Table of Contents

1. The GEMSTAT model

GEMSTAT estimates the probability of gene expression from the ensemble of all possible configurations of bound TFs and BTM. To this end GEMSTAT computes a "statistical weight" Z_c for each configuration c from the energies of the protein-DNA, BTM-DNA, protein-BTM, and protein-protein interactions in c (Shea and Ackers 1985). Below we elaborate on the computation of Z_c.

For a configuration c, the expression for Z_c has terms reflecting binding of TFs to their cognate sites and those reflecting TF-TF interactions. If c is a BTM-bound configuration, then Z_c will have additional terms reflecting TF-BTM interactions and the binding of BTM to promoter (Figure 1). Through these four types of terms, Z_c captures the energy of various binding and interaction events occurring in c, where the

ground state for computing the energies is a configuration where no TF or the BTM is bound. We explain below how different types of binding and interaction events are accommodated in the formulation of Z_c.

A. TF-DNA binding

For a given site S, the binding of a TF f at S contributes a statistical weight of $q_{f,S} = K_{f,S}[f]$ to Z_c. Here $K_{f,S}$ is the equilibrium constant of the DNA-binding reaction between f and S, and $[f]$ is the concentration of f. Let S^f_{max} denote the strongest binding site of f and $K(S^f_{max})$ denote the association constant of the TF-DNA binding reaction between f and S^f_{max}. Then we can re-write $K_{f,S}$ as $K(S^f_{max})\exp(-\beta\Delta E_{f,S})$, where β is the Boltzmann constant and $\Delta E_{f,S}$ denotes the "mismatch energy" of the site S relative to S^f_{max} for f. According to the theory of Berg and von Hippel (Berg and von Hippel 1987), we can estimate $\exp(-\beta\Delta E_{f,S})$from $\exp(-LLR(f,S) + LLR(f,S^f_{max}))$, where $LLR(f,\cdot)$ is the log likelihood ratio score of a site, computed based on the known position weight matrix (PWM) of f and the background nucleotide distribution (Stormo 2000).

The concentration $[f]$ of the TF f is in arbitrary units and essentially can be re-written as $v[f]_{rel}$ where $[f]_{rel}$ is the concentration of f relative to some unknown reference value v. The expression for $q_{f,S}$ then becomes:

$$q_{f,S} = K(S^f_{max})v[f]_{rel}\exp(LLR(f,S) - LLR(f,S^f_{max}))$$

where both $K(S^f_{max})$ and v are unknown quantities. We take their product $K(S^f_{max})v$ as a free parameter in our model and refer to it as the "DNA-binding parameter" for the TF f. We note that, the estimated values of different TFs' DNA-binding parameters in our model are not biochemically comparable since this parameter represents a product of a biochemical parameter (i.e., $K(S^f_{max})$) and an unknown reference value (i.e., v). We also note that, owing to this formulation, we can fit our model using the *relative* levels of mRNA and TF expression.

In case the site S is a signaling pathway response element and the signaling activity is known to attenuate the DNA binding affinity of f, then using the concentration of a chemical species $Sgnl$ whose spatial distribution correlates with the signal's level of activity, we model a modification of $q_{f,S}$ as follows.

$$q_{f,S,Sgnl} = K(S^f_{max})v[f]_{rel}\exp(LLR(f,S) - LLR(f,S^f_{max}) - \varphi([Sgnl]))$$

We used $\varphi([\mathrm{dpERK}]) = C \times [\mathrm{dpERK}]$ in this study, where C is a free parameter, to model an attenuation of Cic's DNA binding affinity under the influence of ERK. This mechanism is suggested by the recent studies of ERK-dependent Cic regulation in mammalian cells (Dissanayake, Toth et al. 2011). Note that, the above formulation for $q_{f,S,Sgnl}$ can also be interpreted as one where the mismatch-energy related term is not modified, rather the association constant $K(S^f_{max})$ is modified as $K'(S^f_{max}) = K(S^f_{max})\exp(-\varphi([Sgnl]))$.

B. TF-TF interaction

If two TFs f_1 and f_2 interact when bound to closely located sites (with no other TF bound between them), as opposed to one TF binding independently of the other (Shea and Ackers 1985), then for each such instance of f_1 and f_2 bound in a configuration c, the statistical weight Z_c includes an extra multiplicative term ω_{f_1,f_2}. This term essentially represents the energy of interaction between f_1 and f_2. As such, $\omega_{f_1,f_2} > 1$ or < 1 depending on whether the interaction between f_1 and f_2 enhance or diminish their occupancy in those closely located sites. Note that, f_1 and f_2 may denote the same TF.

C. TF-BTM interaction

We assume each TF f to impart on the BTM a "transcriptional effect", which essentially represents the energy of interaction between f and the BTM. As such, for each instance of f binding to one of its cognate sites in a BTM-bound configuration c, the statistical weight Z_c includes an extra multiplicative term α_f. If f facilitates the recruitment of BTM, then f is a transcriptional activator and $\alpha_f > 1$. Similarly, $\alpha_f < 1$ if f is a transcriptional repressor.

D. BTM-promoter binding

We include a parameter q_{BTM} in Z_c for every BTM-bound configuration c to capture the energy of BTM binding to promoter.

Considering all the possible binding events occurring in a configuration c, we then write the term Z_c as:

$$Z_c = \left(\prod_S \left({q_{f,S}}^{\sigma_{f,S}} \prod_{\substack{S' < S \text{ and} \\ N(S',S)=0}} {\omega_{f,g}}^{\sigma_{f,S} \times \sigma_{g,S'}} \right) \right) \left(q_{BTM} \prod_S {\alpha_f}^{\sigma_{f,S}} \right)^{\sigma_{BTM}}$$

Where

- Sites in the enhancer are ordered according to their location in a scan of the enhancer (either 5' to 3' or 3' to 5'),
- $\sigma_{f,S}$ is an indicator variable (0/1) to denote where TF f binds to site S,
- σ_{BTM} is an indicator variable (0/1) to denote whether c is a BTM-bound configuration, and
- $N(S', S)$ denotes the number of TF-occupied sites located between two specific sites S' and S where $S' < S$.

As mentioned in the Results section, we ultimately compute the probability of BTM-bound configurations, i.e.,

$$\text{P(bound BTM)} = \frac{Z_{\text{bound}}}{Z_{\text{unbound}}+Z_{\text{bound}}},$$

where the denominator $Z_{\text{unbound}} + Z_{\text{bound}}$ equates Z, the partition function. An efficient computation of the partition function involves application of dynamic programming and the relevant formulations are given in detail in (He, Samee et al. 2010).

2. Annotation of TF binding sites

To annotate a TF's binding sites in an enhancer, we first compute the log likelihood ratio (LLR) score of each *k*-bp window in the enhancer, where *k* denotes the length of the TF's motif (represented by a position weight matrix, PWM) and the two likelihoods in the ratio are computed from the PWM and a uniform background distribution. A window is then annotated as a binding site for the TF if the window's LLR score is at least half the LLR score of the TF's optimal site (maxLLR). In our experience of working with other datasets of *D.mel* developmental gene regulation, this threshold is weak enough to include the experimentally annotated sites for each TF, while maintaining efficiency in model optimization. The motif PWMs used in this model were all collected from FlyFactorSurvey (Zhu, Christensen et al. 2011).

3. Note on the assumption of Zld-Dl cooperativity

We noted the presence of five Zld sites in the *ind* enhancer, with two pairs of adjacent Dl-Zld sites located < 25 bps apart (Figure 2D), and similarly spaced Dl-Zld sites in orthologous sequences in other *Drosophila* species (Figure S1; the pattern is more obvious in the species related closely to *D.mel*), suggesting the inclusion of Dl-Zld cooperativity in our model. Dl-Zld cooperativity is expected to allow the uniformly expressed Zld to accentuate Dl activation and could in principle lead to a steeper dorsal boundary of *ind* expression (Kanodia, Liang et al. 2012), mirroring a similar mechanism in the *sog* enhancer (Liberman and Stathopoulos 2009). Including Dl-Zld cooperativity can also act as a surrogate for chromatin-mediated effect of Zld on Dl activation, as suggested in our recent work (Cheng, Kazemian et al. 2013).

4. Parameter ranges and parameter sampling scheme

Ranges of our model parameters (assumed based on our previous studies on thermodynamic modeling of enhancer readout (He, Samee et al. 2010, Duque, Hassan Samee et al. 2013, Samee and Sinha 2013, Samee and Sinha 2014)) and whether the sampling was performed in the logarithmic scale or in the linear scale for the corresponding parameter is given below. K: range between 10^{-2} and 10^{4}, sampled in logarithmic scale; α (for activators): range between 1 and 10, sampled in linear scale; α (for repressors): range between 10^{-5} and 1, sampled in logarithmic scale; ω: range between 1 and 100, sampled in logarithmic scale; q_{BTM}: range between 10^{-3} and 10^{-2}, sampled in logarithmic scale; Cic_{ATT}: range between 0 and 32, sampled in linear scale.

5. Embryo imaging and extracting spatial fluorescent data of protein- and mRNA- expression

Following fly strains were used: Oregon-R as a wild type, *sna*^{IIG05}, *ind*$^{1.4WT}$-*lacZ*, *ind*$^{1.4Cicmut}$-*lacZ*, *ind*$^{1.4Zldmut}$-*lacZ*, and *ind*$^{1.4Dlmut}$-*lacZ*. *Ind*$^{1.4Zldmut}$ flies were generated using a transcriptional reporter construct (*ind*$^{1.4Zldmut}$-*lacZ*) containing the *ind*$^{1.4}$ enhancer with each of its four Zld binding sites [CAGG (T/C) A (G/A)] mutated to CCAACAA via recombinant PCR. *Ind*$^{1.4Dlmut}$ flies were generated using a reporter construct (*ind*$^{1.4Dlmut}$-*lacZ*) where three evolutionarily conserved Dl sites (AGGAAAATTCC, TGGGAAATTCCC, CCGAAATTCC) in the *ind*$^{1.4}$ enhancer were mutated to TGATATCT. WT and mutant *ind*$^{1.4}$-*lacZ* reporters were assembled in the *placZattB* vector (http://www.flyc31.org/sequences_and_vectors.php) and transformed into flies by phiC31-mediated integration (Bischof, Maeda et al. 2007) at chromosomal position 86Fb, thereby permitting direct comparison of their expression levels.

Mouse anti-Dl (DSHB), rabbit anti-dpERK (Cell Signaling), rat anti-Vnd (gift from Ze'ev Paroush, Hebrew University), rabbit anti-Cic (gift from Celeste Berg, U of Washington), guineapig anti-Sna (gift from Eric Wieschaus, Princeton University), rat anti-Zld (gift from Chris Rushlow, NYU), sheep anti-digoxigenin (Roche), and mouse anti-biotin (Jackson Immunoresearch) were used as primary antibodies. DAPI (Vector laboratories) was used to stain for nuclei, and Alexa Fluor conjugates (Invitrogen) were used as secondary antibodies. See Supplementary Text for the details of embryo imaging.

Nikon A1-RS scanning confocal microscope and 60x Plan-Apo oil objective was used for imaging. Optical cross-sections of vertically oriented embryos were taken at about 90μm from the posterior or anterior pole of an embryo. Spatial profile of proteins and mRNAs were automatically extracted along the entire DV axis from raw fluorescent images using a customized MATLAB script, as described previously (Lim, Samper et al. 2013).

6. Smooth functions to approximate spatial profiles from fluorescent data

We describe below our choice for the smooth function f for each regulator/gene (also see Figure S1C). In the following mathematical expressions, we use x to denote location along the D/V axis (x increases from the ventral-most to the dorsal-most point). For $h =$ Vnd, dpERK, or *ind*, we use x_h to denote the location of the peak expression of h. Moreover, for $h =$ Vnd and dpERK, we use x_V to denote location along D/V axis from the ventral-most point to x_h, and we use x_D to denote location along D/V axis from x_h to the dorsal-most point.

- For Dl and *ind*, we chose $f = a \times \exp(-\frac{(x-b)^2}{2c^2})$, leaving a and c as free parameters and setting $b = 0$ for Dl and $= x_{ind}$ for *ind*.
- For Sna, we chose $f = \frac{a}{1+\exp(mx-c)}$ and left a, m, and c as free parameters.

- For Vnd and dpERK, we chose $f = \frac{a_V}{1+\exp(-m_V x_V + c_V)} + \frac{a_D}{1+\exp(m_D x_D - c_D)}$ and left the $a, m,$ and c parameters as free. Note that our data shows asymmetry in the profiles of Vnd and dpERK with respect to the locations of their peak expression (x_h) and hence, in these two cases we use different parameterizations of the smoothing function on the ventral and the dorsal side of x_h.

For Zld and Cic, we do not explicitly fit any function g. Rather we set f to an arbitrarily chosen constant value. In this study, we set f to the maximum value observed in the mean expression profile of the respective TF.

7. Computation of histograms in Figure 4G

The *LacZ* profiles show the range of intensity values across all embryos are similar. Each histogram in Figure 4G shows the mean intensity value computed from a bootstrapped data set of 1000 intensity profiles. The pixel-count for each intensity value in a bootstrapped profile is a uniform sample from the pixel-counts for the same intensity value in N_{emb} real intensity profiles, where N_{emb} (= the number of embryos imaged) was 20 and 16 for the Zld and Dl site mutagenesis experiments, respectively. The plots in Figure 4K show the smoothed histograms from the WT and mutant *LacZ* profiles, where each histogram was created from N_{emb} profiles (one profile from each embryo) on 256 bins, i.e., one bin for each possible value of intensity, and each bin records the count of pixels having the corresponding intensity.

8. Effect on *LacZ* expression domain driven by site mutagenized *ind* enhancers

LacZ expression domains driven by the site mutagenized *ind* enhancers are not uniform across the anterior-posterior axis: expressions are weaker near the poles and at the middle (e.g., see Figure 4F,I in main text). Our model predictions – computed on data collected from planes close the poles (Lim, Samper et al. 2013) – conform to the observation about reduced expression near the poles. In particular, the "width" of *ind* expression (defined as the number of consecutive bins where relative *ind* expression is at least 1E-3) is 20 bins along the D/V axis in our data. Our models suggest that for Dl and Zld site mutations, the width of *LacZ* expression driven by mutagenized enhancers should be 7 and 9 bins (mean values), respectively. Thus we predict 65% and 55% reduction in *LacZ* expression width upon mutating Dl and Zld sites, respectively.

To quantify how site mutagenesis changes the entire *LacZ* expression domains from the WT to the mutagenized enhancers, we computed the distributions of pixel-intensities within the portions of *LacZ* domains that overlap with the expression domain of *ind* mRNA (Figure S3). For both mutagenesis experiments, we found an appreciable reduction in the *LacZ* domains from the WT to the mutagenized enhancer. In the experiment on Dl site mutagenesis, we found the *LacZ* domains under the WT and the mutagenized enhancers recapitulate 72% and 45% (mean values) of the *ind* mRNA expression domain, respectively. In the experiment on Zld site mutagenesis, the numbers were found to be 71% and 50%,

respectively. Thus the experimental data suggest ~40% and ~30% reduction in *LacZ* expression domain upon mutating Dl and Zld sites, respectively.

9. Note on dorsal repression of *sog* expression

We note that, the *sog* enhancers do not have any binding site for Cic; and although a recent result suggests the possibility of a repressive input defining the dorsal boundary of *sog* expression (Ozdemir, Ma et al. 2014), the identity or necessity of such a dorsal repressor has not been confirmed yet through site mutagenesis experiments.

10. How is an experiment to mutate all four Cic sites informative?

We discuss below how the different possible outcomes of mutating all four Cic sites in the *ind* enhancer are informative. First, a non-basal level of *ind* expression in dorsal-ectoderm upon mutating Cic sites will imply a more direct activating input from Zld and: (a) help us reject the models whose predictions are grossly different (e.g. by several standard deviations) than the experimental result, and (b) suggest a second experiment where spacing between Dl-Zld sites is disrupted so that the effect of Dl-Zld cooperativity can be quantified. On the other hand, a basal level of *ind* expression will suggest Zld playing only a chromatin remodeling role, as has been suggested in (Foo, Sun et al. 2014).

11. Supplementary Figure Legends

Figure S1. Related to Figure 2 and Figure 5. (A) Location of computationally identified binding sites for Dl, Sna, Vnd, Zld, and Cic in the *D.mel ind* enhancer and its orthologs in ten other *Drosophila* species (each bar represents a binding site; height of a bar represents the ratio LLR/maxLLR, see the note on Annotation of TF binding sites; schematic created using the insite tool, downloaded from: https://www.cs.utah.edu/~miriah/insite/). Orthologous sequences were determined using the UCSC liftOver tool (Kuhn, Haussler et al. 2013). The genome release and the exact coordinates of the sequences from each species are as follows. *D.mel*: Apr. 2006 (BDGP R5/dm3), chr3L:15032420-15033835 and Apr. 2004 (BDGP R4/dm2), chr3L:15004579-15005994. *D.sim*: Apr. 2005 (WUGSC mosaic 1.0/droSim1), chr3L:14353571-14354969. *D.sec*: Oct. 2005 (Broad/droSec1), super_0:7149065-7150441. *D.yak*: Nov. 2005 (WUGSC 7.1/droYak2), chr3L:15124657-15125963. *D.ere*: Aug. 2005 (Agencourt prelim/droEre1), scaffold_4784:14991553-14993166. *D.ana*: Aug. 2005 (Agencourt prelim/droAna2), scaffold_13337:17951578-17952873. *D.pse*: Nov. 2004 (FlyBase 1.03/dp3), chrXR_group6:701248-702519. *D.per*: Oct. 2005 (Broad/droPer1), super_12:1384376-1385670. *D.vir*: Aug. 2005 (Agencourt prelim/droVir2), scaffold_13049:17677662-17678656. *D.moj*: Aug. 2005 (Agencourt prelim/droMoj2), scaffold_6680:22706091-22707098. *D.gri*: Aug. 2005 (Agencourt prelim/droGri1), scaffold_25023:1155716-1156282. (B) Predictions of the wild-type ensemble models. Shown is the mean (red) and the range (shaded red area around the curve) of the expression values predicted by the ensemble. (C) Fit between background-eliminated, mean fluorescence data (red) and smoothed functions (blue).

Figure S2. Related to Figure 4. (A-C) Expression patterns of *ind*, Vnd, and dpERK visualized wild-type and *sna* mutant embryos. (D) Predictions of the filtered ensemble models upon mutating all Dl sites. Semantics of the plot is the same as that of Figure 4A. (E) The three evolutionarily conserved Dl sites and their orthologous sequences in six *Drosophila* species.

Figure S3. Related to Figure 4. Effect of (A) Dl and (B) Zld site mutagenesis on *LacZ* expression domains. From all embryos in each experiment (n = 16 for Dl and n = 20 for Zld site mutagenesis), we computed the distribution of pixel-intensities within the portion of the *LacZ* domain that overlaps with expression domain of *ind* mRNA.

Figure S4. Related to Figure 4. Predictions of the filtered ensemble models upon mutating all Cic sites. Semantics of the plot is the same as that of Figure 4A.

References

Berg, O. G. and P. H. von Hippel (1987). "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters." J Mol Biol **193**(4): 723-750.

Bischof, J., R. K. Maeda, M. Hediger, F. Karch and K. Basler (2007). "An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases." Proc Natl Acad Sci U S A **104**(9): 3312-3317.

Cheng, Q., M. Kazemian, H. Pham, C. Blatti, S. E. Celniker, S. A. Wolfe, M. H. Brodsky and S. Sinha (2013). "Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy." PLoS Genet **9**(8): e1003571.

Dissanayake, K., R. Toth, J. Blakey, O. Olsson, D. G. Campbell, A. R. Prescott and C. MacKintosh (2011). "ERK/p90(RSK)/14-3-3 signalling has an impact on expression of PEA3 Ets transcription factors via the transcriptional repressor capicua." Biochem J **433**(3): 515-525.

Duque, T., M. A. Hassan Samee, M. Kazemian, H. N. Pham, M. H. Brodsky and S. Sinha (2013). "Simulations of enhancer evolution provide mechanistic insights into gene regulation." Mol Biol Evol.

Foo, S. M., Y. Sun, B. Lim, R. Ziukaite, K. O'Brien, C. Y. Nien, N. Kirov, S. Y. Shvartsman and C. A. Rushlow (2014). "Zelda potentiates morphogen activity by increasing chromatin accessibility." Curr Biol **24**(12): 1341-1346.

He, X., M. A. Samee, C. Blatti and S. Sinha (2010). "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression." PLoS Comput Biol **6**(9).

Kanodia, J. S., H. L. Liang, Y. Kim, B. Lim, M. Zhan, H. Lu, C. A. Rushlow and S. Y. Shvartsman (2012). "Pattern formation by graded and uniform signals in the early Drosophila embryo." Biophys J **102**(3): 427-433.

Kuhn, R. M., D. Haussler and W. J. Kent (2013). "The UCSC genome browser and associated tools." Brief Bioinform **14**(2): 144-161.

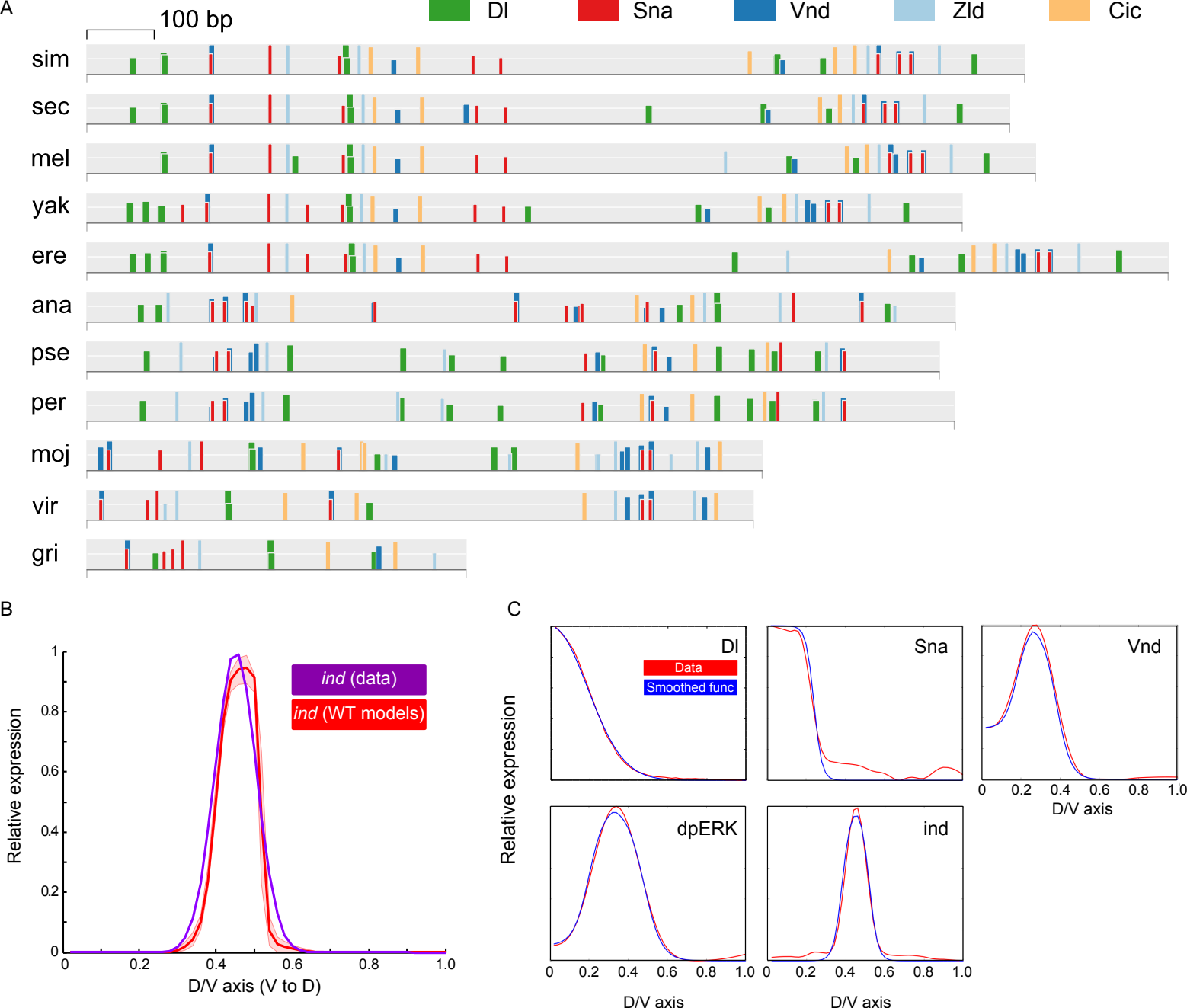
Liberman, L. M. and A. Stathopoulos (2009). "Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence." Dev Biol **327**(2): 578-589.

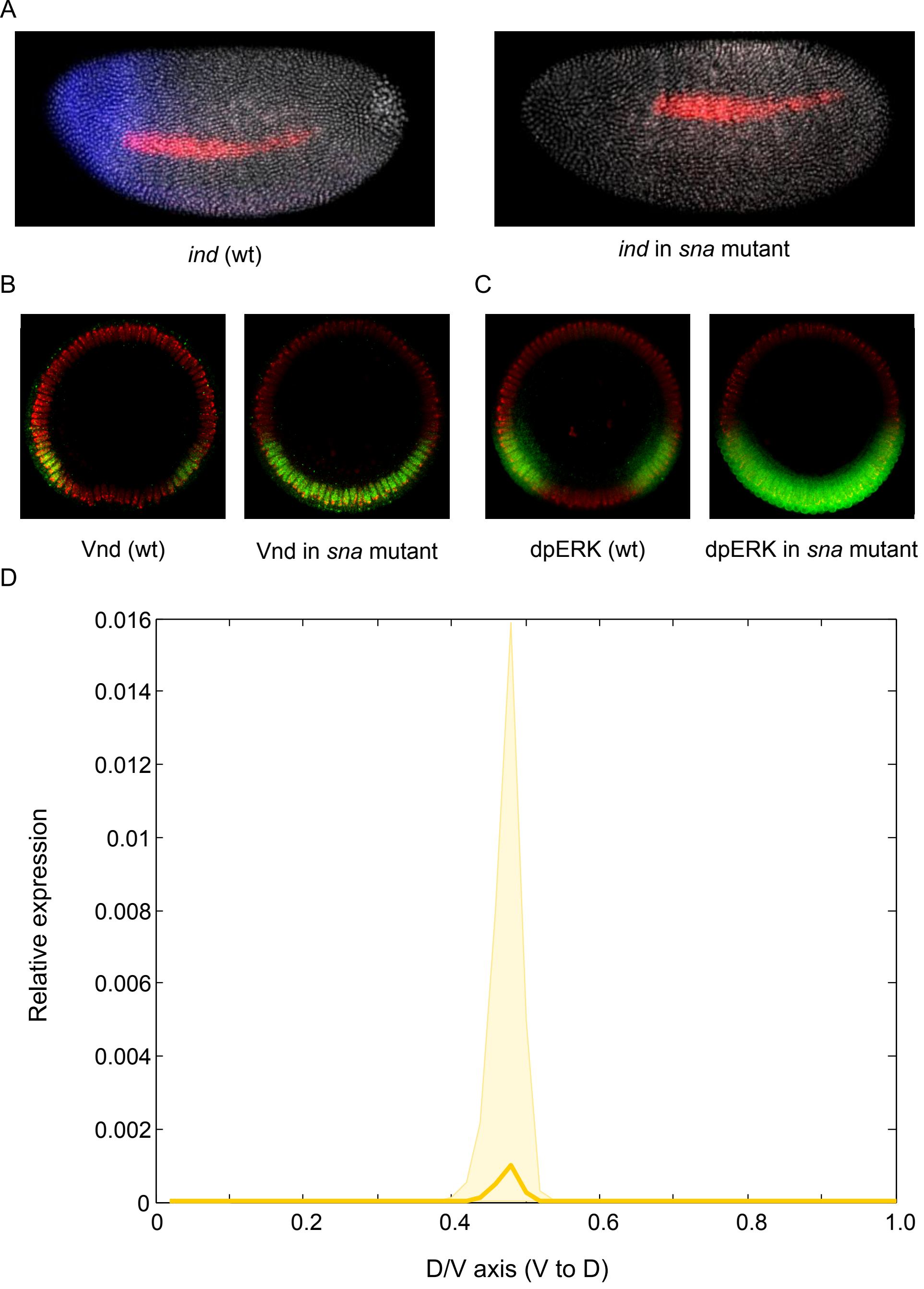
Lim, B., N. Samper, H. Lu, C. Rushlow, G. Jimenez and S. Y. Shvartsman (2013). "Kinetics of gene derepression by ERK signaling." Proc Natl Acad Sci U S A **110**(25): 10330-10335.

Ozdemir, A., L. Ma, K. P. White and A. Stathopoulos (2014). "Su(H)-mediated repression positions gene boundaries along the dorsal-ventral axis of Drosophila embryos." Dev Cell **31**(1): 100-113.

Samee, M. A. and S. Sinha (2013). "Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data." Methods **62**(1): 79-90.

Samee, M. A. and S. Sinha (2014). "Quantitative modeling of a gene's expression from its intergenic sequence." PLoS Comput Biol **10**(3): e1003467.
Shea, M. A. and G. K. Ackers (1985). "The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation." J Mol Biol **181**(2): 211-230.
Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.
Zhu, L. J., R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. A. Wolfe and M. H. Brodsky (2011). "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system." Nucleic Acids Res **39**(Database issue): D111-117.

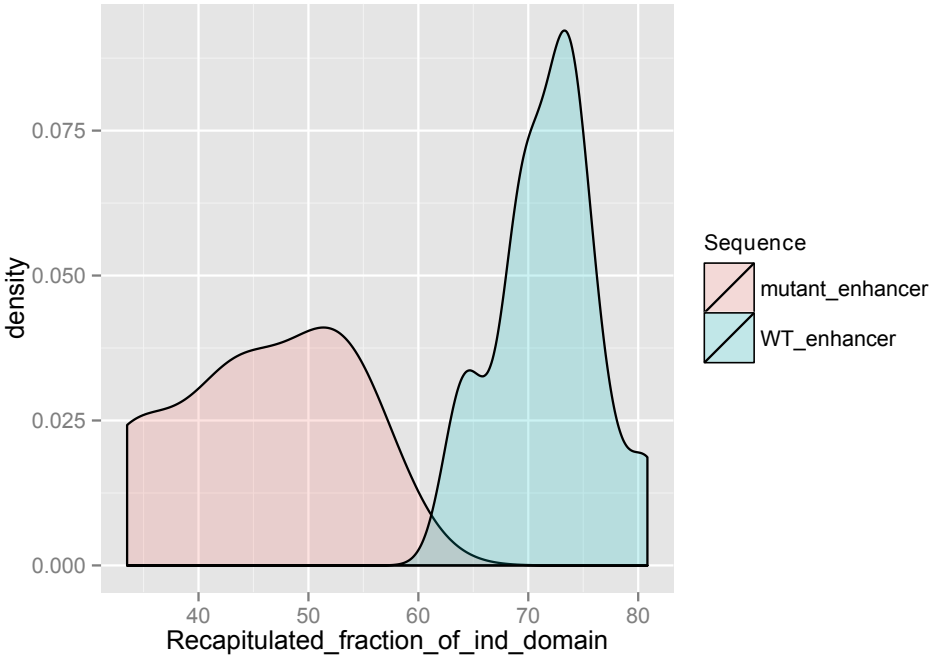

A
100 bp
Dl
Sna
Vnd
Zld
Cic
sim
sec
mel
yak
ere
ana
pse
per
moj
vir
gri
B
ind (data)
ind (WT models)
Relative expression
D/V axis (V to D)
C
Dl
Data
Smoothed func
Sna
Vnd
dpERK
ind
Relative expression
D/V axis
D/V axis



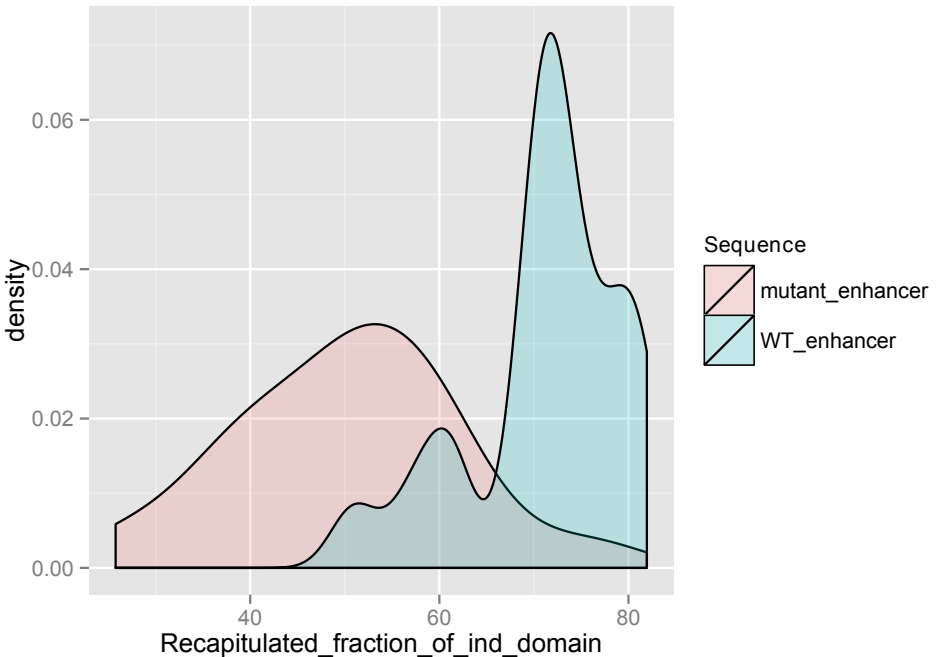
E

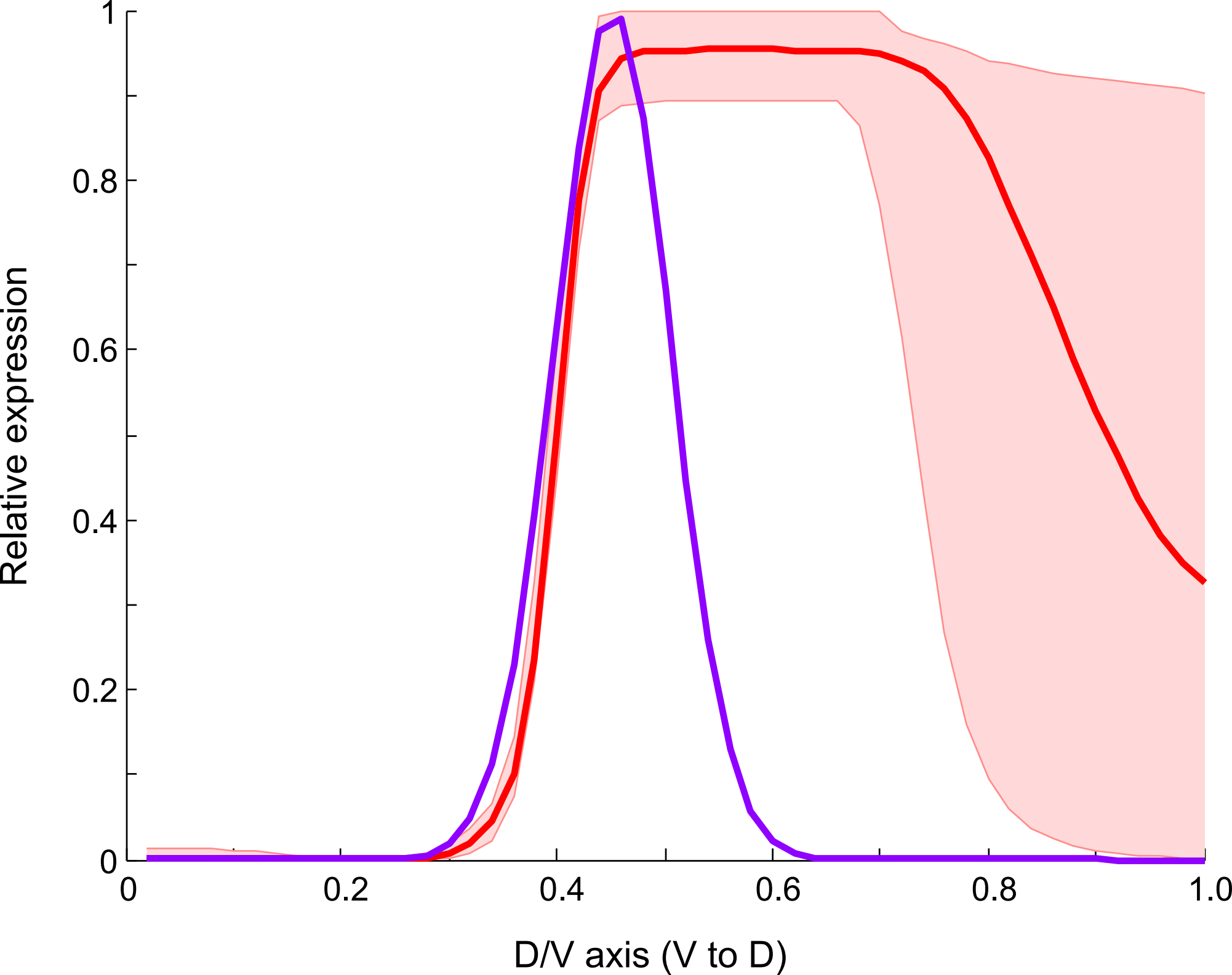
	Sites# 1,2 (two overlapping sites)	Sites# 4,5,6 (three overlapping sites)	Site# 9
mel:	AGGAAAATTCC	TGGGAAATTCCC	CCGAAATTCC
sim:	....	....	....
sec:	....	....	....
yak:	....	....	....
ere:	....	....	G........G
ana:	T......A...	....	TA.......A
pse:	T.A.....A.A	....G..A..T.	....

(A) Change in expression domain due to Dl site mutation



(B) Change in expression domain due to Zld site mutation




Relative expression
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
D/V axis (V to D)