Evolution of transcription control in the segmentation gene network of *Drosophila*

Saurabh Sinha[3]*, Michael Pearce[1]*, Ulrich Unnerstall[1], John Fak[1], Monica Dandapani[1], Mark D. Schroeder[1], Eric D. Siggia[2]* and Ulrike Gaul[1]*#

Derived from EvoDevo_EDSv1.doc

[1]Laboratory of Developmental Neurogenetics,
[2]Center for Studies in Physics and Biology,
Rockefeller University,
New York, NY 10021-6399, USA
[3]Dept. of Computer Science,
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

* equal contribution

#corresponding author
email: gaul@mail.rockefeller.edu
phone: 212 327 7621
fax: 212 327 7923

Abstract

Evolution is largely driven by changes in regulatory sequence, but the underlying mechanisms are not well characterized. Here we carry out a comprehensive computational and experimental analysis of transcriptional regulation in the segmentation gene network in two closely related *Drosophila* species, *melanogaster* and *pseudoobscura*, and correlate molecular and expression changes for 16 orthologous *cis*-regulatory modules. While the expression of the participating genes is very stable, their regulatory sequence is in strong evolutionary flux. Binding site content rapidly shifts along the DNA, altering length, position, and composition of modules. A functional module can shift its location to violate alignment-based homology, a module whose function is redundant with other modules can disappear, and large insertion/deletions can carry binding sites in or out of homologous modules, thereby altering functionality. However, redundancy among binding sites within one module and between modules regulating the same gene results in conservation of gene expression despite divergence at the sequence level. Aside from large indels, sequence-level changes appear driven jointly by point mutation and short tandem repeat expansion/contraction.

Developmental genes are important agents of evolution; however, their proteins are typically strongly conserved, suggesting that evolutionary divergence is driven primarily by changes in their spatio-temporal expression, and thus by changes in transcriptional regulation[1-4]. Instances of regulatory evolution are documented across a wide range of evolutionary time scales[5-10], but only in rare cases have phenotypic differences been mapped to molecular changes [4, 11-16]. The study of molecular evolution has in fact focused almost exclusively on protein-coding sequence, largely because the genetic code provides a simple framework for determining the effect of point mutations on protein function. By contrast, the functional units in non-coding sequence, namely transcription factor binding sites and *cis*-regulatory elements, are much less well defined, impeding the application of molecular evolutionary theory. Given the importance of regulatory evolution, a framework for understanding the underlying molecular events is urgently needed.

In this study, we investigate the evolution of the segmentation gene network of *Drosophila*. The network consists of a hierarchy of transcription factors that lay out the anterior-posterior axis of the embryo in a stepwise refinement of expression patterns[17, 18]. Most of the major participating factors, their expression patterns and binding site preferences are known. Many *cis*-regulatory elements, in particular those receiving input from maternal and early zygotic gap factors, have been identified[19]. They are typically organized as discrete modules of DNA sequence, about 1kb in length, that contain multiple binding sites for multiple transcription factors; binding site composition determines the expression pattern along the anterior-posterior axis. Exploiting this local clustering of binding sites, computational methods have been applied successfully to detect and characterize segmentation modules[19-22]. By several measures, the segmentation gene network is evolving fast: within the ~250 Myrs separating basal (*Anopheles*) from higher (*Drosophila*) Dipterans, new components are introduced, expression patterns are altered, and the non-coding sequence is completely diverged[23-26].

To investigate regulatory evolution within the segmentation gene network, we compared *D.melanogaster (D.mel)* and *D.pseudoobscura (D.pse*). The two species lie at an intermediate evolutionary distance of 25-30 Myrs and are close enough to reliably align the non-coding sequence[27, 28], yet far enough to expect measurable change in binding site composition of modules and their function. (Measured by the rate of synonymous codon divergence, these two species are comparable to human-chicken divergence[29].) We examine the top tiers of the network, specifically 39 validated modules in the non-coding regions of the zygotic gap and pair rule genes[19] (Supplement 1) and the input they receive from four maternal (Bicoid (Bcd), Caudal

(Cad), D-Stat, Tor-RE) activators and five gap gene repressors (Hunchback (Hb), Krüppel (Kr), Giant (Gt), Knirps (Kni), Tailless (Tll)). The protein sequence of these transcription factors is highly conserved between the two species (~85% amino acid identity), in particular, the DNA binding residues are identical (A. Morozov, pers. comm, and Morozov & Siggia [30])), suggesting that any evolutionary change in the network arises primarily from the divergence of *cis*-regulatory sequence. We compare orthologous modules (orthology defined by alignment, see Methods) between *D.mel* and *D.pse* at the levels of sequence, binding site composition and *in vivo* expression, as well as the expression of the endogenous genes. We find that the expression of the maternal and gap factors and those of their target genes show no or very mild differences in spatio-temporal distribution, yet the *cis*-regulatory modules driving this expression show massive sequence change. While noted previously for individual modules[11, 31, 32], this striking discrepancy between functional conservation and molecular divergence of regulatory sequence is not well characterized or understood and begs the question: *how is functionality maintained in the face of such high sequence-level change?* The network-wide analysis we have undertaken here provides intriguing answers.

Sequence changes

The alignment of regulatory sequence between the two species reveals substantial divergence, with a typical salt-and-pepper alternation of aligned and unaligned stretches. In the orthologous segmentation modules, only 30-60% of the sequence is in conserved blocks (≥10bp, ungapped, ≥70% identity; see Methods) (Fig. S1 and Supplementary file 2). Module lengths differ by about 16% (median) between the two species (Fig. 1a). A significant fraction of module sequence consists of inexact tandem repeats of 5-10bp in length and with 2-3 copies (median 14% in *D.mel* and 23% in *D.pse*); in both species, tandem repeat coverage is substantially higher in unaligned sequence than in conserved blocks[33] (Fig. S2), thus implicating tandem repeats as potential carriers of sequence change[34]. 8 of the 39 modules show substantial (>300bp) indels (insertions or deletions) – 1 insertion into and 5 deletions from *D.mel*, and 2 insertions in *D.pse*, detected using *D. virilis* as outgroup. The notable bias towards indels that reduce the relative size of the *D.mel* modules was confirmed by three way *D.mel/D.ananassae/D.pse* comparisons (see Supplement 3).

Binding site composition changes

We used the Stubb algorithm[20, 35-38] to predict the binding site composition of sequence-orthologous modules. This algorithm computes the most likely binding free energy between all of the given regulatory factors and a segment of regulatory DNA, using no factor-specific parameters. It weights strong and weak sites in accordance with their binding affinity, and returns a fractional occupancy for each predicted site, as well as an integrated "profile value" for each factor representing its total site content. (See Supplementary file 5 for assessment of our binding site prediction with this method. Also see http://veda.cs.uiuc.edu/evodevo-supp/windowfits/index.html for a sample output.) Overall binding site content of the module is measured by the "free energy" value returned by the program. Additionally, Stubb may be run on aligned sequence from two species, in which case it places binding sites in homologous sequence blocks according to an evolutionary model, and consistently scores sites in unaligned sequence. Since our model of evolution presumes consistent function, the quality of the fit (positive "synergy", see Methods) is a measure of overall functional conservation.

Using the Stubb output, we used two distinct methods for defining the change in binding site composition between orthologous modules. The first, called "netchange I", computes for each factor the difference in integrated profile values, (see Methods) and reveals a very substantial site-level change, with a median of 39% (Fig. 1b). The second method, "netchange II", conservatively predicts loss or gain of sites block by block as defined by the sequence alignment (see Methods) and yields lower values of net change (median 16%), but a similar overall distribution (Fig. 1b). Binding site change does not correlate well with sequence-level change (Fig. 1d), and is only slightly more likely to occur in nonaligned sequence than in conserved blocks (Fig. S1). A summary of our expression data, Fig 1d, shows that the free energy is a much better measure of module expression change than is sequence conservation. We point out here that the above two measures reflect the net difference in binding site contents of orthologous modules, and neither captures any "movement" of binding sites within modules.

We ran Stubb on the original set of 39 modules and selected 16 modules representing various levels of binding site change and types of molecular change to investigate whether such changes translate into differences in module expression. To this end, orthologous modules from *D.mel* and *D.pse* were fused to a *lacZ* reporter and examined in *D.mel*; thus input factor distributions are held constant, and any differences in expression are directly attributable to differences in the *cis*-regulatory sequence (Methods). For each tested module, we attempted to relate observed expression changes (or lack thereof) to computational observations of binding site-level change, such as difference in free energy (Stubb's measure of binding site number and strengths, see Methods), "netchange I and II" values, presence of insertions (or deletions) with predicted binding sites, and computational prediction of significant changes in binding site content for individual transcription factors.

Compensatory change within modules

The *kni_(-1)* orthologous modules (derived from *kni_kd*[18]) show an abrupt sequence change in the form of a 386 bp insertion in *D.mel*, which contains multiple binding sites, for the factors Hb and Gt, both of which are known to repress knirps expression[18, 39] However, the overall binding site content of the orthologs is similar (Fig. 1b), with the only significant change (Fig. 2) being the number of DStat sites (0.9 in *D.pse* and 0.1 in *D.mel*). Accordingly, both modules drive correct expression in the *kni* posterior domain (Fig. 3a and Fig. 4a). When the above-mentioned insertion is deleted from the *D.mel* module, its expression is significantly broadened (Fig. 3a “kni_(-1)_del” and Fig. 4a), consistent with the loss of repressor sites. The *D.mel* insertion thus contains functionality that was lost elsewhere in the module; such a case of compensatory change within a sequence-orthologous module has previously been described for *eve_stripe 2* [40].

The phenomenon of compensatory loss and gain has been termed as “binding site turnover” by Moses *et al.*[41], but there are few experimental validations of a compensatory effect such as that observed above. We followed up on this observed effect by computationally assessing the extent of compensatory loss and gain in all 39 modules. This phenomenon was statistically significant in 17 cases, very close to the ~16 expected by chance at the significance level used. (Supplementary File 6.)

Binding site dispersal into adjacent sequence

The *hairy_stripe_5 (h_5)* modules[42] show a marked change in the free energy (*D.mel*>*D.pse*) and binding site composition (Fig. 1b and Fig. 3b, left panel). There is significant difference in binding site content for DStat, Bcd, Kr, and Kni (Fig. 2). Consistent with these differences, the *D.mel* ortholog drives a prominent *h* stripe 5 (as well as weak ectopic anterior and posterior expression)[42], while expression of the *D.pse* ortholog is severely muted (Fig. 3b and Fig. 4b). Given that in *D.mel* the *hairy_stripe_1 (h_1)* module is directly adjacent, we asked whether in *D.pse* binding site content required for stripe 5 formation may in part have shifted to this neighboring region. Using the Stubb free energy profile, we re-delineated the modules to include this adjacent free energy peak and found that the extended modules of both species nicely produce *h* stripes 1 and 5 (Fig. 3b and Fig. 4c).

A similar, less dramatic, case is *even skipped_3+7 (eve_3+7)*[43]: The two orthologs show very similar binding site composition (Fig. 1b), and none of the transcription factors undergo significant change in binding site content (Fig. 2). Both modules drive a strong and correctly positioned stripe 3, but the *D.mel* module produces a weak and overly narrow stripe 7, while the *D.pse* module produces a strong and overly wide stripe 7 (Fig. 3c and Fig 4d). Inspection of the Stubb free energy profiles (Fig. 3c, left panel) reveals a substantial broadening of the *eve_3+7* peak in *D.pse* compared to *D.mel*. We re-delineated the modules to fully encompass the free energy peaks of both species and found that both extended modules produce a stripe 3+7 pattern with largely correct positioning (Fig. 3c and Fig. 4e). (These two re-delineated modules also show low net change; Supplement 1.) Figure 1b classifies eve_3+7 as the exceptional case of expression change with no sequence change, potentially because we scored just the original (improperly delineated) modules.

These two examples show that, to a remarkable extent, binding sites are able to redistribute along the DNA and disperse beyond the boundaries defined by orthologous seqence. We analysed this phenomenon systematically by separately delineating modules in the two species based on Stubb free energy profiles (see Methods). These species-specific 'functional' delineations reveal differences in the length of equivalent modules, with a median difference of 26%, which is markedly higher than the length variation resulting from simply aligning orthologous sequence (Fig. 1a). Thus, flexibility in module length and positioning provides an additional mechanism for preserving functional output despite evolutionary sequence-level change. Neighboring sequence is readily populated with binding sites and co-opted into the module, revealing the fleeting nature of the association of binding sites with specific DNA segments.

Redundancy between modules

The *gt_(-1)* modules[19] show a marked difference in free energy (*D.mel*>*D.pse*) and binding site composition (Fig. 1b and Fig. 3d, left panel), with significant difference in binding site content for DStat and Tll. The *D.mel* ortholog drives strong correct expression in the two main *gt* domains, while the *D.pse* ortholog drives expression only in a slightly widened anterior domain; endogenous gene expression is the same in both species (Fig. 3d). Interestingly, the *gt_(-1)* module is redundant with two other modules in the *gt* control region: the adjacent *gt_(-3)*, which drives posterior expression, and *gt_(-10)*, which drives anterior expression[19]; these two modules drive proper expression in both species (Fig. 3d). Thus, (partial) redundancy between modules permits evolutionary modification of the composition and expression of individual modules without altering overall gene expression. To our knowledge, this is the first documented case of a segmentation module losing/gaining an entire expression domain.

The *sloppy2_(-3) (slp2_(-3))* modules[19] present a similar case. The two orthologs differ markedly in their free energy (*D.pse*>*D.mel*) and binding site composition (Fig. 1b**)**, due in part to a very large (505 bp) insertion in *D.pse* that carries both activator (TorRE, Bcd, Cad) and repressor (Hb, Kni) sites. Correspondingly, the *D.pse* module shows much stronger and earlier expression (Fig. 4f). The endogenous gene expression of *slp2* (and *slp1*) are identical between the two species (Fig. 4f), suggesting that *D.mel* receives functional compensation elsewhere; indeed two additional modules driving the same expression have been identified[22].

Concomitant change in module and gene expression

The regulatory region of Kr presents a complex case. In addition to its central domain, Kr is dynamically expressed in secondary patterns at the anterior and posterior tip of the embryo. These secondary patterns, which arise later in the blastoderm and are incompletely mapped to regulatory sequence[44], show subtle differences between *D.mel* and *D.pse*. We compared the three known blastoderm modules related to Kruppel and detected strong binding site-level change for the two modules driving the anterior secondary pattern (*Kr_CD2_AD1, Kr_AD2*, Fig. 1b and Fig. 2). Both modules show modest differences in expression between *D.mel* and *D.pse* that are consistent with those observed in the endogenous gene pattern (Fig. 3e and Fig. 4f,g). Interestingly, some binding site content of the *Kr_CD2_AD1* module appears to have shifted into the neighboring *Kr_CD1* module in *D.pse*, resulting in expression of this module in an anterior cap that is not present in the *D.mel* ortholog (Fig. 3e). If the *D.pse Kr_CD1* module is truncated in accordance with the Stubb free energy profile, this additional anterior expression is lost (Fig. 3e). (The removed sequence has predicted binding sites for the anterior activator Bcd, as well as sites for Hb and Kni.)

The *nubbin_(-2) (nub_(-2))* module of *D.mel*[19] has a significant number of Kni repressor sites outside of conserved blocks which are absent in *D.pse* (Fig. 5b and Fig. 2). The *D.mel* module correctly drives expression in a posterior band that quickly resolves into two stripes (Fig. 5a). The expression domain of Kni coincides with the nub interstripe (Fig. 5c); RNAi reveals that Kni is indeed responsible for inter-stripe repression (Fig. 5a). With the *D.pse* module, stripe formation is slower and much less pronounced (P-value 0.0002, see methods), consistent with the absence of Kni sites (Fig. 5a,c). Interestingly, this change in module expression is mirrored in the endogenous gene expression, and thus not compensated elsewhere (P-value 0.0499, Fig. 5a,c). In RNAi-mediaed *kni*$^{-}$ experiments, the difference between the *D.mel* and *D.pse* modules is absent (P-value 0.26).

Expression change correlates with informatic measures of change

Overall, we find that 9 of the 16 tested orthologous modules show an appreciable difference in expression between *D.mel* and *D.pse.* We can now ask which molecular features correlate with such change in expression. While sequence conservation correlates poorly (Fig. 1d), "netchange" of total binding site content (by both measures) correlates well with change in expression (Fig. 1b), supporting the notion that sequence conservation is not a good indicator for function, but binding site content is: of the 5 tested modules with lower predicted binding site-level change, 4 show no change in expression (eve_1, gt_(-10), gt_(-3), kni_(-1)). Of the 11 tested modules with higher predicted change, 8 show a corresponding change in expression pattern. (Fisher's exact test p-value 0.077.) (The three that do not show expression change are h_6, gt_(-6) and kni_(-5).) Low levels of "synergy", which assesses evolutionary conservation of binding site content by comparing the free energy scores of single-species and two-species Stubb runs (see Methods), are also highly predictive (Fisher's exact test p-value 0.054) of expression change (Fig. 1c). Cases where we observe high binding site-level change without change in expression may reflect module robustness, with binding site change below the phenocritical threshold, or over-prediction of site change. Stubb's prediction of binding site clustering and module position along the sequence, represented in the free energy profile, has proven remarkably precise. In cases where module boundaries were initially defined by *D.mel*-based experimentation, redelineations of homologous modules based on Stubb free energy profiles dramatically improve the expression pattern (*eve, h, Kr*).

Conclusions

Our combined computational and experimental analysis of segmentation modules, carried out under uniform criteria and at network-wide scale, has made possible the detection of both rare events and subtle global trends, thus affording unprecedented new insight into the evolution of regulatory sequence. Our results reveal a remarkable plasticity in the transcriptional regulation of segmentation genes and trace some of the compensatory mechanisms that ensure conservation of gene expression despite large changes at the molecular level. We find that about 30% (16% by more conservative measure) change in binding site content is readily tolerated without affecting module expression, providing a quantitative measure of module robustness. Note however that these quantifications are specific to the evolutionary divergence between the two species studied here. Redistribution of sites along the sequence, and thus shifting of module boundaries beyond those defined by sequence orthology, is a recurring phenomenon. Typically, for each transcription factor multiple binding sites are dispersed at variable distances throughout the module, which diminishes the importance of individual sites and of their spacing relative to one another. We also observe significant evolutionary modulation of functionally redundant modules, specifically in the regulatory regions of the gap genes. Notably, these (partially) redundant modules are either adjacent (*Kr*[44]; also *tll*[45]) or split between a distal region and one proximal to the basal promoter (*gt*[19]). These features of intra- and inter-module redundancy and plasticity permit change in regulatory sequence without affecting the overall expression of the gene. Interestingly, the few uncompensated changes we observe (*nub, Kr*) do not affect primary expression domains, but rather secondary patterns in the late blastoderm[46], whose functional significance for the embryo has not been established. It is therefore unclear whether these changes represent genuine newly emerging functions or inconsequential epi-phenomena of sequence-level change.

The very high rate of sequence divergence between *D.mel* and *D.pse*, which also includes length polymorphisms, cannot be driven by point mutations alone. Apart from the relatively rare occurence of large indels, the pervasive presence of tandem repeats, particularly in the non-aligned sequence, suggests that short-range duplication/deletion and point mutations jointly produce the bulk of the sequence change[33, 34]. The observed repeat length of 5-10 bp not only suits DNA topology but also matches the typical size of transcription factor binding sites. However, the fact that the tandem repeats are not preferentially associated with the maternal and gap factor sites suggests that duplication/deletion, presumably by replication slippage, is a generic mutagenic mechanism[47, 48] that creates new sequence, which can then be turned into binding sites by point mutation[49]. Such slippage events may facilitate not only the local re-distribution of binding site content, but also larger scale recombination events such as module duplication. Similar mechanisms, i.e. repeat expansion and gene duplication, have been proposed to account for the rapid divergence of protein sequence in larger-scale evolutionary transitions[50-52]. It thus appears that regulatory regions, less constrained by sequence content and spacing rules, are able to explore the sequence space very rapidly, with multiple intrinsic compensation mechanisms, such as the ones described here, ensuring the (near-) neutrality of most changes.

Methods

Sequence analysis. We analysed 39 non-redundant segmentation modules with validated and (largely) faithful expression from the collection in Schroeder et al.[21] (Supplement 1), using release 3 coordinates of the *D.melanogaster* genome[53]; orthologs were extracted from the February 2003 assembly of the *D.pseudoobscura* genome (Baylor Human Genome Sequencing Center). Sequences were aligned using LAGAN[54] (parameters "-mt 1 –ms -2 –gs -6 –gc 0"), ungapped blocks of ≥10bp and ≥70% sequence identity were defined as conserved blocks, the remainder as unaligned/non-conserved sequence[55]. In case a module boundary falls outside a conserved block, equal distance to the nearest included block was used to delimit the *D.pse* module. The precise criteria for defining aligned blocks do not significantly affect our assessment of binding site changes (data not shown). Tandem repeat coverage in segmentation modules was assessed using Tandem Repeats Finder[56], with parameters "2 3 5 80 10 25 500 -m –d", and Mreps[57], with parameters "-fasta -res3 -minperiod 3", as described in Sinha and Siggia[33]. All quantitative data on sequence-level change, tandem repeat coverage, binding site content and binding site change are collected in Supplement 2.

Stubb algorithm and prediction of free energy of modules. Position weight matrices (PWMs) for the transcription factors Bicoid, Caudal, Hunchback, D-Stat, Giant, Krüppel, Knirps, and Tailless, and for the torso response element (torRE) were obtained from[19, 20]. The Stubb program[35] computes the likelihood of a sequence being generated by a probabilistic model that samples binding sites from the input PWMs and normalizes it against a suitable background model (first order Markov model trained on the sequence). The normalized likelihood is called the *free energy* of a given sequence and provides a measure of the density and strength of binding sites in the sequence, without free parameters. The *free energy profile* is the result of moving a sliding window of 500 bp along the sequence in 50 bp increments and plotting the free energy score for each window. For the free-energy-based or functional delineation of segmentation modules, we concatenated overlapping windows above a free energy threshold of three standard deviations above the genome-wide mean, with slight adjustment for *D.pse* to produce equal average module length (7.95 for *D.mel* and 8.8 for *D.pse*; at these thresholds, not all modules are recovered in both species, therefore n=29). Stubb can also be run in two-species mode, producing a compound free energy score based on a probalistic model of binding site evolution, with sites in blocks up- or downweighted depending on their degree of conservation. The difference between this multi-species free energy score and the sum of the single-species scores, termed ‘synergy’, provides a measure of binding site conservation in a given module, with high values signifying strong conservation.

Computation of binding site content of modules and “netchange”. Stubb predicts both the position and strength (fractional occupancy) of binding sites. The binding site composition of modules is determined by adding, for each transcription factor, the fractional occupancy values of all sites above a threshold (0.1), termed the *integrated profile value* for that factor. Binding site-level change between orthologous modules in *D.mel* and *D.pse* can then be computed as the sum of the absolute values of the differences between the integrated profile values for all nine input factors, and expressed as a fraction of the total binding site content of the module in both species (‘netchange I’). This measure permits compensation of site loss by site gain elsewhere in a module, and is thus a measure of the net difference in binding site composition. It ignores changes in site position.

For a more conservative measure of binding site-level change that disregards differences between the species in site strength, we defined a binding site as lost if it is present (fractional occupancy > 0.1) in one species but absent at the orthologous position in the other (for intra-block sites), or if it is present in one species and not matched by a corresponding site within the unaligned sequence of the other (for out-of-block sites). Net site loss is then calculated separately for each factor based on the fractional occupancy values of all lost sites, and summed over all factors to produce the net change for the module (‘netchange II’). For the purposes of this calculation, sites predicted by both single-species and two-species Stubb were considered. Stubb was run with a first order Markov background trained on over 800 Kbp of intergenic sequence around segmentation genes

Predicting significant change in binding site content for individual transcription factors
For each module M, and for each transcription factor T, the integrated profile value $\sigma(M, T)$ was computed as above to measure the number of binding sites of T in the sequence. The raw score $\sigma(M, T)$ was then *normalized* as follows: 500 random sequences were generated from the background model, of the same length as M, and the score σ was computed for each, thereby producing a null distribution. The mean and standard deviation of this distribution were used to obtain the *normalized* score $\sigma_N(M, T)$. For each pair of orthologous modules M_{mel} and M_{pse}, the normalized scores $\sigma(M_{mel}, T)$ and $\sigma(M_{pse}, T)$ were computed, and modules that scored above a threshold $\tau_1 = 3$ in one species, and below a threshold $\tau_2 = 2$ in the other species were reported

as having changed significantly with respect to motif transcription factor T. We also performed the entire exercise with a different measure of binding site content, the free energy differential (defined below), with τ_1 = 2 and τ_2 = 1. For each measure of change, we tried three different background models (Markov order 0, 1 and 2), for a total of six methods. We then considered only those changes that were predicted by at least two different methods. *Definition of free energy differential:* For a particular transcription factor T, the free energy differential is the difference in the free energy computed by Stubb, when run with all PWM's and when run with all PWMs *but T*. This represents the contribution of PWM T to the overall free energy of the module.

Analysis of expression patterns. Module expression was analysed as described[19], except that *exactly* delineated genomic DNA fragments were generated by secondary PCR and cloned into *hs43GAL*[58]. A Fasta file with the primers and cloned regions is available in Supplement 1. Modules of both species were tested in *D.mel*; for each construct, at least three independent insertions were analysed. Endogenous gene expression patterns for both *D.mel* (*w^{1118}*) and *D.pse* (*14011-0121.94*, Tucson stock center) were determined using *D.mel* RNA probes. RNAi was carried out as described[59]. To generate the quantitative expression profiles in Figures 4 and 5, we used an automated image processing tool that detects embryo boundaries and measures averge pixel intensity in blocks arrayed along the dorsal and/or ventral peripheries and projected onto the center line (=longest distance along the anteroposterior axis) such that they divide it into 200 segments of equal length. For Figure 5, the resulting profile curves were normalized and averaged over 8-16 embryos per genotype.

Statistical significance of change in *nub_(-1)* or *nub* endogenous expression pattern

Each expression profile has two peaks of expression separated by a "cleft". The average height h_p of the peaks and the height h_c of the lowest point in the cleft are calculated (with baseline expression as the origin). The ratio h_c/h_p is used to measure the extent of the dip in gene expression. This statistic is collected from 17 experimental replicates representing *nub_(-1)* in *D.mel* and 14 replicates from *nub_(-1)* in *D.pse* and the two samples are compared with a one-tailed t-test. Similarly, 5 replicates from each of *D.mel* and *D.pse* in *kni^{-}* are compared. Eight replicates from *D.mel* and 9 replicates from *D.pse* are used to compare endogenous gene expression. (Raw data available in Supplementary Materials.)

Supplementary Information

Supplement 1. Sequence information for all 39 *D.mel* and *D.pse* segmentation modules used in the study and 4 modules extended/shortened based on functional delineation, in fasta format.

Supplement 2. Quantitative data on sequence-level change, tandem repeat coverage, binding site content and binding site change in segmentation modules.

Supplement 3. Analysis of large (>300 bp) indels in segmentation modules.

The gbrowse display of free energy profiles for genome-wide Stubb runs can be viewed at http://edsc.rockefeller.edu/cgi-bin/gbrowse_nature/cgi-bin/gbrowse?source=fly3; windowfit displays giving a graphical representation of binding site position and strength for all 39 orthologous modules can be accessed at http://veda.cs.uiuc.edu/evodevo-supp/

Supplement 4. Raw data used to assess statistical significance of changes in nubbin and *nub_(-1)* expression (Fig. 5).

Supplement 5. Assessment of sensitivity of binding site prediction. (See legend within.)

Supplement 6. Methods and results for statistical analysis of the extent of binding site turnover.

Supplementary Figure S1. Histogram depicting the fraction of sequence, binding sites, and binding site-level change falling in conserved bloc0ks in orthologous modules in *D.mel* and *D.pse*; medians are indicated by triangles.

Supplementary Figure S2. Histogram showing the fractions of total sequence, conserved blocks, and binding sites covered by tandem repeats in *D.pse* modules.

Acknowledgements

We thank Steve Small for the gift of cDNAs and members of the lab for comments on the manuscript. This work was supported by grants from the NIH (E.D.S and U.G.), NSF (E.D.S), and a W.M. Keck Foundation fellowship (S.S.).

Figures

Figure 1. Quantitative analysis of sequence and binding site change in 39 orthologous segmentation modules. a. Histogram showing distribution of module length differential and module shift for sequence-orthologous and functionally delineated modules in *D.mel* and *D.pse*, expressed as fraction of total module length in *D.mel.* **b**. Scatter plot comparing two measures of total binding site-level change showing that either measure serves to classify those modules that change expression. The dashed lines are medians **c**. Scatter plot showing that binding sites are slightly more likely to reside in conserved blocks, and the lack of correlation between binding site and sequence-level change. **d.** The fraction of sequence in conserved sequence blocks does not correlate with expression changes, while the change in free energy score or synergy (Methods) works much better.

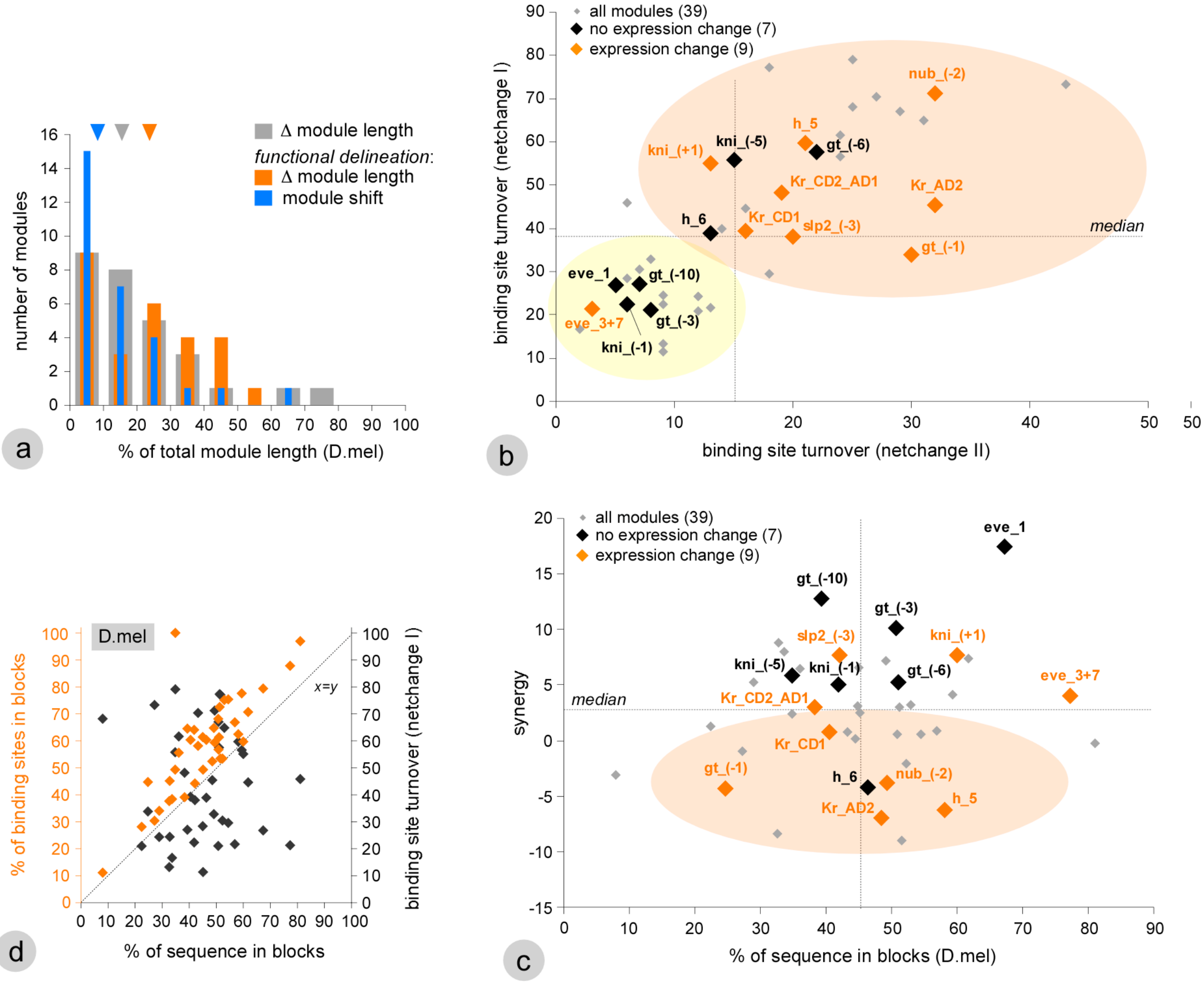


Figure 2. Significant changes in binding site content for individual transcription factors. Six different methods were used to predict if a transcription factor's binding site content changes between orthologous modules, for each of the 16 experimentally tested modules. The number of methods that report a significant change is shown (minimum of 2). Green: more sites in *D.mel*, Yellow: more sites in *D.pse.*

module	change	at least two methods								
		Dstat	**TorRE**	Tll	Bcd	**Hb**	**Gt**	**Kr**	Kni	**Cad**
eve1	no									
gt_(-10)	no									
gt_(-3)	no									
kni_(-5)	no				2	2				
h_6	no								2	
kni_(-1)	no	3								
gt_(-6)	no			5				6	2	
Kr_CD2_AD1	anterior domain stronger in Dpse			4	2	3		4		
eve_3+7	Stripe 7 stronger, broader in Dpse									
Kr_AD2	Dmel expr at posterior (amnioserosa)			6						
kni_(+1)	Dpse more native-like; Dmel aberrant								2	2
slp2_-3	Bolder anterior domain in Dpse									2
Kr_CD1	anterior tip only in Dpse					4	2			
nub_(-2)	quicker resolution of 2 strps in Dmel					4			4	4
h_5	weaker in Dpse	2			4			6	5	
gt_(-1)	weaker in Dpse	2		2						

Figure 3. Experimental analysis of segmentation modules. Expression of orthologous segmentation modules, as revealed by reporter gene fusions (*module-basal promoter-lacZ*), and of the endogenous genes, *D.mel* (blue frame, top) and *D.pse* (green frame, bottom). Notable differences between the two species are indicated by arrowheads and discussed in the text. Panels on the left depict the genomic regions surrounding the genes, with single-species Stubb free energy profiles (*D.pse* profiles are projected onto *D.mel* coordinates based on LAGAN alignments), and the position of modules. Modules shown are highlighted in dark grey, modules redelineated based on Stubb free energy profiles in orange. a. *kni* b. *h* c. *eve* d. *gt* e. *Kr* f. *slp*2.

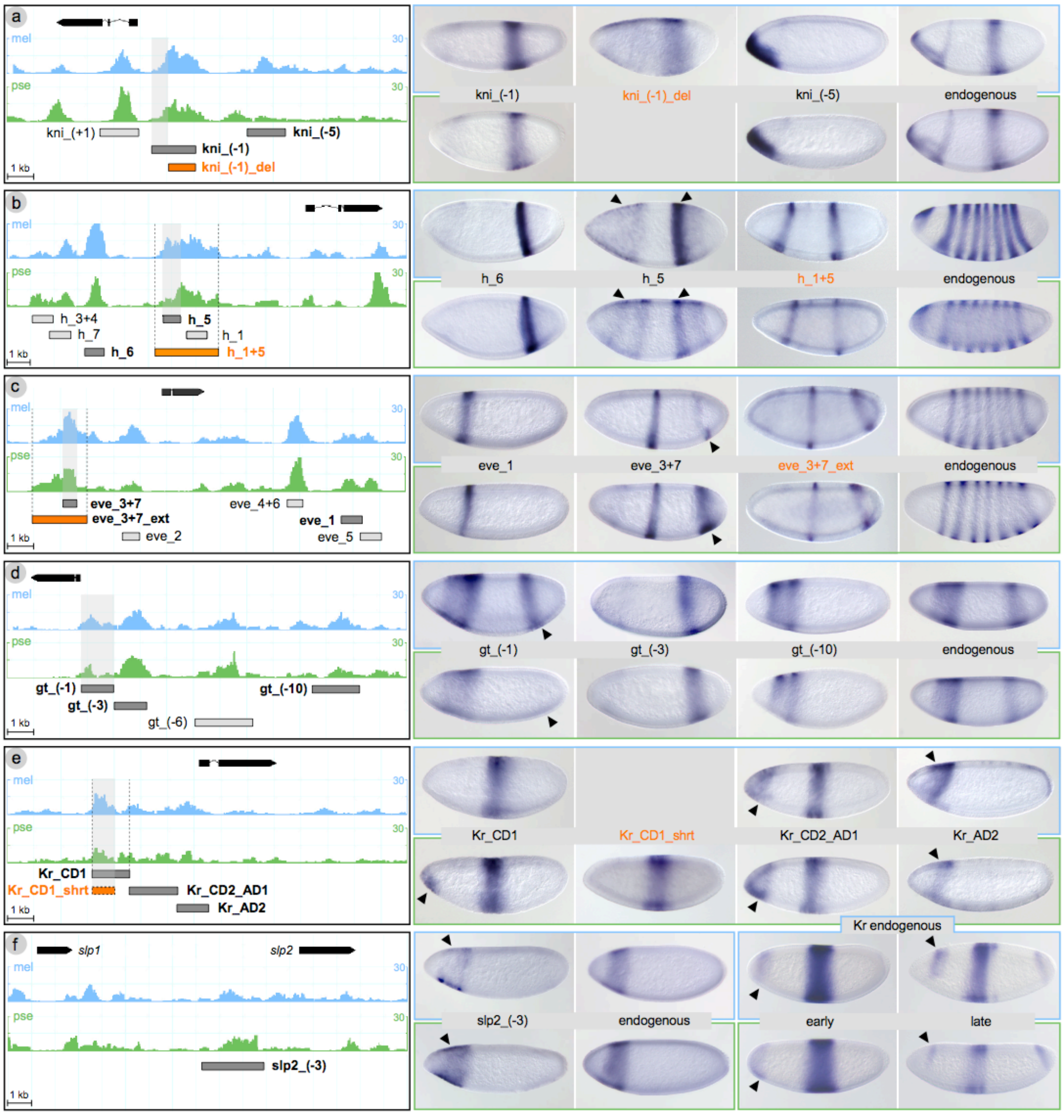

a
mel
30
pse
30
kni_(+1)
kni_(-5)
kni_(-1)
kni_(-1)_del
1 kb
kni_(-1)
kni_(-1)_del
kni_(-5)
endogenous
b
mel
30
pse
30
h_3+4
h_5
h_7
h_1
h_6
h_1+5
1 kb
h_6
h_5
h_1+5
endogenous
c
mel
30
pse
30
eve_3+7
eve_3+7_ext
eve_4+6
eve_1
eve_2
eve_5
1 kb
eve_1
eve_3+7
eve_3+7_ext
endogenous
d
mel
30
pse
30
gt_(-1)
gt_(-3)
gt_(-10)
gt_(-6)
1 kb
gt_(-1)
gt_(-3)
gt_(-10)
endogenous
e
mel
30
pse
30
Kr_CD1
Kr_CD1_shrt
Kr_CD2_AD1
Kr_AD2
1 kb
Kr_CD1
Kr_CD1_shrt
Kr_CD2_AD1
Kr_AD2
Kr endogenous
f
slp1
slp2
mel
30
pse
30
slp2_(-3)
1 kb
slp2_(-3)
endogenous
early
late

Figure 4. Graphical representation of module expression profiles. These were generated as described in Methods. The quantification was done along either the dorsal or the ventral periphery (projected onto the central line) or along the central line; for each panel, all curves were extracted in the same manner.

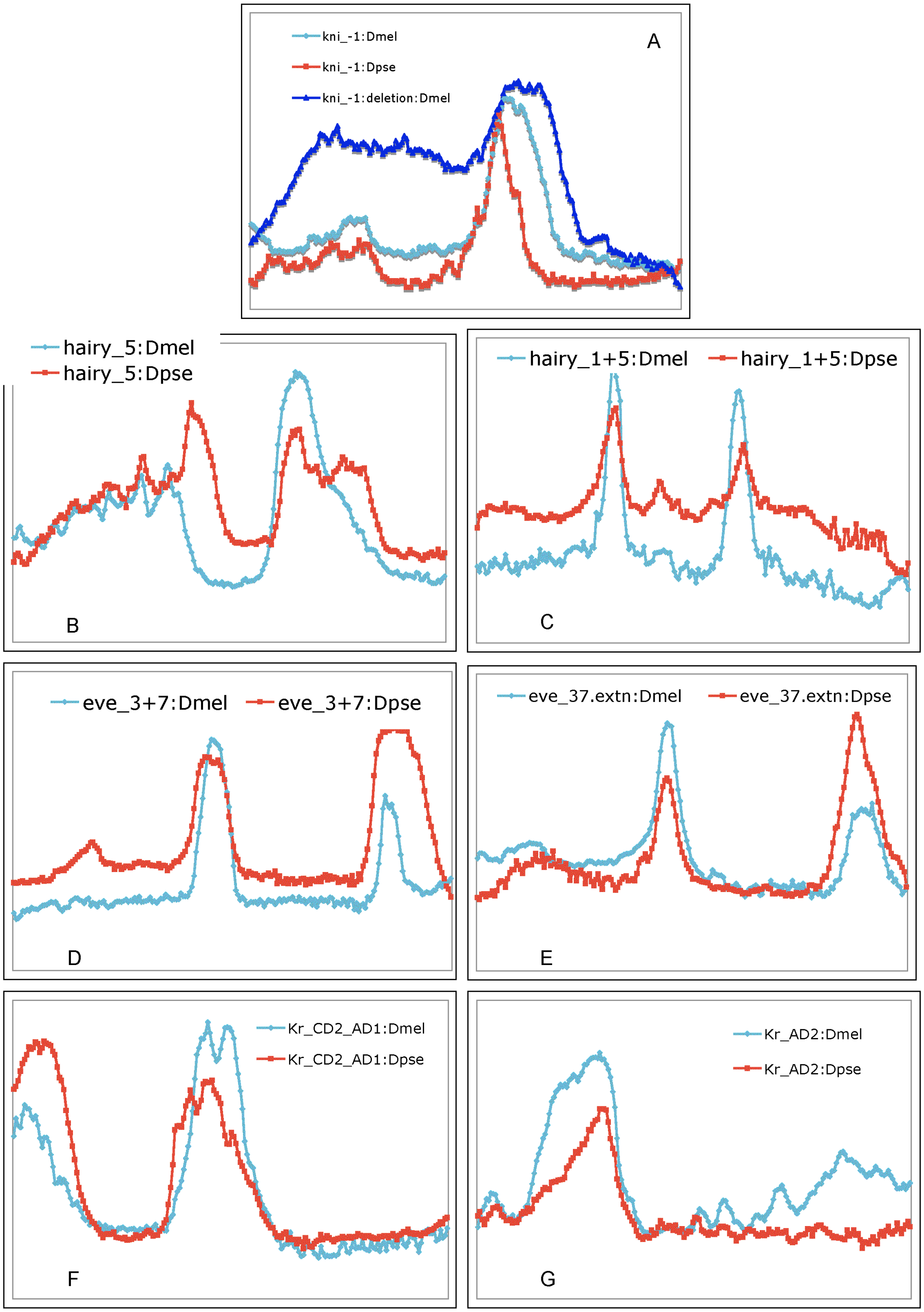
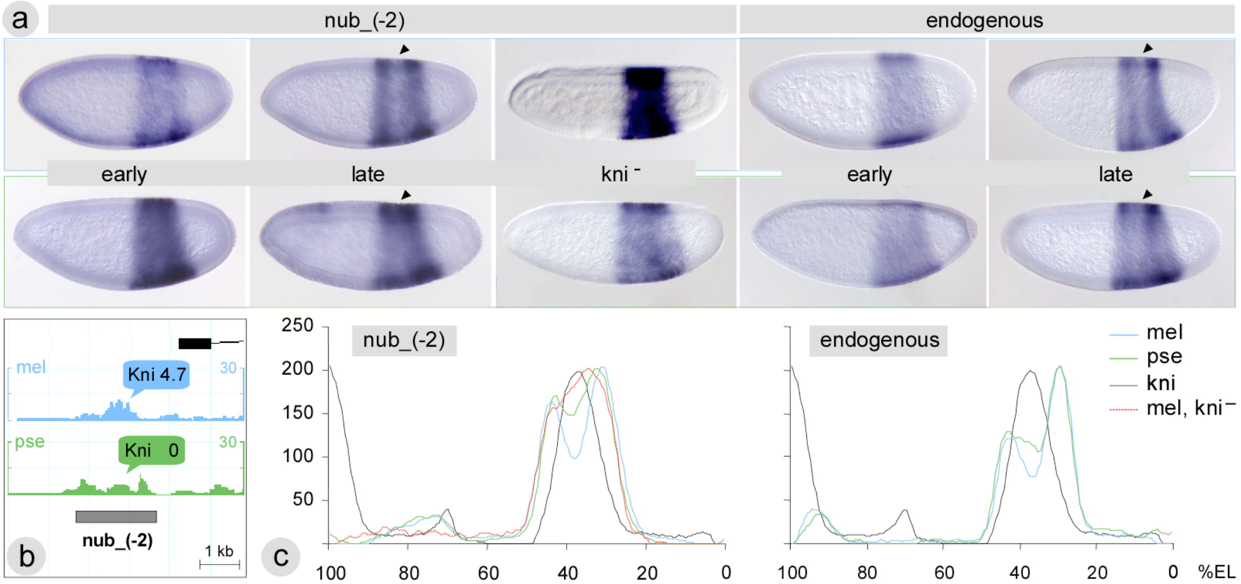

kni_-1:Dmel
kni_-1:Dpse
kni_-1:deletion:Dmel
A
hairy_5:Dmel
hairy_5:Dpse
B
hairy_1+5:Dmel
hairy_1+5:Dpse
C
eve_3+7:Dmel
eve_3+7:Dpse
D
eve_37.extn:Dmel
eve_37.extn:Dpse
E
Kr_CD2_AD1:Dmel
Kr_CD2_AD1:Dpse
F
Kr_AD2:Dmel
Kr_AD2:Dpse
G

Figure 5. Nub module and gene expression. a. Expression of the *nub_(-2)* module and the nub gene in mid and late blastoderm, *D.mel* (top) and *D.pse* (bottom), showing resolution of the contiguous domain into two distinct stripes in *D.mel* but not in *D.pse* or under kni loss of function conditions (RNAi). b. Genomic region surrounding the *nub_(-2)* module with Stubb free energy profiles and predicted Kni profile value (=number/strength of Kni binding sites). c. Expression profiles for *nub_(-2)* module and *nub* gene in late blastoderm, based on measurements of 8-16 embryos per genotype (see Methods). The profiles are almost identical, but clearly show *D.mel* forming a strong interstripe that coincides with the expression of the repressor Kni and is largely absent in *D.pse*, consistent with lack of Kni binding sites.



References

1. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. From DNA to diversity (Blackwell Science, Malden, MA, 2001).
2. Davidson, E. H. Genomic regulatory systems (Academic Press, San Diego, CA, 2001).
3. Simpson, P. Evolution of development in closely related species of flies and worms. Nat Rev Genet 3, 907-17. (2002).
4. Wray, G. A. et al. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20, 1377-419. Epub 2003 May 30. (2003).
5. Averof, M. & Patel, N. H. Crustacean appendage evolution associated with changes in Hox gene expression. Nature. 388, 682-6. (1997).
6. Cohn, M. J. & Tickle, C. Developmental basis of limblessness and axial patterning in snakes. Nature. 399, 474-9. (1999).
7. Kopp, A., Duncan, I., Godt, D. & Carroll, S. B. Genetic control and evolution of sexually dimorphic characters in Drosophila. Nature. 408, 553-9. (2000).
8. Skaer, N., Pistillo, D. & Simpson, P. Transcriptional heterochrony of scute and changes in bristle pattern between two closely related species of blowfly. Dev Biol. 252, 31-45. (2002).
9. Sucena, E. & Stern, D. L. Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. Proc Natl Acad Sci U S A. 97, 4530-4. (2000).
10. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. Nature. 430, 85-8. (2004).
11. Bonneton, F., Shaw, P. J., Fazakerley, C., Shi, M. & Dover, G. A. Comparison of bicoid-dependent regulation of hunchback between Musca domestica and Drosophila melanogaster. Mech Dev. 66, 143-56. (1997).
12. Shaw, P. J., Wratten, N. S., McGregor, A. P. & Dover, G. A. Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. Evol Dev. 4, 265-77. (2002).
13. Ludwig, M. Z. et al. Functional evolution of a cis-regulatory module. PLoS Biol. 3, e93. Epub 2005 Mar 15. (2005).
14. Hersh, B. M. & Carroll, S. B. Direct regulation of knot gene expression by Ultrabithorax and the evolution of cis-regulatory elements in Drosophila. Development. 132, 1567-77. (2005).
15. Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. & Carroll, S. B. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature. 433, 481-7. (2005).
16. Wray, G. A. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8, 206-16 (2007).
17. St Johnston, D. & Nusslein-Volhard, C. The origin of pattern and polarity in the Drosophila embryo. Cell 68, 201-19. (1992).
18. Pankratz, M. J., Busch, M., Hoch, M., Seifert, E. & Jackle, H. Spatial control of the gap gene knirps in the Drosophila embryo by posterior morphogen system. Science. 255, 986-9. (1992).
19. Schroeder, M. D. et al. Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol 2, E271. Epub 2004 Aug 31. (2004).
20. Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U. & Siggia, E. D. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics 5, 129. (2004).

21. 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. Epub 2004 Aug 20. (2004).
22. Ochoa-Espinosa, A. et al. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A 102, 4960-5. Epub 2005 Mar 25. (2005).
23. Lynch, J. & Desplan, C. 'De-evolution' of Drosophila toward a more generic mode of axis patterning. Int J Dev Biol 47, 497-503. (2003).
24. Lynch, J. A., Brent, A. E., Leaf, D. S., Pultz, M. A. & Desplan, C. Localized maternal orthodenticle patterns anterior and posterior in the long germ wasp Nasonia. Nature. 439, 728-32. (2006).
25. Goltsev, Y., Hsiong, W., Lanzaro, G. & Levine, M. Different combinations of gap repressors for common stripes in Anopheles and Drosophila embryos. Dev Biol 275, 435-46. (2004).
26. Zdobnov, E. M. et al. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science 298, 149-59. (2002).
27. Richards, S. et al. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res 15, 1-18. (2005).
28. Clark, A. G. et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, 203-18 (2007).
29. Stark, A. et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, 219-32 (2007).
30. Morozov, A. V. & Siggia, E. D. Connecting protein structure with predictions of regulatory sites. Proc Natl Acad Sci U S A 104, 7068-73 (2007).
31. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development. 125, 949-58. (1998).
32. Dover, G. How genomic and developmental dynamics affect evolutionary processes. Bioessays 22, 1153-9 (2000).
33. Sinha, S. & Siggia, E. D. Sequence turnover and tandem repeats in cis-regulatory modules in drosophila. Mol Biol Evol 22, 874-85. Epub 2005 Jan 19. (2005).
34. Tanay, A. & Siggia, E. D. Sequence context affects the rate of short insertions and deletions in flies and primates. Genome Biol 9, R37 (2008).
35. Sinha, S., van Nimwegen, E. & Siggia, E. D. A probabilistic method to detect regulatory modules. Bioinformatics 19, i292-301. (2003).
36. Sinha, S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics 22, e454-63 (2006).
37. Sinha, S., Ling, X., Whitfield, C. W., Zhai, C. & Robinson, G. E. Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees. Proc Natl Acad Sci U S A 103, 16352-7 (2006).
38. Sinha, S., Adler, A. S., Field, Y., Chang, H. Y. & Segal, E. Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res 18, 477-88 (2008).
39. Eldon, E. D. & Pirrotta, V. Interactions of the Drosophila gap gene giant with maternal and zygotic pattern-forming genes. Development 111, 367-78 (1991).
40. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403, 564-7. (2000).
41. Moses, A. M. et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2, e130 (2006).

42. Langeland, J. A. & Carroll, S. B. Conservation of regulatory elements controlling hairy pair-rule stripe formation. Development. 117, 585-96. (1993).
43. 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, 413-22 (1989).
44. Hoch, M., Schroder, C., Seifert, E. & Jackle, H. cis-acting control elements for Kruppel expression in the Drosophila embryo. Embo J 9, 2587-95. (1990).
45. Rudolph, K. M. et al. Complex regulatory region mediating tailless expression in early embryonic patterning and brain development. Development 124, 4297-308. (1997).
46. Tautz, D. & Nigro, L. Microevolutionary divergence pattern of the segmentation gene hunchback in Drosophila. Mol Biol Evol. 15, 1403-11. (1998).
47. Lovett, S. T. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol Microbiol 52, 1243-53. (2004).
48. Sinden, R. R., Hashem, V. I. & Rosche, W. A. DNA-directed mutations. Leading and lagging strand specificity. Ann N Y Acad Sci 870, 173-89 (1999).
49. Dermitzakis, E. T., Bergman, C. M. & Clark, A. G. Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol. 20, 703-14. Epub 2003 Apr 2. (2003).
50. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 579, 3346-54. Epub 2005 Apr 8. (2005).
51. Ohno, S. Evolution by gene duplication (Springer, New York, NY, 1970).
52. Taylor, J. S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38, 615-43. (2004).
53. Celniker, S. E. et al. Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol 3, RESEARCH0079. Epub 2002 Dec 23. (2002).
54. Brudno, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13, 721-31 (2003).
55. Emberly, E., Rajewsky, N. & Siggia, E. D. Conservation of regulatory elements between two species of Drosophila. BMC Bioinformatics 4, 57. (2003).
56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-80 (1999).
57. Kolpakov, R., Bana, G. & Kucherov, G. mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 31, 3672-8 (2003).
58. Thummel, C. S. & Pirrotta, V. Technical Notes: New pCaSperR P-element vectors. Drosophila Information Newsletter 2, available at http://flybase.bio.indiana.edu/docs/news/DIN/dinvol2.txt (1991).
59. Kennerdell, J. R. & Carthew, R. W. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. Cell 95, 1017-26. (1998).