predicted directly from sequence in conjunction with information about relevant transcription factors and their binding specificity. The results of this paper show that our model has achieved this capability on a problem that is difficult if not impossible to solve by currently available methods. The proximal 1.1 kb of the *eve* control region contains diffuse binding sites that are not clustered, and yet it is required for the expression of stripe 7 from the proximal 1.7 kb of *eve* control region. The distal portion of this region of DNA, while containing clusters, is unable to drive stripe 7 expression. Moreover, we have shown that our model is predictive. Dividing control regions into separable functional units remains a major goal for the understanding and manipulation of gene expression. The work reported here is a natural generalization of previous efforts to attain this goal. It is based on a specific hypothesis proposed to explain the existence of CRMs and minimal elements, namely that their capability to act independently stems from the limited range of action of quenchers.

This advance in methodology relied on a close integration of modeling and experiment. The data required for the model are themselves a significant experimental advance, which revealed new phenomena because of quantification and high spatial resolution. Our ability to resolve temporal changes with a resolution of 6.5 min also adds a new dimension to the study of promoter-reporter constructs. We found that stripe 7 activity driven from p1.7*eve-lacZ* attains maximum expression levels about 10 min later than does stripe 2 driven from the same fragment, a phenomenon that is perhaps associated with increased expression in the posterior *hb* domain. We also noted that the time course of stripe 2 expression differs depending on whether it is driven by p1.7*eve-lacZ* or by MSE2 alone without proximal sequences. Expression from MSE2 alone reaches a maximum at T6 and stays at a high constant level until after gastrulation, whereas expression from p1.7*eve-lacZ* reaches maximum at T5 and then declines. This decline—a change of stripe 2 expression in the time, rather than space, domain—is mediated by sequences outside of MSE2, a finding incompatible with the standard paradigm.

This study provides new insights into *eve* transcriptional control. With respect to stripe 2, this work further confirms the classic experimental analysis[11,16] and extends it with the following new results. The 34-site model predicts that 7 transcription factors bind to the 1.7 kb of *eve* upstream regulatory sequences. In addition to Bcd, Hb, Kr and Gt, which have been previously shown to bind to the MSE2 region, the model predicts sites for Tll, Cad and Kni. Sites for the latter two have been predicted by others[3].

With respect to stripe 7, we show that the timing of the formation of stripe 7 is determined by the rise in expression of an activator, Cad, and not by the decrease of a repressor. Repression by Tll is necessary to define the position of the posterior border of stripe 7. However, stripe 7 driven by the MSE3 CRM is absent from *tor* and *tll* mutants[13]. This effect may be due to indirect repression from Kni, as suggested by our model and by the facts that the posterior Kni domain expands posteriorly in *tor* and *tll* mutants[21,22] and that there is no posterior Kni domain in *hs-tll* mutants[23]. Moreover, Kni is known to directly repress MSE3[13]. The fact that Tll expression does not overlap with stripe 7 further supports a role for Tll as a repressor.

A critical question remains: how are the activities of MSE3, the proximal 1.7 kb and other sequences integrated to give rise to the behavior of the native stripe 7? The answer to this question will require the quantitative modeling of expression driven from the 4.8 kb of DNA 5′ to the *eve* transcription start site. This region of DNA, driving the *eve* coding region rather than *lacZ*, has been shown to confer full biological activity on stripes 2, 3 and 7 in flies deficient for native *eve*[24]. Modeling this region of DNA raises the question of corepression and coactivation. In coactivation, a coactivator bound close to a repressor causes the repressor to behave as an activator (and similarly for corepression). In our study, Hb was treated as an activator, but it is known to be a repressor on MSE3[13]. However, it is also believed that *bcd* coactivates *hb* on MSE2[19]; hence, the effects of coactivation and corepression are likely to be important. These mechanisms can be represented in the model using the same ideas employed for quenchers.

Although coactivation, corepression and other mechanisms remain to be incorporated into the model, the feasibility of our approach is supported by the ability to construct models from sequence data using PWMs. The quality of PWMs is limited by the number of experimentally identified binding sites used to construct them. Despite this limitation we believe that the collection of sites used in this study is reasonably reliable. We base this assertion on the robustness of our results with respect to the total number of sites included as determined by the threshold of significance of PWM scores (**Supplementary Fig. 3** and **Supplementary Table 2** online). Besides using the PWMs to identify binding sites, we also used them to predict ligand affinity K (**Supplementary Table 1** online). Fitting all K values independently did not significantly improve the results (**Supplementary Table 2** online). This is important because it means that only a small number of parameters, $2L + 2$, is needed, where L is the number of ligand species. If each binding site required an independent affinity, the number of model parameters would scale with the number of binding sites and hence with the length of the DNA segment modeled. Instead it scales with the number of ligands. In summary, the full set of binding site affinities is provided by the DNA sequence itself.

METHODS

In situ hybridization. Embryos bearing the –1.7 kb or –1.55 delta 1.1 (MSE2) *eve-lacZ* fusion genes[11] were collected, fixed and stained for *lacZ* mRNA by *in situ* hybridization and for Eve protein by immunostaining using modified standard protocols[25,26]. Fixation was done in 1× PBS + 50 mM EGTA + 10% formaldehyde (Tousimis) with an equal volume of heptane. Acetone was used for permeabilization of the embryos as described[27]. The *lacZ* riboprobe used a 2.5-kb *Pvu*II *lacZ* fragment blunt-cloned into the *Eco*RV site of pBluescriptIIKS+ (gift from S. Small) and was labeled with fluorescein by transcription using T3 polymerase. After hybridization, *lacZ* mRNA was visualized by sequential incubation with rabbit antibody to fluorescein (Molecular Probes), followed by antibody to rabbit labeled with Alexa Fluor 647 (Molecular Probes). The embryos were simultaneously incubated with guinea pig antibody to Eve[28] and antibody to guinea pig labeled with Alexa Fluor 555 (Molecular Probes) to detect endogenous Eve protein. After antibody incubations, each embryo was stained with PicoGreen (Molecular Probes) for 20 min to visualize the DNA. All antibody incubations and washes were done in PBS + 0.1% Tween20. Blocking was done in Western Blocking Reagent (Roche), diluted 5 times. All secondary antibodies were preabsorbed by incubating them with 0- to 12-h-old wild-type *D. melanogaster* embryos for at least 2 h at 4 °C. Embryos were mounted in 40 µl mounting medium (4% n-propylgallate and 90% glycerol in 1× PBS buffer, pH 8.0) and covered with a 22 × 30 mm cover glass (No. 1 1/2). Detection of *lacZ* mRNA in the absence of Kni protein was done in embryos homozygous for the deficiency Df(3L)ri-79C (Bloomington Stock Center) and homozygous for p1.7*eve-lacZ*. Lack of Kni was confirmed by immunostaining for Kni protein (data not shown).

Quantitative expression data. Scanning of fluorescently stained embryos was carried out as described[29]. Fluorophores were excited with three different laser wavelengths (488, 543 and 633 nm) and the detection was done in a filterless spectral separation system with nonoverlapping wavelength windows of 500–545 nm, 560–645 nm and 650–715 nm, respectively. Image segmentation was carried out as described[29]. Embryos were classified temporally as belonging to either C13, or one of eight time classes (T1-T8), each about 6.5 min long, in cycle 14A (C14A), as described[30]. Background removal of RNA signal was as described[31] with an additional smoothing step by wavelets[32] performed before finding the background parabola; smoothed data at this step were only used for estimating background. We were able to remove background from young embryos by finding individual nonexpressing nuclei[31]. Registration was performed by registering to preexisting integrated *eve* data[32]. Ligand data used here have been described[30], with the addition of new Tll data starting with C13, averaged from at least ten embryos per time class. Data from the middle 10% of dorsoventral positional values of each embryo were averaged for each time class. Nuclei were grouped into 50 (C13) and 100 (C14A) equal-sized bins according to their position along the AP axis. Data for C13 were then duplicated to obtain the same number of data points as in C14A. Finally, averaged data for C13 were smoothed using singular spectrum analysis[33]. The numbers of embryos N per time class used to generate the average expression profiles for the 1.7 *eve-lacZ* construct were 8 (C13), 21 (T1), 12 (T2), 18 (T3), 11 (T4), 22 (T5), 18 (T6), 13 (T7) and 10 (T8). The model was fit to ligand data from 35% to 92% AP.

Selection of binding sites. We ran the program patser-v3d with options -A a:t 0.297 c:g 0.203 -ls -c -d1 (refs. 3,34) to convert alignment matrices into PWMs and to search 1.7 kb of *eve* upstream regulatory sequences for binding sites for Hb, Kr, Bcd, Cad[3], Kni and Tll[6]. The alignment matrix for Cad was shortened by 2 bp at the 5′ end. We regarded the 17 footprinted sites as highly reliable[11,16,18,19] and added predicted sites to the model in order of decreasing P value. If overlapping sites for the same ligand were found, the site with the highest score was kept. Additional sites for Gt were found by searching for sites closely matching the two known Gt binding sites in regulatory sequences in the *Kr* gene[35], resulting in three putative sites. In a final step, the length of each binding site was set to 14 bp, except for Gt, where the length of the binding sites is 28 bp. Lists of binding sites used in the different models are given in **Supplementary Table 1** online.

Computation and optimization. The model Equations, shown in **Box 1**, were implemented in C. Parameters were determined by minimizing the summed squared difference between the model output and the data, which consisted of 406 sets of transcription factor concentrations and RNA output over seven different times. Optimization was performed using the Lam simulated annealing schedule[36–38]. For Gt, $K_{i[m_i,n_i;\mathrm{Gt}]}$ was determined by optimization for each site i. For every other protein a only the $K_{j[m_j,n_j;a]}$ belonging to the site j with the highest PWM score was determined by optimization, with other K values determined from K_j by taking the exponent of the difference in PWM scores.

Parameter search spaces were set by explicit search limits for K_a, Θ, R_0 and C_a, with $f^{AF} = 0.99$ and $Q = 1$ (**Box 1**). Each annealing run required from 1 to 10 d of computation on a single P4 (2.8 GHz) or Xeon (2.6 GHz) processor. Runs were repeated 5 times with different random seeds for each optimization problem. The quality of the runs was judged by its root mean square score and by visual observation of the expression pattern.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank P. Gergen for comments on the manuscript. E.M., H.J., J.J., A.K. and J.R. were supported by US National Institutes of Health grant 2 ROI RR07801. D.S., H.J. and S.H. were supported in part by contract W-7405-ENG-36 from the US Department of Energy. H.J. received partial support from the State University of New York.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
2. Yuh, C.-H., Bolouri, H. & Davidson, E.H. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128**, 617–629 (2001).
3. Berman, B.P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **99**, 757–762 (2002).

4. Berman, B.P. *et al.* Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* **5**, R61 (2004).
5. Clyde, D.E. *et al.* A self-organizing system of repressor gradients establishes segmental complexity in *Drosophila*. *Nature* **426**, 849–853 (2003).
6. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E.D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **3**, 30 (2002).
7. Schroeder, M.D. *et al.* Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2**, e271 (2004).
8. Gray, S., Szymanski, P. & Levine, M. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes Dev.* **8**, 1829–1838 (1994).
9. Goto, T., MacDonald, P. & Maniatis, T. Early and late periodic patterns of *even-skipped* expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell* **57**, 413–422 (1989).
10. Harding, K., Hoey, T., Warrior, R. & Levine, M. Autoregulatory and gap gene response elements of the *even-skipped* promoter of *Drosophila*. *EMBO J.* **8**, 1205–1212 (1989).
11. Small, S., Blair, A. & Levine, M. Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. *EMBO J.* **11**, 4047–4057 (1992).
12. Small, S., Arnosti, D.N. & Levine, M. Spacing ensures autonomous expression of different stripe enhancers in the *even-skipped* promoter. *Development* **119**, 767–772 (1993).
13. Small, S., Blair, A. & Levine, M. Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Dev. Biol.* **175**, 314–324 (1996).
14. Ludwig, M. *et al.* Functional evolution of a cis-regulatory module. *PLoS Biol.* **3**, e93 (2005).
15. Reinitz, J., Hou, S. & Sharp, D.H. Transcriptional control in *Drosophila*. *ComPlexUs* **1**, 54–64 (2003).
16. Stanojevic, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science* **254**, 1385–1387 (1991).
17. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
18. Stanojevic, D., Hoey, T. & Levine, M. Sequence-specific DNA-binding activities of the gap proteins encoded by *hunchback* and *Krüppel* in *Drosophila*. *Nature* **341**, 331–335 (1989).
19. Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. Transcriptional regulation of a pair-rule stripe in *Drosophila*. *Genes Dev.* **5**, 827–839 (1991).
20. Arnosti, D.N., Barolo, S., Levine, M. & Small, S. The *eve* stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 (1996).
21. Pankratz, M.J., Hoch, M., Seifert, E. & Jäckle, H. *Krüppel* requirement for *knirps* enhancement reflects overlapping gap gene activities in the *Drosophila* embryo. *Nature* **341**, 337–340 (1989).
22. Rothe, M., Wimmer, E.A., Pankratz, M.J., González-Gaitán, M. & Jäckle, H. Identical transacting factor requirement for *knirps* and *knirps-related* gene expression in the anterior but not in the posterior region of the *Drosophila* embryo. *Mech. Dev.* **46**, 169–181 (1994).
23. Steingrimsson, E., Pignoni, F., Liaw, G.J. & Lengyel, J.A. Dual role of the *Drosophila* pattern gene *tailless* in embryonic termini. *Science* **254**, 418–421 (1991).
24. Fujioka, M., Jaynes, J.B. & Goto, T. Early *even-skipped* stripes act as morphogenetic gradients at the single cell level to establish *engrailed* expression. *Development* **121**, 4371–4382 (1995).
25. Hughes, S.C. & Krause, H.M. Single and double FISH protocols for *Drosophila*. in *Confocal Microscopy Methods and Protocols: Methods in Molecular Biology* Vol. 122 (ed. Paddock, S.W.) 93–101 (Humana Press, Totowa, New Jersey, 1998).
26. Wu, X., Vasisht, V., Kosman, D., Reinitz, J. & Small, S. Thoracic patterning by the *Drosophila* gap gene *hunchback*. *Dev. Biol.* **237**, 79–92 (2001).
27. Nagaso, H., Murata, T., Day, N. & Yokoyama, K.K. Simultaneous detection of RNA and protein by *in situ* hybridization and immunological staining. *J. Histochem. Cytochem.* **49**, 1177–1182 (2001).
28. Kosman, D., Small, S. & Reinitz, J. Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Dev. Genes Evol.* **208**, 290–294 (1998).
29. Janssens, H. *et al.* A high-throughput method for quantifying gene expression data from early *Drosophila* embryos. *Dev. Genes Evol.* **215**, 374–381 (2005).
30. Jaeger, J. *et al.* Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster. Genetics* **167**, 1721–1737 (2004).
31. Myasnikova, E., Samsonova, M., Kosman, D. & Reinitz, J. Removal of background signal from *in situ* data on the expression of segmentation genes in *Drosophila*. *Dev. Genes Evol.* **215**, 320–326 (2005).
32. Myasnikova, E., Samsonova, A., Kozlov, K., Samsonova, M. & Reinitz, J. Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. *Bioinformatics* **17**, 3–12 (2001).
33. Elsner, J. & Tsonis, A. *Singular Spectrum Analysis: a New Tool in Time Series Analysis* (Plenum, New York, 1996).
34. Hertz, G.Z. & Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
35. Capovilla, M., Eldon, E.D. & Pirrotta, V. The *giant* gene of *Drosophila* encodes a b-ZIP DNA-binding protein that regulates the expression of other segmentation gap genes. *Development* **114**, 99–112 (1992).
36. Lam, J. & Delosme, J.-M. *An efficient simulated annealing schedule: derivation. Technical Report 8816* (Yale Electrical Engineering Department, New Haven, Connecticut, 1988).
37. Lam, J. & Delosme, J.-M. *An efficient simulated annealing schedule: Implementation and evaluation. Technical Report 8817* (Yale Electrical Engineering Department, New Haven, Connecticut, 1988).
38. Reinitz, J. & Sharp, D.H. Mechanism of *eve* stripe formation. *Mech. Dev.* **49**, 133–158 (1995).