What Does It Take to Evolve an Enhancer? A Simulation-Based Study of Factors Influencing the Emergence of Combinatorial Regulation

Thyago Duque[1] and Saurabh Sinha[1,2,*]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign

[2]Institute for Genomic Biology, University of Illinois at Urbana-Champaign

*Corresponding author: E-mail: sinhas@illinois.edu.

Accepted: May 5, 2015

Abstract

There is widespread interest today in understanding enhancers, which are regulatory elements typically harboring several transcription factor binding sites and mediating the combinatorial effect of transcription factors on gene expression. The evolution of enhancers poses interesting unanswered questions, for example, the evolutionary time taken for a typical enhancer to emerge or the factors shaping its evolution. Existing approaches to *cis*-regulatory evolution have often ignored the combinatorial nature and varied biochemical mechanisms of gene regulation encoded in enhancers. We report on our investigation of enhancer evolution through the use of PEBCRES, a framework for evolutionary simulation of enhancers that employs a mechanistic and well-supported sequence-to-expression model to assign fitness to the evolving enhancer genotype. We estimated the time necessary to evolve, from genomic background, enhancers capable of driving complex gene expression patterns similar to those involved in early development in *Drosophila*. We found the time-to-evolve to range between 0.5 and 10 Myr, and to vary greatly with the target expression pattern, complexity of the real enhancer known to encode that pattern, and the strength of input from specific transcription factors. To our knowledge, this is the first estimate of waiting times for realistic enhancers to evolve. The in silico evolved enhancers had, with a few interesting exceptions, site compositions similar to those seen in real enhancers for the same patterns. Our simulations also revealed that certain features of an enhancer might evolve not due to their biological function but as aids to the evolutionary process itself.

Key words: evolution, gene regulation.

Introduction

The evolution of regulatory sequences is an active area of research today, with important questions such as the following: 1) How long does it take for regulatory sequences to evolve under various assumptions? (Stone and Wray 2001; Carter and Wagner 2002; Gerland and Hwa 2002; MacArthur and Brookfield 2004; Durrett and Schmidt 2007, 2008), 2) do specific regulatory sequences display signatures of positive selection or negative selection? (Moses et al. 2004; Moses 2009; He et al. 2011), 3) what is the evolutionary history of a particular regulatory sequence? (Francois et al. 2007; Josephides and Moses 2011), and 4) how to mathematically model the evolution of regulatory sequences? (Berg et al. 2004; Kim et al. 2009; Nourmohammad and Lässig 2011).

The time scale for evolutionary changes in regulatory systems is an important question that has puzzled biologists for many years, especially since it was first noticed that the general organization of regulatory sequences can be maintained for tens of millions of years (Damjanovski et al. 1998; Ludwig et al. 1998) (also see review in Maeso et al. 2013), despite evidence that functional differences can evolve over significantly shorter time scales (Ross et al. 1994), and sequence comparisons showing that transcription factor (TF) binding sites could appear and disappear among closely related species and even within a population (Damjanovski et al. 1998; Segal et al. 1999). Such observations led Stone and Wray (2001) to ask the following question as a first step towards solving this puzzle: "What time period would be required for new transcription factor binding sites to evolve . . . as a consequence of local point mutations . . . under the

assumption of neutral evolution?" (Stone and Wray 2001). They estimated that new binding sites can emerge due to point mutations alone on extremely short time scales, for example, about 24 years in *Drosophila* or about 6,000 years in humans, even in the absence of selection. The work by Stone and Wray opened a debate over the time scales necessary for the emergence of single binding sites. Durrett and Schmidt (2007) revisited the question while accounting for dependencies present in the population due to common descent, and found that the average time for an 8-bp binding site to appear in humans can range between as little as 60,000 years and as much as 650 Myr, depending on whether a perfect match is required or a fuzzy match suffices, and on whether a partial match pre-existed.

Although these initial studies theorized (Stone and Wray 2001; Durrett and Schmidt 2007) about the evolution of a single binding site, the time scale of major *cis*-regulatory innovations is a more involved question, because in reality the function and evolution of binding sites depends on their context (Duque et al. 2014). *Cis*-regulatory modules (CRMs, also called enhancers) are approximately 1-kb long sequences that typically harbor multiple binding sites for one or more TFs and mediate the combinatorial influence of those TFs on gene expression (Davidson 2010). The emergence of *cis*-regulatory functionality is strongly linked to the appearance of CRMs, and needs to be studied in this context. This is an important issue because, as Stone and Wray (2001) point out, it is unlikely that the dozens of binding sites present in, say, typical *Drosophila* CRMs could emerge one by one, under neutral evolution, without older binding sites getting destroyed. The significance of studying binding site evolution in context is also exemplified by the work of Carter and Wagner (2002), who considered the phenomenon of compensatory mutations leading to a turnover of binding sites (Ludwig et al. 1998; Sinha and Siggia 2005; Moses et al. 2006; Swanson et al. 2011; Kim et al. 2013). They showed the need to explicitly model the evolution of pairs of binding sites, instead of single binding sites, in order to explain relative rates of binding site turnover in invertebrate and vertebrate populations (Aparicio et al. 1995; Ludwig et al. 1998; Swanson et al. 2011) (supplementary note S1, Supplementary Material online). Similarly, Durrett and Schmidt (2008) specifically examined the case of pairs of sites and estimated that in *Drosophila* a pair of mutations can inactivate a binding site and activate another on the time scale of several million years, consistent with empirical observations (Ludwig et al. 1998; Swanson et al. 2011).

Although the above-mentioned studies were important steps in the right direction, they were not meant to tackle the more ambitious questions of how fast can entire CRMs evolve and what factors might influence this "time-to-evolve." CRMs are usually composed of more than a pair of binding sites; as noted above, they harbor several binding sites for multiple TFs that exhibit different roles (e.g., activator or repressor) and potency (strong or weak) and are expressed in different spatiotemporal domains. In fact, the creation of a CRM to perform a specific regulatory function cannot generally be reduced to the emergence of predetermined numbers and types of TF binding sites. Comparison of orthologous CRMs (Kim et al. 2009; Swanson et al. 2011) and quantitative modeling of CRM function (Zinzen et al. 2006; He et al. 2010) suggests considerable flexibility in the "*cis*-regulatory code," in that the same function can be achieved by different combinations (types, numbers, and strengths) of binding sites. Perhaps due to these complexities, there does not exist a computational model capable of estimating the time necessary to evolve a realistic CRM involving multiple TFs with distinct roles, under a range of population genetic and mechanistic assumptions. This is the gap that we attempt to bridge in this work.

We considered approximately 30 bona fide CRMs involved in anterior–posterior (*A/P*) axis specification in the blastoderm stage *Drosophila* embryo and asked how long it might take for these CRMs or other CRMs of comparable functional complexity to appear under strong positive selection for the expression pattern encoded by them. We then explored a variety of factors that might influence this time-to-evolve. Our analyses were enabled by a flexible, simulation-based model of CRM evolution, called PEBCRES (He et al. 2012; Duque et al. 2014). It relies on a state-of-the-art sequence-to-expression model called GEMSTAT (He et al. 2010) that can predict the spatial expression pattern driven by an arbitrary CRM-length sequence, given sufficient information about the trans context. The PEBCRES framework then uses a specialized function (Samee and Sinha 2013; Duque et al. 2014) to compare the predicted gene expression pattern with an ideal expression pattern, and thus estimate the fitness of the evolving sequence. The fitness value then plugs into a standard evolutionary simulation. Previous work has demonstrated the accuracy of GEMSTAT for modeling A/P patterning CRMs (He et al. 2010; Samee and Sinha 2013, 2014), thereby lending credibility to inferences based on its use in assigning fitness values for CRM genotypes in this study. Furthermore, in recent work by Duque et al. (2014), we used PEBCRES to model the evolution of A/P patterning CRMs under negative selection and showed that it can 1) explain rates of conservation and loss of binding sites within CRMs and 2) correctly predict mechanistic properties such as cooperative DNA binding by specific TFs. This previously demonstrated success of PEBCRES (He et al. 2012; Duque et al. 2014) justifies its use as a model to study the appearance of complex developmental CRMs during evolution.

We estimate that CRMs of considerable complexity, for example, CRMs that have the information required to drive A/P patterns in the early *Drosophila* embryo, may evolve in as little as approximately 0.5 Myr, but this time-to-evolve can range widely, with some expression patterns requiring up to 30 times longer. Our simulations suggest that the time-to-evolve depends in part on the complexity of the combinatorial

logic encoded by the CRM, and may be strongly influenced by the requirement for binding sites of a single TF. In silico evolved CRMs by and large resemble the corresponding *Drosophila melanogaster* CRM, with a few interesting exceptions where CRMs evolved in our simulations appear to be more parsimonious than their natural counterparts. To account for uncertainties in our model, we repeat our experiments under varying assumptions about (parameter values of) the underlying population genetics and mechanistic models. In doing so we find that insertions and deletions seem to be more disruptive than constructive, in a manner consistent with our previous work (He et al. 2012; Duque et al. 2014). We also find that changes to the selection strength and mutation rate have their expected effects: Decreasing either parameter increases the time-to-fit estimates. Finally, we note that activators that are uniformly expressed across all modeled cell types may reduce the time-to-evolve, even though their regulatory function need not be essential to the target gene's expression pattern. We speculate that this may be an example of evolution of evolvability (Wagner and Altenberg 1996), in which a genotypic feature evolves not due to its functional effect, but due to its effect on the ability of DNA to evolve more quickly.

Materials and Methods

Procedure for Selecting Expression Patterns for Simulation

To select the 28 patterns used in our experiments reported in figures 1–5, we repeated the following procedure on each of the 37 expression patterns predicted by GEMSTAT (denoted by "EP"), noting that these are the same expression patterns whose evolution was previously modeled using PEBCRES (Duque et al. 2014). First, we generated a set of random sequences that are fed into our fitness function. The fitness of each sequence was calculated using EP as the target expression pattern. If the average fitness of the random sequences was above 0.2, EP was not considered further. In other words, we chose to work with 28 expression patterns for which a random sequence has low fitness. At the same time, we know that the selected expression patterns are "achievable" in our framework because in each case we have a real CRM for which GEMSTAT predicts that expression pattern. We refer to each of these expression patterns by the name of the CRM from *D. melanogaster* that generated the pattern. The 28 expression patterns used for our experiments are shown supplementary figure S1, Supplementary Material online.

Fitness Estimation and Selection Scale

In our PEBCRES simulations, the GEMSTAT model was run with six TFs—BICOID (BCD), CAUDAL (CAD), KRUPPEL (KR), HUNCHBACK (HB), KNIRPS (KNI), and GIANT (GT)—as regulators, and configured to include self-cooperativity for BCD, CAD, and KNI, exactly as in the experiments described in Duque et al. (2014). The fitness estimation in PEBCRES compares the GEMSTAT-predicted expression for a sequence with the target expression profile to assign a numeric fitness between 0 and 1 to that sequence. This numeric fitness is used by a Wright–Fisher simulator to decide the evolutionary fate of arbitrary sequences in the population.

More specifically, PEBCRES uses the weighted Pattern Generating Potential (wPGP) function (Samee and Sinha 2013; Duque et al. 2014), which assigns fitness to a genotype g with predicted pattern p given a target pattern t (Samee and Sinha 2013; Duque et al. 2014). After computing wPGP(p, t), it computes a "fitness functional" as $f(g) = [max(0, wPGP(p, t))]^2$ and then converts this to a fitness value as $F(g) = 1 + Kf(g)$, where K is a free parameter called the selection scale. The wPGP function has a range between −1 and +1, which is mapped to a fitness functional between 0 and 1, as in Duque et al. (2014) and Samee and Sinha (2013). The selection scale parameter (K) may be interpreted as the selection coefficient (typically denoted by s in the literature) in the case where two competing genotypes (a and b) have fitness $F(a) = 1$ and $F(b) = 0$.

Because the fitness functional $f(g)$ (and consequently the fitness $F(g)$) is dependent on the assumptions of the sequence-to-expression model, it is unviable to determine the value of the selection scale experimentally. Instead, we rely on data to find the best fit value of K. This was done by Duque et al. (2014) and we use the value obtained there. Duque et al. (2014) also offer a detailed explanation of the wPGP function and its advantages over alternative functions (such as SSE [Sum of Squared Errors] or correlation), as well as a detailed explanation of the relationship between the selection scale K and the selection coefficient s.

Time Rescaling

To speed up simulation time, we use time rescaling (Hoggart et al. 2007) as in our previous work (He et al. 2012; Duque et al. 2014). The time-rescaling procedure enables the simulation of λt generations of unscaled time in t generations of simulated time by scaling the population size by a factor of $1/\lambda$ and the selection coefficient and mutation rate by a factor of λ, keeping the population genetics parameters $2N\mu$ and $4Ns$ (population-scaled mutation rate and selection coefficient, respectively) unchanged.

Unless otherwise stated, we use time-scaling factor $\lambda = 1{,}000$, time-scaled population size $2N = 1{,}000$, and mutation rate $\mu = 10^{-5}$. These parameters correspond to an unscaled population size of $2N = 10^6$ and mutation rate of $\mu = 10^{-8}$, both within estimated ranges from the literature ($2N \sim 10^5$–10^6 [Thornton and Andolfatto 2006] and $\mu \sim 10^{-9}$–10^{-8} [Drake et al. 1998]). As mentioned in the previous section, we do not explicitly model a selection coefficient s, but rather use a selection scale parameter K, which is similarly scaled by λ.

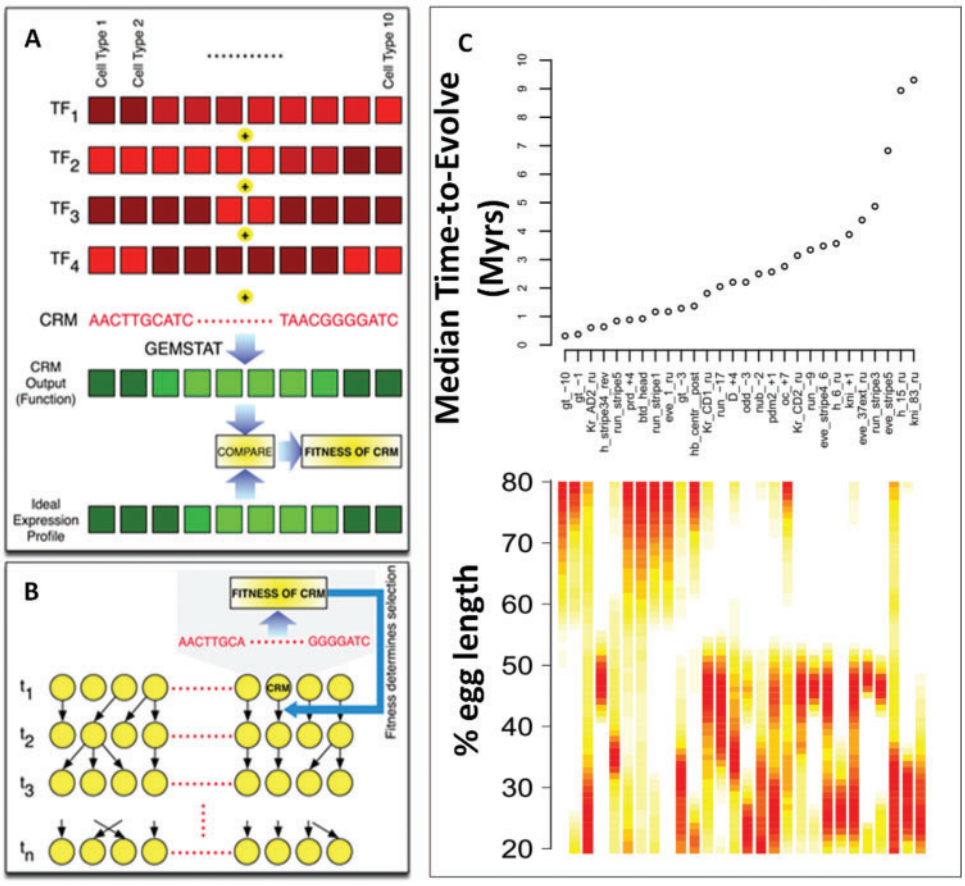

Fig. 1.—Estimating the time necessary for a CRM to evolve. (*A*, *B*) Methodology. A schematic representation of the PEBCRES framework describing how it is used to estimate the time necessary for a CRM to evolve from genomic background. Expression readout of the evolving CRM is predicted using GEMSTAT, producing a fitness value (*A*), which is then plugged into a Wright–Fisher simulation with selection (*B*). (*C*) Top panel: Time-to-evolve estimates (*y* axis), in Myr, for each of the 28 target expression patterns (*x* axis). Bottom panel: A representation of the 28 A/P expression patterns that serve as target patterns in our simulations, sorted by time-to-evolve estimate (same order as in top panel). Each expression pattern is represented by a column in the heatmap, with red representing high expression and white representing absent expression. The anterior end of the embryo is at the top and posterior end at the bottom. Only 20–80% egg length interval is shown.

To convert time-to-evolve estimates from number of generations to time in years, we first calculate the median of the number of generations (t) until a sufficiently fit phenotype emerges in the population (the median is calculated among all simulations of the evolution of the same expression pattern). This is equivalent to λt generations in real time. Finally, we multiply this number by the average length (in years) of a generation. In the case of *Drosophila*, this number is approximately 0.027 years per generation or 10 days per generation.

Occupancy Calculation

Computation of TF occupancy uses the efficient dynamic programing implementation of GEMSTAT (He et al. 2010) to calculate the relative probability of each binding site being in the bound state, called the fractional occupancy of a binding site. The probability here refers to the Boltzmann distribution over all configurations of sites in the sequence being bound or free, following a statistical mechanics treatment. We compute the occupancy of a TF by summing the fractional occupancy of all putative binding sites for that TF in the given sequence. The procedure to calculate the fractional occupancy is described in He et al. (2010), and this method of estimating TF occupancy was also used in our earlier work (He et al. 2012).

Procedure for Selecting Starting Sequences

By default, our simulations begin with a random sequence. In the section "Dependence on initial conditions and the possibility of exaptation," we report on two sets of experiments where the initial sequence was not random. Here, we first calculate the correlation between the expression patterns of

every pair of CRMs and simulate the evolution of sequences targeting the expression pattern of one CRM while initializing the population with the sequence of another CRM. In one set of experiments, we evolved an expression pattern from a sequence that already drove a similar pattern, by choosing the initial sequence for each simulation randomly from one of the 5 CRMs whose expression is most correlated with the target expression pattern. In the other set of experiments, we evolved a pattern from a sequence that drove a very different pattern, by choosing the initial sequence randomly from one of the five CRMs whose expression is most anticorrelated with the target expression pattern.

Procedure for Comparing Simulations with Different GEMSTAT Model Specifications

Our evolutionary simulations require a model of CRM function, which is provided by GEMSTAT, to help define the fitness function. The model in GEMSTAT can be specified to include or exclude a specific regulator, and once a regulator is added to the model, all parameters are learnt from appropriate training data. Our goal was to compare evolutionary simulations made with two different specifications of the GEMSTAT model: One that includes a ubiquitous activator and one that does not. However, there are a couple of concerns to be addressed before such comparisons can be made.

Recall that the baseline model, that is, the model without the universal activator, is used to define the target expression pattern of a simulation. Specifically, as noted in the section "Uniformly expressed activators can speed up emergence of CRMs," we take a *D. melanogaster* CRM, use the baseline model to predict its expression pattern (say "*T*"), and use this pattern as the target of a PEBCRES simulation. This ensures that the simulation is using a fitness function such that there is at least one sequence with perfect fitness. Now, we may train a new GEMSTAT model (say "M_U") that includes the universal activator, and perform simulations using this new model to define fitness. These simulations must target the same expression pattern (*T*) as before, to make claims about the role of the ubiquitous activator in shaping the evolutionary dynamics. However, there is no guarantee that there exists a sequence with perfect fitness when using the new model M_U. That is, there may not exist a sequence for which model M_U predicts expression pattern *T* exactly. This makes the comparison unfair, because the existence of a perfect solution is only guaranteed for one of the models. An alternative is to run both the simulations (with baseline model or with M_U) with a new target expression pattern (say *T'*), set to be the prediction of M_U on the *D. melanogaster* CRM. This guarantees that the simulations with M_U as fitness function can in principle find a sequence with perfect fitness, but the new pattern *T'* may require the use of the ubiquitous activator and simulations with the baseline model may not have any chance of finding the perfectly fit CRM. Our hypothesis is that ubiquitous activators reduce the time necessary to evolve certain expression patterns, even if the same pattern might have been evolved without utilizing the ubiquitous activator. To test this hypothesis, we need a setup where the fitness function includes regulatory input by a ubiquitous activator but the latter is not necessary for a solution to have high fitness. To this end, we use the experimental setup described below:

1. Start with the baseline GEMSTAT specification (TFs: BCD, CAG, GT, HB, KNI, KR; self-cooperativity for BCD, CAD, KNI).
2. Train on all 37 CRMS an alternative GEMSTAT specification that includes a ubiquitous activator (either DSTAT or ZLD). All other assumptions of the baseline model are maintained. The alternative specification should be trained to match the expression patterns predicted by the baseline model. This will result in an alternative model (say M_U) whose predicted expression for each of the 37 CRMs is very close to the predictions from the baseline model.
3. For each CRM, merge the predicted expression patterns from the baseline model and from M_U by taking their average. The merged expression pattern is thus equally achievable by either model.
4. Repeat the experiment to determine median time-to-evolve per CRM using the baseline model as the fitness function, but targeting the merged expression pattern. This is only done for the 28 CRMs shown in figure 1.
5. Repeat the experiment to determine median time-to-evolve per CRM using M_U as the fitness function, again targeting the merged expression pattern.
6. Compare median time-to-evolve per CRM for simulations from steps 4 and 5.

Our experimental setup still does not guarantee that there exists a solution with fitness of 1 during simulations, but manual inspection assured us that in each simulation, whether it uses the baseline model or M_U, there is at least one sequence with fitness approximately 1 with respect to the target expression pattern defined as above. We repeated the above procedure for two alternative models of M_U, the first one including ZLD and the second including DSTAT as the additional ubiquitous activator.

Results

Overview of Simulations

We used the PEBCRES simulation framework (He et al. 2012; Duque et al. 2014) (fig. 1*A* and *B*) to evolve sequences that drive a predetermined expression pattern, simulating the process of evolutionary adaptation under a variety of scenarios. The main simplifying features of a PEBCRES simulation are the following: 1) A constant-sized population of 2*N* haploid individuals evolves as per the Wright–Fisher model (Wright 1931; Fisher 1999), 2) each individual's genotype is a DNA sequence 500–2,000 bp long (typical length of a CRM), 3) mutations occur at a fixed rate and independently at each nucleotide,

and 4) no recombination occurs. Selection is modeled so that an individual *i* spawns an expected number of offspring proportional to $1 + KF_i$, where *K* is a constant called the "selection scale" and F_i is the fitness of individual *i* on a scale of 0 (unfit) to 1 (fit). Additional details are in Materials and Methods section and in Duque et al. (2014) and He et al. (2012).

The distinguishing feature of a PEBCRES simulation is its calculation of a fitness value (*F*) for any given CRM-length sequence and a given expression pattern called the "target pattern." The target pattern is prespecified as a (say *M* dimensional) vector of gene expression values on a scale of 0 to 1 (fig. 1*A*). The sequence is mapped to the expression pattern it encodes (also an *M*-dimensional vector) by the statistical thermodynamics-based GEMSTAT model (He et al. 2010). Note that the parameters of GEMSTAT, representing the trans context, are trained before and outside of PEBCRES simulations. (Also see next paragraph for comments about reliability of these parameters.) The predicted expression pattern corresponding to the sequence is then compared with the target pattern by a specialized function called "weighted Pattern Generating Potential" or wPGP (Samee and Sinha 2013; Duque et al. 2014) to produce a fitness value between 0 and 1, which is 1 if and only if the two pattern vectors are identical. (See supplementary fig. S6, Supplementary Material online, for a visual depiction of predicted expression profiles across the spectrum of fitness values.)

To decide on the target expression patterns to use in our study, we considered a set of 37 bona fide CRMs from *D. melanogaster* that drive well-characterized A/P patterns in the blastoderm stage embryo. These 37 CRMs were the subject of a detailed modeling exercise in our previous work (He et al. 2010), and the accuracy of model fits for a majority (see supplementary note S2, Supplementary Material online) of CRMs in that exercise assures us that the genotype-to-phenotype mapping used in PEBCRES simulations here is a reasonable approximation of reality. Furthermore, in Duque et al. (2014) we analyzed the evolutionary changes within these 37 CRMs across the *Drosophila* subfamily (12 sequenced species separated by $\leq$65 Myr) and were able to accurately model these changes using PEBCRES simulations of a functionally constrained CRM. The selection strength on CRMs (i.e., the selection scale parameter *K* mentioned above) estimated in that study as providing the best fits between models and data was used as the default value in this study. We selected 28 of the 37 A/P patterning CRMs as the subject of our analyses (see Materials and Methods for selection criterion), predicted their expression patterns using GEMSTAT, and used these 28 predicted patterns (see supplementary fig. S1, Supplementary Material online), which are in approximate agreement with experimental CRM readouts, as the target patterns in PEBCRES simulations. We will refer to each target expression pattern by the name of the *D. melanogaster* CRM associated with that pattern.

Thus, using a carefully constructed fitness function and with target patterns representing the typical complexity of a developmental CRM, we hoped that our simulations will provide meaningful insights into what it takes to evolve an enhancer.

Estimating the Time to Evolve a CRM

Our first goal was to estimate how long it might take for a typical developmental CRM to evolve from genomic background, under a variety of assumptions. We simulated the evolution of random sequences targeting each of the 28 target patterns (at least 30 simulations for each pattern) and recorded the time-to-evolve for each simulation, that is, the earliest generation in which an individual with fitness above 0.8 emerged in the population (we noted that fixation quickly follows the emergence of a fit genotype). Using ideas from population genetics theory and properly accounting for the time rescaling used by PEBCRES (see Materials and Methods) (Hoggart et al. 2007; He et al. 2012), we converted the time-to-evolve value from generations to an estimate of time in millions of years of *Drosophila* evolution. Finally, we examined the median over all of our simulations for each target pattern. The results of this computational experiment are presented in figure 1*C* and discussed below.

Our simulations predict that, under strong selection, functional CRMs for complex spatial patterns could evolve in surprisingly short evolutionary times. For example, the average time necessary to evolve the pattern for gt_-10, as per our simulations, is only approximately 0.3 Myr. As a point of reference, this is nearly 10 times smaller than the divergence between *D. melanogaster* and *Drosophila simulans* (2.5 Myr [Ranz et al. 2003], synonymous substitution rate of approximately 0.04 [Bedford and Hartl 2008]), predicting that even between these two closely related species there could be lineage-specific CRMs driving simple expression patterns defined by the response to a single TF. (The gt_-10 pattern is mediated by activating sites of the BCD TF.) Other quickly evolving patterns were mostly BCD-driven anterior patterns like gt_-10, but also included more central patterns such as "h_stripe_34_rev" and "run_stripe5" (fig. 1*C*), which are regulated by two or more TFs (supplementary fig. S1, Supplementary Material online).

On the other hand, some target patterns require much longer time to evolve, with the longest time being about 9 Myr (median) for the expression pattern "kni_83_ru," roughly 30 times longer than that for gt_-10. There is a clear trend of anterior patterns to have lower time-to-evolve estimates, while central and posterior patterns have larger estimates (fig. 1*C*, bottom). We noticed that half of the expression patterns have time-to-evolve estimates that are higher than the distance between *D. melanogaster* and *D. simulans* (2.5 Myr, the closest of the currently sequenced species; Ranz et al. 2003)), while all the patterns have time-to-evolve that is

shorter than the divergence between *D. melanogaster* and *Drosophila yakuba* (13–17 Myr; Satta et al. 1987; Satta and Takahata 1990). This suggests an opportunity for future studies to compare these sequenced genomes, which are amenable to high-quality alignments, for the existence and function of many lineage-specific CRMs. Our theoretical findings are also supported by the recent discovery of hundreds of CRMs (driving expression in *Drosophila* S2 cells) being gained since the *D. melanogaster–D. yakuba* split (Arnold et al. 2014).

Evolutionary Sampling of the Fitness Landscape: Real Versus In Silico Evolved CRMs

We next examined the in silico evolved CRMs (also called "simulated" CRMs below) from the previous section more closely, with a view to gain deeper insights into the "fitness landscape" (Berg et al. 2004) associated with each target expression pattern. Our primary goal was to determine 1) if these simulated CRMs resemble the real *D. melanogaster* CRM associated with the target pattern, as might be expected and 2) whether cases that deviate from this expectation provide clues about shortcomings in our models of CRM function (Duque et al. 2014), reveal signatures of the evolutionary process (He et al. 2012), or suggest multiple optima in the fitness landscape. For this investigation, we chose to describe a CRM by the estimated "occupancy" of each TF in the CRM (see Materials and Methods; He et al. 2012), which is an integrated score reflecting the total number of binding sites, both strong and weak, of that TF (supplementary note S3, Supplementary Material online). It also enables easy comparison of two CRMs for similarity of *cis*-regulatory logic. We compared any two CRMs, real or evolved, by the Euclidian distance between their respective six-dimensional vectors of TF occupancy counts (GEMSTAT modeling was based on six TFs, see Materials and Methods).

We first examined all in silico evolved CRMs for all 28 target patterns and noted that CRMs associated with similar expression patterns are closer to each other than distinctly expressed CRMs (supplementary fig. S2, Supplementary Material online), as expected. We then asked if in silico evolved CRMs for the same target pattern cluster in the vector space, and how tight these clusters are. Table 1 presents two relevant metrics to answer these questions. The first metric, d_{intra}, represents the average distance between any pair of evolved CRM for a particular target pattern, and a second metric, d_{inter}, denotes the average distance between CRMs for a specific expression pattern and CRMs representing other patterns. (We restricted the other patterns to be those that are least correlated with that pattern, because several of the target patterns are highly similar to each other.) As table 1 shows, the ratio d_{inter}/d_{intra} is almost always ≥ 2, indicating that distinct target patterns are associated with well-clustered simulated CRMs. A few examples are depicted in figure 2 (note black circles in each panel), which further confirms this observation.

Table 1

Average Pairwise Distance between CRMs Evolved In Silico for the Same Expression Pattern (d_{intra}) and for Distinct Patterns (d_{inter})

Target Pattern	d_{WT}	d_{intra}	d_{inter}	d_{inter}/d_{intra}	d_{inter}/d_{WT}
h_15_ru†	6.13	2.36	4.66	1.97	0.76
kni_83_ru†	4.35	1.67	3.76	2.25	0.87
h_6_ru†	3.57	0.86	3.71	4.34	1.04
Kr_CD2_ru†	3.44	1.26	3.61	2.86	1.05
run_stripe3†	4.21	2.20	4.52	2.05	1.07
eve_37ext_ru†	4.19	1.68	5.03	3.00	1.20
kni_+1	2.62	1.36	3.54	2.61	1.35
D_+4	2.54	1.10	3.43	3.11	1.35
Kr_CD1_ru	2.40	0.79	3.37	4.29	1.41
run_-9	2.92	1.48	4.13	2.79	1.41
gt_-3	2.39	1.08	3.38	3.13	1.41
gt_-1	2.26	0.73	3.25	4.48	1.44
odd_-3	2.33	1.12	3.59	3.19	1.54
pdm2_+1	2.10	1.10	3.30	3.00	1.57
eve_stripe4_6	2.23	1.03	3.63	3.53	1.63
nub_-2	2.02	1.11	3.40	3.07	1.68
oc_+7	1.92	1.15	3.44	3.00	1.79
h_stripe34_rev	2.44	1.31	4.55	3.48	1.87
run_-17	1.81	1.15	3.86	3.36	2.13
gt_-10	1.51	0.73	3.68	5.03	2.43
hb_centr__post	1.48	1.05	3.59	3.41	2.44
btd_head	1.51	0.65	4.01	6.15	2.67
run_stripe5	1.19	1.14	3.40	2.97	2.86
prd_+4	1.16	0.56	3.51	6.23	3.01
Kr_AD2_ru	1.03	1.08	3.16	2.92	3.07
eve_stripe5	1.07	1.07	3.59	3.35	3.35
run_stripe1	1.04	0.61	4.24	6.90	4.06
eve_1_ru	0.81	0.40	3.77	9.37	4.65

Note.—Patterns marked with † are those where the real CRM falls outside the cluster of simulated CRMs.

We next asked if in silico evolved CRMs for a target pattern are similar to the *D. melanogaster* CRM (henceforth, "real" CRM) associated with that pattern. For this, we calculated a metric, d_{WT}, as the average distance between the real CRM and all simulated CRMs for each target pattern, and compared it with the intercluster distances d_{inter} as well as intracluster distances d_{intra} defined above. A large relative value of d_{WT} indicates that CRMs resulting from evolutionary simulation are different from the real CRM. As table 1 shows, d_{WT} is in most cases slightly larger than d_{intra} but smaller than d_{inter}, indicating that the real CRM falls more or less within the cluster of evolved CRMs for the same expression pattern (fig. 2*A–D*). There were a few interesting exceptions to this trend, marked with a "†" in the table. For example, the in silico evolved CRMs for kni_83_ru and h_15_ru (fig. 2*E* and *F*) seem to be distinctly more parsimonious than the real CRM, although GEMSTAT predicts their functionality to be the same (supplementary note S4, Supplementary Material online). We may speculate on why high occupancy evolved in the real CRM for these patterns. One hypothesis is that the evolutionary history of the real CRMs is more complicated than our

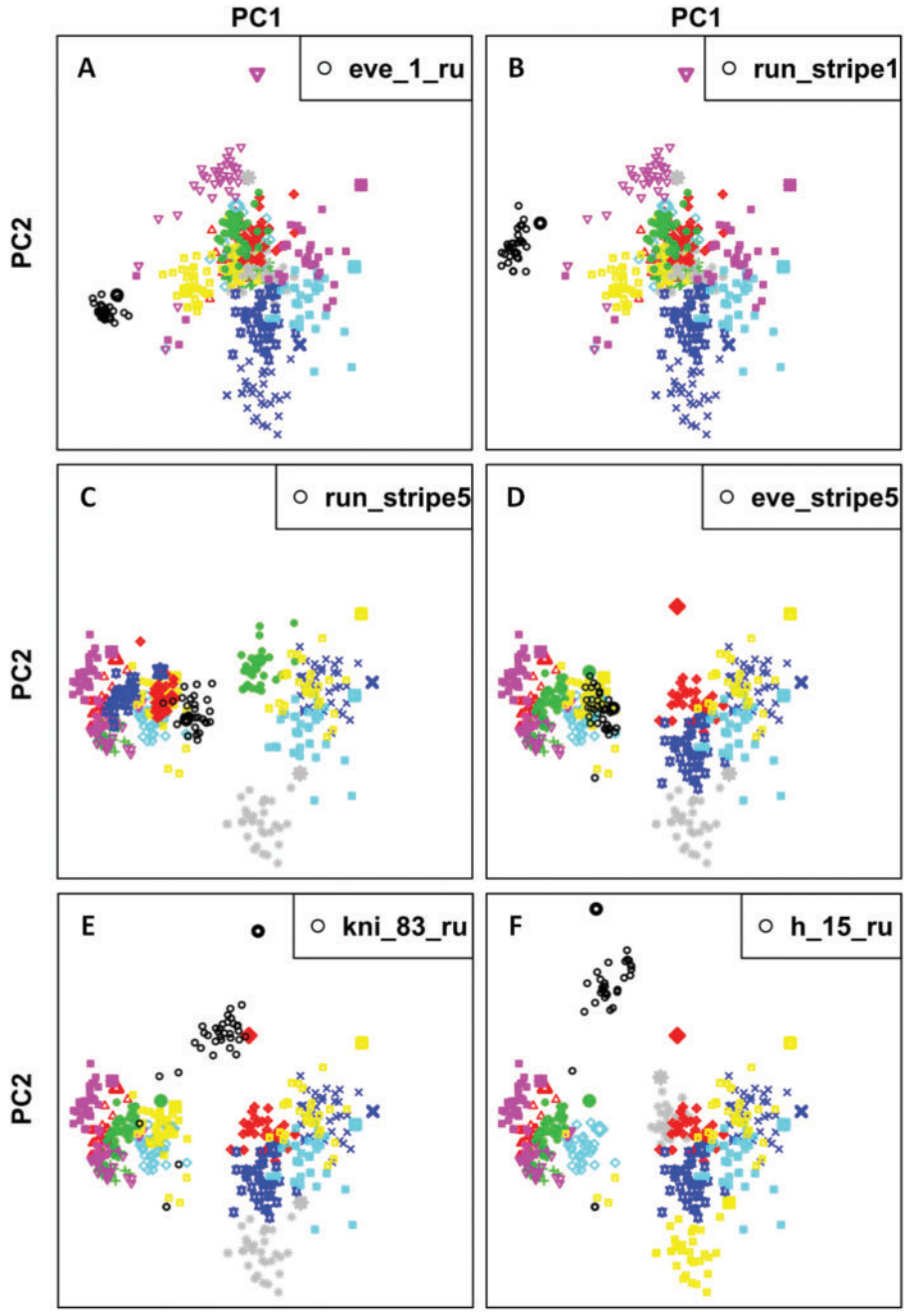

Fig. 2.—Visual representation of real and in silico evolved CRMs. A two-dimensional projection of the six-dimensional "TF occupancy" space occupied by these CRMs. The axes represent the first and second principal components. The panels correspond to CRMs for patterns eve_1_ru (*A*), "run_stripe1" (*B*), "run_stripe5" (*C*), "eve_stripe5" (*D*), kni_83_ru (*E*), and "h_15_ru" (*F*). In each panel, simulated CRMs of respective pattern are shown in small black circles, and the real *Drosophila melanogaster* CRM for that pattern as a larger black circle; points in other colors represent simulated CRMs (smaller icons) and the real CRM (larger icon, same color) for other target patterns.

simple simulations assume, for example, they have been "exapted" (de Souza et al. 2013) from other functional sequences to perform a different function. An alternative possibility is that the high occupancy values seen in the real CRM are functionally necessary due to some unknown mechanism not modeled by GEMSTAT.

Features of CRM Composition May Influence Its Time-to-Evolve

As noted above, time-to-evolve estimates for CRMs of different expression patterns vary greatly, by at least one order of magnitude. We sought to determine the factors that can explain such variability, focusing on two classes of potential

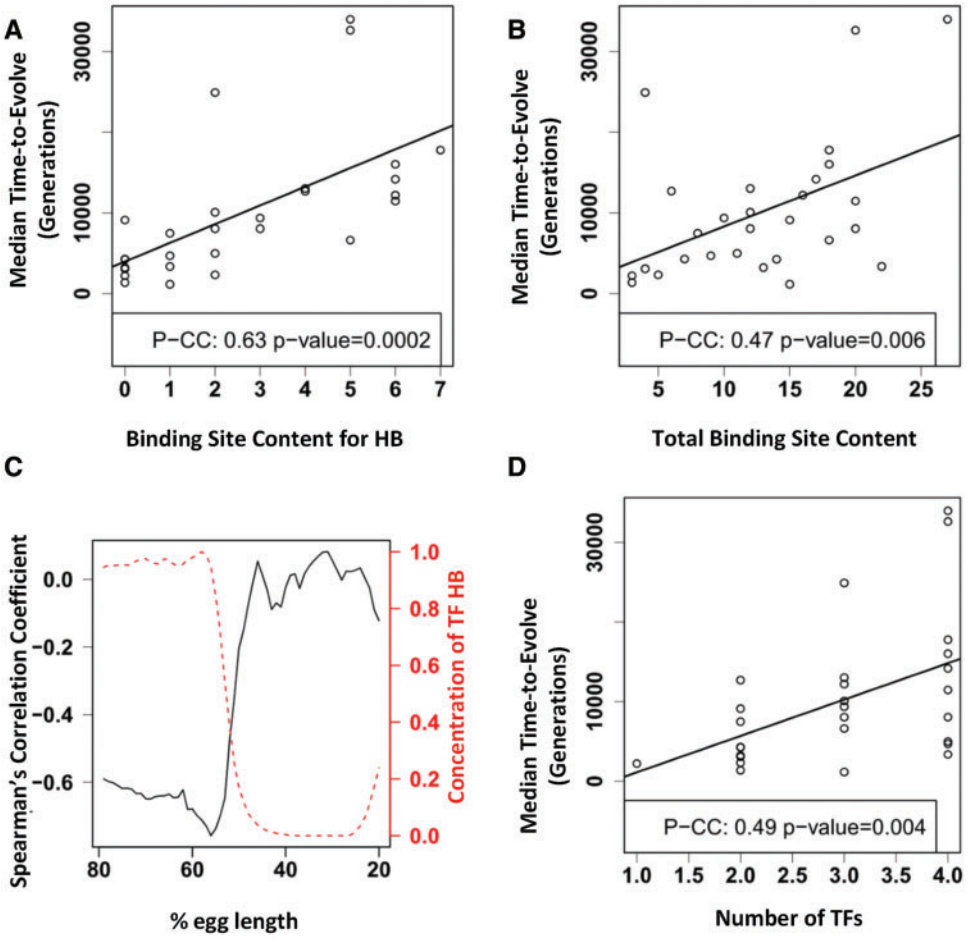

Fig. 3.—Features of CRMs that influence time-to-evolve. (*A*) A scatter plot relating the estimated occupancy of HB in a real CRM (*x* axis) and the median estimated time to evolve a CRM for the corresponding pattern (*y* axis). The Pearson's correlation coefficient between the two variables is of 0.7, which is significant at a *P* value of 1.4×10^{-4}. The best fit line is also shown (solid line). (*B*) A scatter plot relating the estimated TF occupancy in a CRM, summed over all TFs used in the model (*x* axis), and the median estimated time-to-evolve for the corresponding pattern (*y* axis). Pearson CC = 0.47, *P* = 0.005. However, the partial correlation, discounting the contribution of HB sites, is not significant (*P* = 0.45). The best fit line is also shown (solid line). (*C*) Time-to-evolve estimates of CRMs are highly negatively correlated with expression level in anterior parts of the embryo. The *y* axis shows for each position along the *A*/*P* axis ("%egg length," *x* axis) the Spearman's correlation coefficient between a target pattern's expression level at that axial position and the time-to-evolve estimate for that pattern. The concentration profile of HB across the axis is also shown (dashed line). (*D*) Scatter plot relating the number of TFs with at least one binding site present in the *D. melanogaster* CRM (*x* axis) and the median estimated time-to-evolve for the corresponding pattern (*y* axis). The Pearson's correlation coefficient between the two variables is 0.49, *P*-value = 0.004, indicating the number TFs acting in a pattern correlates with the time-to-evolve that pattern.

determinants: Binding site content of the CRM and features of the target expression pattern itself.

We first tested for a correlation between time-to-evolve for a target pattern and each TF's binding site count (or estimated occupancy) in the real CRM associated with that pattern. We found that binding site content of the TF HB has a strong positive correlation with time-to-evolve estimates (Pearson CC = 0.70, $P = 1.5 \times 10^{-5}$; fig. 3*A*). We also found that total binding site content of a CRM, aggregated over all six TFs, significantly positively correlates with time-to-evolve estimates (fig. 3*B*); however, this effect can be attributed mostly to HB site content, as indicated by a weak partial correlation coefficient (Johnson et al. 1992) with *P* of 0.45.

We next asked if certain aspects of the target pattern make it harder to evolve. A visual inspection (fig. 1*C*) suggested that expression in the anterior domain of the embryo marks smaller time-to-evolve estimates. To probe this point further, we calculated the Spearman's correlation coefficient between the expression level of a CRM at a fixed position along the *A*/*P* axis and the time-to-evolve estimate of that CRM, and repeated this procedure for every axial position. We found strong negative correlation at anterior positions (fig. 3*C*), that is, anterior expression patterns appear to be easier to evolve. We also noted that the plot of correlation coefficients in figure 3*C* very closely a resembles "flipped" version of the expression pattern of HB (fig. 3*C*, dashed line), suggesting

again that the faster evolution of CRMs with anterior patterns may be related to their HB binding levels. This is consistent with the fact that HB is modeled in GEMSTAT as a repressor, and therefore high levels of expression in the anterior end of the embryo indicate absence of HB sites in the CRM, which in turn correlates with shorter time-to-evolve estimates. We find it surprising that a single TF correlates so strongly with time-to-evolve estimates, and speculate that it may be due to the repeat-like T-rich motif of HB (supplementary fig. S3, Supplementary Material online), or an artifact of mechanistic details about HB regulation not captured in GEMSTAT (see Discussion).

Finally, we find that the number of TFs involved in generating the pattern also correlates with high time-to-evolve estimates for that pattern (fig. 3*D*), with Pearson's correlation coefficient of 0.49 ($P = 0.004$). We calculated the number of TFs needed to generate a pattern as the number of TFs that have at least one site above the LR threshold of 0.25 (He et al. 2012), but the correlation remains significant for other thresholds on the strength of sites (data not shown).

Dependence on Initial Conditions and the Possibility of Exaptation

Recall that each of our simulations begins with a random sequence. If the initial random sequences have a higher fitness value for certain target patterns, perhaps due to a greater frequency of random occurrence of certain binding sites necessary for that pattern, then such patterns may be quicker to evolve. This is the reason why we selected only 28 expression patterns out of the 37 A/P expression patterns modeled in Duque et al. (2014) (see Materials and Methods). Even within these 28 target patterns, we observed a significant positive correlation between the average fitness of random (initial) sequences and median time-to-evolve estimate (supplementary fig. S4, Supplementary Material online). However, a partial correlation analysis (Johnson et al. 1992) revealed that this correlation with fitness of initial sequences is not significant if we discount the already noted correlation with HB site counts in the real CRM. This was not true when partialing out the effect of other TFs' site counts (data not shown). Moreover, the correlation between HB site content and estimated time-to-evolve remains significant after partialing out the effect of initial fitness (data not shown). We interpret these observations to suggest that the number of HB sites in the initial random sequences influences the fitness of those sequences for certain target patterns, and therefore their time-to-evolve estimates.

Simulations beginning with random sequence represent an extreme scenario of evolution of regulatory sequences. In reality, features of the initial sequence where a CRM is to arise may strongly influence the waiting time. For instance, as previously noted by Durrett and Schmidt (2007, 2008) and MacArthur and Brookfield (2004), the composition of the genomic background affects the time required to evolve binding sites and regulatory sequences. Dermitzakis et al. (2003) noted that CRMs have short words that are close to becoming functional sites, and thus have the potential to quickly gain new function. Taking this line of reasoning further, one might argue that a CRM may readily evolve by transformation of a sequence that already contains several relevant binding sites (Prud'homme et al. 2007; Okada et al. 2010; Emera et al. 2012; de Souza et al. 2013), a scenario that may be considered as an example of exaptation, also known as co-opted evolution (Hoekstra 2006).

We designed two computational experiments to explore the effect of initial sequences on time-to-evolve. The first experiment simulates evolution under the favorable scenario where a CRM evolves from a sequence that drives an expression pattern very similar to the target pattern (see Materials and Methods). The second experiment explores an opposite scenario, in which a CRM evolves from a sequence that drives a very different pattern (e.g., in which a CRM with anterior expression evolves from a sequence that drives posterior expression). As expected, the time to evolve each of the CRMs in the first experiment is largely reduced (supplementary fig. S5*A*, Supplementary Material online) due to the abundance of binding sites for the necessary TFs. However, simulations from initial sequences that drive a pattern anticorrelated with the target has a negative effect on evolutionary time of several CRMs (supplementary fig. S5*B*, Supplementary Material online). This is due to the contrasting roles that some pairs of CRMs have. Binding sites present in the initial sequence are expected to reduce time-to-evolve only if they are for the right TFs, that is, ones that can contribute to the target pattern. If, on the other hand, the starting sequence has several sites that disrupt the target pattern and few sites that contribute to it, evolution will have to proceed by deconstructing the initial sequence before it can start constructing the target pattern. For example, the kni_83_ru CRM drives expression in a stripe in the posterior end of the embryo (supplementary fig. S1, Supplementary Material online), and contains many binding sites for CAD, GT, HB, and KR. If we used this sequence to initiate simulations for the target pattern "eve_1_ru," a stripe in the anterior end of the embryo, the evolving sequence would have to lose most of its binding sites for CAD and HB, maintaining the sites for KR and gaining new sites for BCD.

Uniformly Expressed Activators Can Speed Up Emergence of CRMs

Patterning of the early *Drosophila* embryo is well known to be achieved by gradients of maternally deposited TFs and by their patterned regulatory targets. Recent studies have focused also on uniformly expressed TFs that function as important activators in patterning systems (Liang et al. 2008; Harrison et al. 2011; Tsurumi et al. 2011). These activators by themselves do

not or may not have the patterning ability of nonuniformly expressed TFs, but can modulate the response of a CRM to a patterned signal (Kanodia et al. 2012). They are present in several regulatory systems including the A/P system (Arbouzova and Zeidler 2006; Liang et al. 2008; Kanodia et al. 2012), the Dorsal–Ventral patterning system (Liang et al. 2008; Kanodia et al. 2012), and other patterning or developmental systems (Arbouzova and Zeidler 2006; Nien et al. 2011; Tsurumi et al. 2011). Here, we pursued the hypothesis that the deployment of uniformly expressed activators in patterning systems also has an evolutionary explanation: That they improve the "evolvability" of target patterns, by increasing the number of viable paths evolution can take from a random initial sequence to a functional CRM.

To explore this hypothesis, we repeated the time-to-evolve simulations from above with a GEMSTAT (CRM function) model specification that includes a ubiquitous activator, and compared the results with those from the original model. We designed a methodology that ensures that there exists a fit solution (CRM) for the target pattern under either function model, with and without the ubiquitous activator, so that any difference in time-to-evolve can be attributed to the evolutionary ramifications of the ubiquitous activator (see Materials and Methods). We tested the effects of two well-characterized ubiquitous activators, ZLD (Liang et al. 2008; Harrison et al. 2011) and DSTAT (Tsurumi et al. 2011), separately. As shown in figure 4, each of these TFs reduces the median time-to-evolve for several target patterns, with the effect of ZLD being clearly more prominent. A two-way analysis of variance supported these observations, with P value of 2×10^{-4} (ZLD) and 0.03 (DSTAT) (supplementary tables S1 and S2, Supplementary Material online), indicating that adding either ubiquitous activator to the model has a statistically significant effect of decreasing time-to-evolve.

For deeper insights into the effect of ubiquitous activators on time-to-evolve, we discuss the example of the target pattern "eve_37ext_ru," which comprises a single stripe of expression peaking at about 49% egg length (fig. 4*D*). (This is the third stripe of eve expression along the *A/P* axis, with stripe 7 being outside the modeled range of 20–80% egg length.) To drive this pattern using the TFs in the baseline model (BCD, CAD, GT, HB, KNI, and KR), whose A/P expression profiles are shown in figure 4*C*, evolution could add activator sites for TFs BCD and CAD, generating expression across all the *A/P* axis, and add repressor sites for KNI and HB to create repression at the anterior and posterior sides of the desired stripe. Indeed, this may be the strategy employed by nature (Goltsev et al. 2004), because these are the TFs for which sites are present in the real CRM from *D. melanogaster* (fig. 4*E*). However, neither BCD nor CAD has maximal concentration around 49% egg length (fig. 4*C*), and to create sufficient activation in the central domain of the *A/P* axis, it would be necessary to add several strong sites to the CRM. On the other hand, if DSTAT is also available (as a ubiquitous activator), evolution could use DSTAT sites to add to the weaker activation by BCD and CAD in the central domain. This offers another avenue for evolution to explore, ultimately leading to a lower time-to-evolve in our simulations. Intriguingly, the *D. melanogaster* CRM for eve_37ext_ru has two DSTAT sites (not shown), suggesting that this may indeed have been the avenue taken by evolution. Our interpretation is in agreement with theories of evolutionary computation (Holland 1975; Goldberg 1989, 2002), according to which if a combinatorial problem has many fit solutions we are more likely to find one of these solutions quickly (Goldberg 2002).

Sensitivity to Evolutionary Parameters

We began this study by estimating the time necessary to evolve 28 different expression patterns starting from a random sequence. These estimates are expected to depend on values of the population genetics parameters used in the simulations, in particular the population size N, the mutation rate μ, and the selection coefficient s. We explored these dependencies next, varying the simulation parameters within reasonable ranges.

All our simulations used a time-rescaling heuristic (Hoggart et al. 2007; He et al. 2012) for speeding up simulations, with a scaling factor $\lambda = 1{,}000$, a time-scaled population size $2N = 1{,}000$, and a time-scaled mutation rate $\mu = 10^{-5}$ (mutations per generation per base pair), resulting in a scaled mutation rate $2N\mu = 10^{-2}$, which is within the estimated range of $10^{-2} - 10^{-4}$ (Drake et al. 1998; Thornton and Andolfatto 2006) for *Drosophila* (see Materials and Methods). We note however that this mutation rate is higher than that used in Duque et al. (2014). The higher mutation rate reduces the computational time required for a simulation and as mentioned is still within the estimated range for *Drosophila*. However, to understand the effect of mutation rate on our results, we repeated the time-to-evolve estimation procedure with values of $2N\mu$ that are an order of magnitude greater or lesser than 10^{-2}. Figure 5*A* shows how the time to evolve a CRM, averaged over the 28 target patterns, changes with the values of $2N\mu$. Changing the scaled mutation rate $2N\mu$ by a factor of 10 results in time-to-evolve estimates that change by less than 10 times, which is not unexpected because different values of $2N\mu$ result in different balances between selection and drift. In particular, reducing $2N\mu$ from 0.01 to 0.001 (a factor of 10) results in average time-to-evolve increasing about 7-fold from approximately 2.1 to approximately 18 Myr, with estimates for individual target patterns ranging between 2.1 and 25 Myr. (As a comparison point, we note that the estimated divergence time between *D. melanogaster* and *Drosophila pseudoobscura* to be 25–55 Myr; Richards et al. 2005.)

Another important population genetics parameter is the selection coefficient s, or equivalently, the population-scaled selection coefficient $4Ns$. In our simulations, the strength of

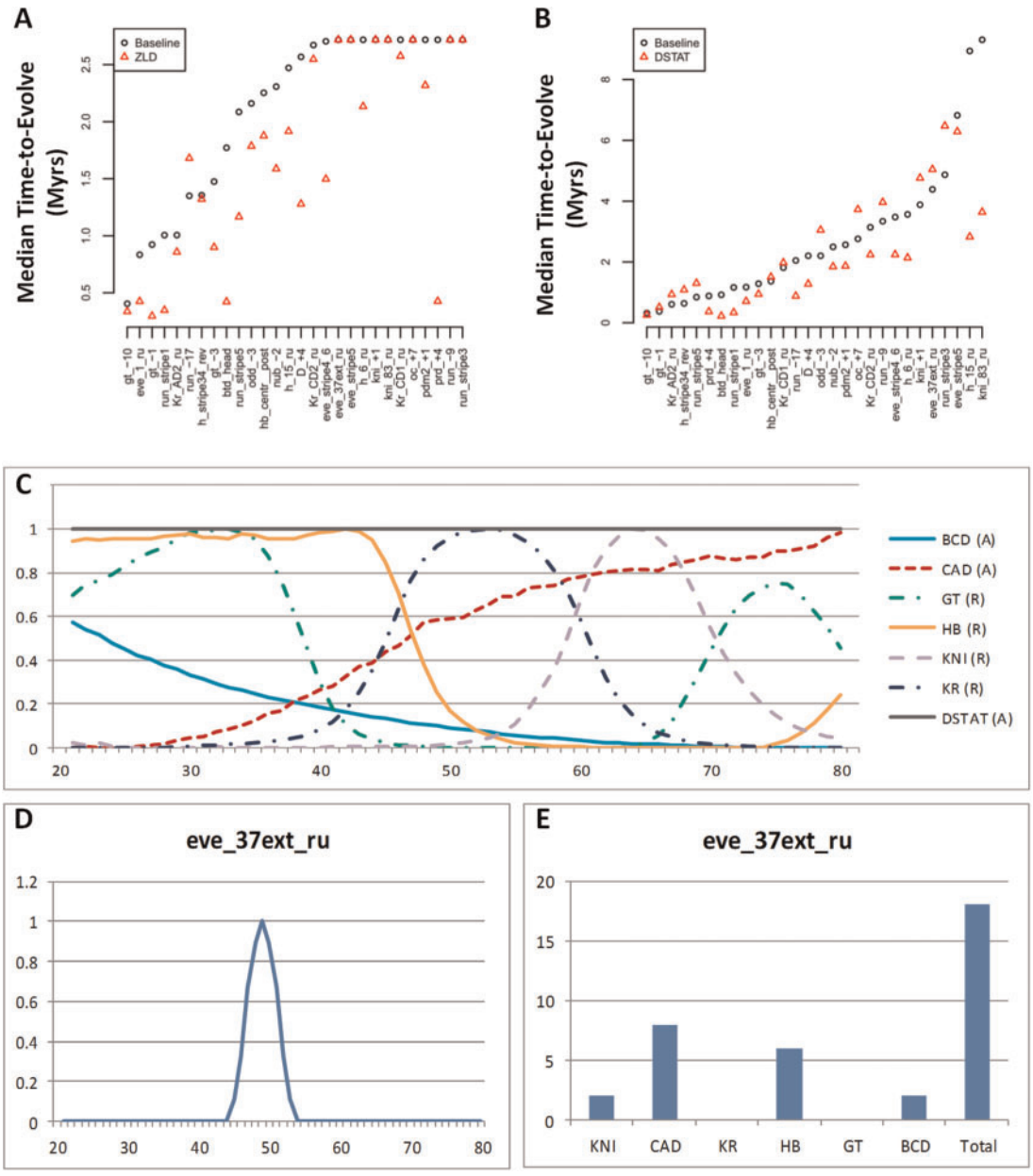

Fig. 4.—The effect of uniformly expressed activators on time-to-evolve for each target pattern. (*A*) Comparison of time-to-evolve estimates between the baseline model and a model that includes ZLD as a uniform activator. Expression patterns are sorted based on time-to-evolve estimates from the baseline model. (*B*) Comparison of time-to-evolve estimates between the baseline model and a model that includes DSTAT as a uniform activator. (*C*) Concentration profiles of seven TFs across the *A/P* axis. Activators are indicated with an *A* and repressors with an *R*. (*D*) Target expression pattern for the CRM eve_37ext_ru. (*E*) Number of sites present in the eve_37ext_ru CRM in *D. melanogaster*, for each of the six TFs (other than DSTAT). Sites are called at relative strength of 0.25 following the procedure described in Materials and Methods.

selection is controlled by the selection scale parameter K, which is analogous to s when two competing individuals have fitness of 0 and 1. For the experiments reported above, we used $K = 50$, which is of the same order as the value determined in Duque et al. (2014) to provide the best fit to real evolutionary data. In the absence of better tools to estimate the actual strength of selection, this value is our best guess in the context of our experiment. Nevertheless, we repeated our experiments with different values of K (2, 5, 25, 50, 100), as shown in figure 5*B*, in part to compensate for our lack of knowledge of the real selection strength and in part to understand how the selection strength influences the time-to-evolve. As expected, smaller values of the selection scale K result in longer times necessary to evolve CRMs. For instance,

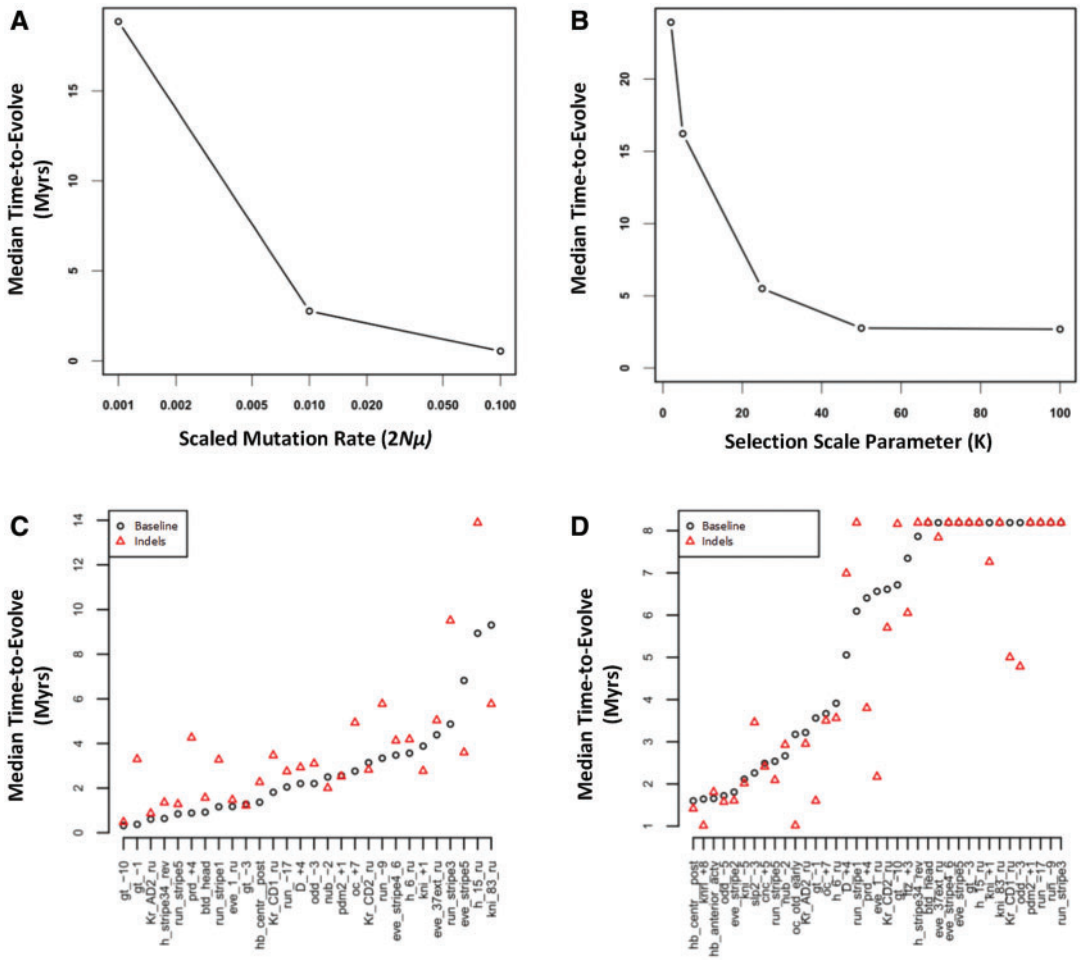

Fig. 5.—Sensitivity of time-to-evolve estimates to simulation parameters. (*A*) Sensitivity to the scaled mutation rate ($2N\mu$). Shown are the average time-to-evolve (median of all simulations for a pattern, averaged over 28 target expression patterns) for three values of $2N\mu$. (*B*) Sensitivity to selection scale parameter (K). (*C*) The effect of indels. The plot shows time-to-evolve estimates (y axis) for each of the 28 target expression patterns (x axis) for an evolutionary model without insertions or deletions (black circles) and an evolutionary model that includes indels (red triangles). Adding indels significantly increases time-to-evolve estimates ($P = 0.0002$). (*D*) Same as (*C*), except that the initial sequences for simulations are real CRMs that drive a pattern anticorrelated with the target pattern. The trend is that including indels in these simulations reduces the time-to-evolve estimates.

reducing the selection scale by a factor of 10 ($K = 5$) results in time-to-evolve estimates increasing by less than a factor of 10. We also found an apparent saturation in the effect of increasing the selection strength from 50 to 100 (fig. 5*B*).

Finally, we note that other assumptions about the evolutionary model might also influence time-to-evolve estimates. For example, insertions and deletions (indels) have been suggested to have important effects on the evolution of regulatory sequences (Sinha and Siggia 2005; Lusk and Eisen 2010; Nourmohammad and Lässig 2011); recombination has also been suggested to influence the rate of adaptation (Schoustra et al. 2007) and even ploidy had been suggested to influence adaptation (Orr and Otto 1994; Zeyl et al. 2003). It is beyond the scope of this work to test for the effect of all such mechanisms, but we examined the effects of indels on time-to-evolve estimates (fig. 5*C*). We find that adding indels (insertions implemented as short tandem repeats, as in He et al. 2012) to our model increases time-to-fit estimates from approximately 2.8 to approximately 3.6 Myr on average (fig. 5*C*), a statistically significant increase (paired t-test with pooled standard deviation, $P = 0.0002$). One way to interpret this is that insertions and deletions are more likely to completely destroy binding sites than point mutations, and therefore are more likely to be selected against. Interestingly, the effect of indels reverses direction when performing the same comparison using simulations where the initial sequence was a real CRM sequence that drives a highly dissimilar pattern. (These simulations were done in a manner similar to that reported above and in supplementary fig. S5*B*, Supplementary Material online, except for the choice of initial sequence.) In particular, the salient trend was for time-to-fit estimates to decrease (fig. 5*D*). It appears that indels might have a role in making

it easier to "exapt" a CRM that starts from a very different expression pattern. Similar tests with initial sequences driving patterns similar to the target pattern are reported in supplementary figure S5C, Supplementary Material online, but show a less clear trend.

Discussion

We used the evolutionary simulation framework of PEBCRES, from our previous work (He et al. 2012; Duque et al. 2014), to study what it might take for a functional CRM to evolve, first asking how long this may take under strong selection, and then investigating various factors that may influence the estimated time-to-evolve. These questions have been addressed in various ways by other authors before us, for instance, by Stone and Wray (2001), MacArthur and Brookfield (2004), and Durrett and Schmidt (2007, 2008). Early approaches to this question focused on the independent evolution of single binding sites (Stone and Wray 2001; Durrett and Schmidt 2007), pairs of binding sites (Durrett and Schmidt 2008), or simple CRMs composed of a single TFs (MacArthur and Brookfield 2004). However, questions regarding CRM evolution assume additional complexity due to the diverse mechanisms and combinatorial nature of gene regulation, which have not been adequately addressed in previous work. In recent work, we developed the PEBCRES evolutionary framework to bridge this gap, and used it to accurately model the evolutionary dynamics of binding sites within CRMs under strong negative selection for a fixed regulatory function (Duque et al. 2014). The success of that work encouraged us to explore here a complementary aspect of CRM evolution—that of emergence of a new CRM under strong adaptive forces.

Before discussing our findings further, we note that our simulation-based inferences are theoretical projections, given the many unknown aspects of regulatory and evolutionary biology that our formalism does not capture. For instance, our fitness function is based on comparing predicted and target expression patterns and the mapping of the resulting deviation to a fitness score is an ad hoc choice. (It is a simple nonlinear function so that fitness rises much more rapidly when the evolving pattern is close to the target.) There is little guidance from the literature regarding a realistic mapping, and this will be a major research direction in itself. The wPGP score used here for comparing two expression patterns is itself subject to questioning, although we have argued in previous work (Samee and Sinha 2013) why this is preferable to more obvious alternatives. Caveats arise not only from the evolutionary aspects but also the *cis*-regulatory aspects of the model, which are far from being fully characterized, even though we intentionally focused on one of the best characterized regulatory systems here. The impact of chromatin structure and nucleosome positioning is not included in our function model, which operates under the assumption that the entire enhancer is an "open chromatin" region; this assumption is true only to a first order of approximation.

We estimated that CRMs that exhibit the combinatorial complexity associated with early developmental enhancers (specifically, those involved in A/P patterning in *Drosophila* embryos) can emerge on fairly short time scales, of the order of few millions of years, even when starting from random sequences of little or no functional ability. A recent study by Arnold et al. (2014) used massively parallel enhancer screens (Arnold et al. 2013) to find that hundreds of novel CRMs have emerged on the scale of approximately 10 Myr, lending credibility to our theoretical findings. Although we are not aware of other previous studies reporting time-to-evolve estimates for CRMs, it is worth noting that Durrett and Schmidt (2007) estimated that an 8-bp long "fuzzy" binding site (one mismatch allowed) might emerge in the human population on a time scale of 60,000 years. A CRM evolved in our simulations for the gt_-10 pattern, for example, has about 7 binding sites on average, and takes about 0.3 Myr to evolve. This agrees roughly with an extrapolation from Durrett and Schmidt (2007) whereby the time for 7 binding sites to emerge should about 0.42 Myr, assuming sites are not lost and sites emerge sequentially. This is a ballpark comparison, because the two estimates are for human and fruitfly populations respectively and contingent on different assumptions about a binding site's information content.

Here, CRMs were evolved in silico to drive predetermined expression patterns along the *A/P* axis. CRMs arising from different simulations for the same target expression pattern tended to cluster strongly in terms of site composition, with distinct expression patterns defining distinct clusters of "fit" enhancers. Importantly, we noted evolved sequences to be similar in site composition to the real *D. melanogaster* CRMs associated with their respective patterns, thus demonstrating agreement between model-based evolutionary simulations and real data. The few exceptions from this general trend were also illuminating, with the evolved CRMs being significantly more parsimonious than their real counterparts, leading to speculations about a more complex evolutionary history of those real CRMs or about missing regulatory mechanisms in the PEBCRES/GEMSTAT framework (Duque et al. 2014). This latter point deserves special mention as missing regulatory mechanisms can shade the findings of simulation-based studies, as was demonstrated in our recent work (Duque et al. 2014). For instance, we note that the A/P patterns used as targets in our simulations lack terminal aspects—we only considered the regulatory function of a CRM in the range 20–80% egg length. This may lead to underestimates of time-to-evolve CRM for certain patterns. Proper modeling of these CRMs requires that the underlying fitness function, specifically GEMSTAT, uses additional TFs, some of which are not known (He et al. 2010). Additionally, previous work on GEMSTAT (He et al. 2010) and PEBCRES (Duque et al. 2014) have produced careful estimates for many of the free parameters used

in this work by modeling expression patterns that excluded the terminal ends. These were two major reasons why we decided to exclude the terminal ends of the embryo from our analysis.

We noted up to a 30-fold variation in the time-to-evolve CRMs for different expression patterns, naturally raising the following question: What causes this variable time-to-evolve? The flexibility inherent in the PEBCRES framework (as opposed to a purely analytical framework; Durrett and Schmidt 2007, 2008) allowed us to explore different aspects of the evolutionary process and regulatory mechanisms, and how they might affect emergence time of CRMs. For example, we asked how these times might be affected if a CRM, instead of evolving from genomic background, arose from a sequence that already drives some expression pattern. Unsurprisingly, we found that if the two expression patterns (that driven by the original sequence and the target pattern) are highly similar, the time to evolve a CRM is greatly reduced; however, perhaps more interestingly, the emergence time can also be significantly greater if the expression patterns are very dissimilar. Dependence of evolution times on initial sequences has been proposed in previous work. For instance, MacArthur and Brookfield (2004) argued that the time to evolve a CRM that drives a certain level of activation by a TF may be influenced by the CG-content of the initial sequence.

As another example of factors affecting time-to-evolve, we found that ubiquitous activators, which are not by themselves capable of patterning a target gene, may work with other TFs and reduce the time to evolve a CRM. We speculate that this may be due to two complementary reasons: 1) Ubiquitous activators provide alternative solutions to the underlying combinatorial optimization problem of finding a fit CRM and 2) ubiquitous activators reduce the number of binding sites necessary to create certain expression patterns, and thus the number of steps (mutations) needed to find a fit solution. Both situations are expected to reduce the time to find one such solution, as per theories of evolutionary computation (Goldberg 2002).

Our simulations suggest that CRMs with more combinatorial regulation (measured by the number of TFs with sites in the *D. melanogaster* CRM for the same pattern) should take longer to evolve. Perhaps more surprisingly, we noted that the binding site content for a particular TF—HB—is one of the strongest predictors of time-to-evolve values. It is possible that this points to shortcomings of our simulation framework. We noted that the HB motif can be characterized as a poly-T repeat (supplementary fig. S3, Supplementary Material online). Such repeat patterns might be easily created through mutational mechanisms that we have not modeled adequately in PEBCRES (Nourmohammad and Lässig 2011). Moreover, there is evidence that HB might play dual roles of activator and repressor depending on the regulatory context (Papatsenko and Levine 2008; Bieler et al. 2011). The absence of this mechanism in our GEMSTAT-based fitness function may be related to the strong correlation noted above, and illustrate more generally how evolutionary modeling may lead us to closer examination of mechanisms encoded in *cis*-regulatory sequences (Duque et al. 2014).

We also found that anterior expression patterns were quicker to evolve sequences than posterior patterns. However, this observation is likely a consequence of the already mentioned influence of HB site counts. Noting that HB is modeled as a repressor and is largely expressed in the anterior end of the embryo, anterior expression correlates with lesser site content for HB, which in turn correlates with shorter time-to-evolve values.

Our application of PEBCRES to understanding the evolution of CRMs can be extended in several ways. For example, our model could be used to shed light on shadow enhancers (Perry et al. 2010; Barolo 2012), by using the GEMSTAT-GL model of locus-level modeling for regulatory function prediction instead of the GEMSTAT model of enhancer function. Other avenues of future exploration include understanding the effect of indirect activators (Kanodia et al. 2012), the effect of local duplications (Sinha and Siggia 2005) on time-to-evolve estimates, exploring the robustness of evolved CRMs to fluctuations in input TF concentrations (Pujato et al. 2013) and how such robustness might evolve (Wagner 2005), and understanding how evolvability (Wagner and Altenberg 1996; Wagner 2005) affects the architecture of *cis*-regulatory sequences and how it evolves in the first place (Wagner and Altenberg 1996).

Supplementary Material

Supplementary figures S1–S6, tables S1–S2, and notes S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was performed with support from the National Science Foundation (DBI-0746303, EFRI-1136913) to S.S. The authors thank Xin He for valuable discussions in the early stages of this work. They also thank Md. Abul Hassan Samee for his assistance in configuring GEMSTAT for our simulations.

Literature Cited

Aparicio S, et al. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. Proc Natl Acad Sci U S A. 92:1684–1688.

Arbouzova NI, Zeidler MP. 2006. JAK/STAT signalling in *Drosophila*: insights into conserved regulatory and cellular functions. Development 133:2605–2616.

Arnold CD, et al. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339:1074–1077.

Arnold CD, et al. 2014. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution. Nat Genet. 46:685–692.

Barolo S. 2012. Shadow enhancers: frequently asked questions about distributed *cis*-regulatory information and enhancer redundancy. Bioessays 34:135–141.
Bedford T, Hartl DL. 2008. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in *Drosophila*. Mol Biol Evol. 25:1631–1638.
Berg J, Willmann S, Lässig M. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol Biol. 4:42.
Bieler J, Pozzorini C, Naef F. 2011. Whole-embryo modeling of early segmentation in *Drosophila* identifies robust and fragile expression domains. Biophys J. 101:287–296.
Carter AJ, Wagner GP. 2002. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc Biol Sci. 269:953–960.
Damjanovski S, Huynh MH, Motamed K, Sage EH, Ringuette M. 1998. Regulation of SPARC expression during early *Xenopus* development: evolutionary divergence and conservation of DNA regulatory elements between amphibians and mammals. Dev Genes Evol. 207:453–461.
Davidson EH. 2010. The regulatory genome: gene regulatory networks in development and evolution. San Diego (CA): Academic Press.
de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements into novel *cis*-regulatory elements: is the evidence always strong? Mol Biol Evol. 30:1239–1251.
Dermitzakis ET, Bergman CM, Clark AG. 2003. Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol. 20:703–714.
Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148:1667–1686.
Duque T, et al. 2014. Simulations of enhancer evolution provide mechanistic insights into gene regulation. Mol Biol Evol. 31:184–200.
Durrett R, Schmidt D. 2007. Waiting for regulatory sequences to appear. Annu Appl Probab. 17:1–32.
Durrett R, Schmidt D. 2008. Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. Genetics 180:1501–1509.
Emera D, et al. 2012. Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. Mol Biol Evol. 29:239–247.
Fisher RA. 1999. The genetical theory of natural selection: a complete variorum edition. Oxford: Oxford University Press.
Francois P, Hakim V, Siggia ED. 2007. Deriving structure from evolution: metazoan segmentation. Mol Syst Biol. 3:154.
Gerland U, Hwa T. 2002. On the selection and evolution of regulatory DNA motifs. J Mol Evol. 55:386–400.
Goldberg DE. 1989. Genetic algorithms in search, optimization and machine learning. Reading (MA): Addison-Wesley.
Goldberg DE. 2002. The design of innovation: lessons from and for competent genetic algorithms. Boston (MA): Kluwer Academic.
Goltsev Y, Hsiong W, Lanzaro G, Levine M. 2004. Different combinations of gap repressors for common stripes in *Anopheles* and *Drosophila* embryos. Dev Biol. 275:435–446.
Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB. 2011. Zelda binding in the early *Drosophila melanogaster* embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet. 7:e1002266.
He BZ, Holloway AK, Maerkl SJ, Kreitman M. 2011. Does positive selection drive transcription factor binding site turnover? A test with *Drosophila* *cis*-regulatory modules. PLoS Genet. 7:e1002053.
He X, Duque TS, Sinha S. 2012. Evolutionary origins of transcription factor binding site clusters. Mol Biol Evol. 29:1059–1070.
He X, Samee MA, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol. 6. pii: e1000935.
Hoekstra HE. 2006. Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity (Edinb) 97:222–234.
Hoggart CJ, et al. 2007. Sequence-level population simulations over large genomic regions. Genetics 177:1725–1731.
Holland JH. 1975. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. Ann Arbor (MI): University of Michigan Press.
Johnson RA, Wichern DW, Education P. 1992. Applied multivariate statistical analysis. Englewood Cliffs (NJ): Prentice Hall.
Josephides C, Moses AM. 2011. Modeling the evolution of a classic genetic switch. BMC Syst Biol. 5:24.
Kanodia JS, et al. 2012. Pattern formation by graded and uniform signals in the early *Drosophila* embryo. Biophys J. 102:427–433.
Kim AR, et al. 2013. Rearrangements of 2.5 kilobases of noncoding DNA from the *Drosophila* even-skipped locus define predictive rules of genomic *cis*-regulatory logic. PLoS Genet. 9:e1003243.
Kim J, He X, Sinha S. 2009. Evolution of regulatory sequences in 12 *Drosophila* species. PLoS Genet. 5:e1000330.
Liang HL, et al. 2008. The zinc-finger protein Zelda is a key activator of the early zygotic genome in *Drosophila*. Nature 456:400–403.
Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. Development 125:949–958.
Lusk RW, Eisen MB. 2010. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. PLoS Genet. 6:e1000829.
MacArthur S, Brookfield JF. 2004. Expected rates and modes of evolution of enhancer sequences. Mol Biol Evol. 21:1064–1073.
Maeso I, Irimia M, Tena JJ, Casares F, Gomez-Skarmeta JL. 2013. Deep conservation of *cis*-regulatory elements in metazoans. Philos Trans R Soc Lond B Biol Sci. 368:20130020.
Moses AM. 2009. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol Biol. 9:286.
Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 5:R98.
Moses AM, et al. 2006. Large-scale turnover of functional transcription factor binding sites in *Drosophila*. PLoS Comput Biol. 2:e130.
Nien CY, et al. 2011. Temporal coordination of gene networks by Zelda in the early *Drosophila* embryo. PLoS Genet. 7:e1002339.
Nourmohammad A, Lässig M. 2011. Formation of regulatory modules by local sequence duplication. PLoS Comput Biol. 7:e1002167.
Okada N, Sasaki T, Shimogori T, Nishihara H. 2010. Emergence of mammals by emergency: exaptation. Genes Cells 15:801–812.
Orr HA, Otto SP. 1994. Does diploidy increase the rate of adaptation? Genetics 136:1475–1480
Papatsenko D, Levine MS. 2008. Dual regulation by the Hunchback gradient in the *Drosophila* embryo. Proc Natl Acad Sci U S A. 105:2901–2906.
Perry MW, Boettiger AN, Bothma JP, Levine M. 2010. Shadow enhancers foster robustness of *Drosophila* gastrulation. Curr Biol. 20:1562–1567.
Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A. 104(Suppl. 1):8605–8612.
Pujato M, MacCarthy T, Fiser A, Bergman A. 2013. The underlying molecular and network level mechanisms in the evolution of robustness in gene regulatory networks. PLoS Comput Biol. 9:e1002865.
Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 300:1742–1745.
Richards S, et al. 2005. Comparative genome sequencing of *Drosophila* *pseudoobscura*: chromosomal, gene, and *cis*-element evolution. Genome Res. 15:1–18.

Ross JL, Fong PP, Cavener DR. 1994. Correlated evolution of the *cis*-acting regulatory elements and developmental expression of the *Drosophila Gld* gene in seven species from the subgroup *melanogaster*. Dev Genet. 15:38–50.
Samee AH, Sinha S. 2013. Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data. Methods 62:79–90.
Samee MA, Sinha S. 2014. Quantitative modeling of a gene's expression from its intergenic sequence. PLoS Comput Biol. 10:e1003467.
Satta Y, Ishiwa H, Chigusa SI. 1987. Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. Mol Biol Evol. 4:638–650.
Satta Y, Takahata N. 1990. Evolution of *Drosophila* mitochondrial DNA and the history of the *melanogaster* subgroup. Proc Natl Acad Sci U S A. 87:9558–9562.
Schoustra SE, Debets AJ, Slakhorst M, Hoekstra RF. 2007. Mitotic recombination accelerates adaptation in the fungus *Aspergillus nidulans*. PLoS Genet. 3:e68.
Segal JA, Barnett JL, Crawford DL. 1999. Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. J Mol Evol. 49:736–749.
Sinha S, Siggia ED. 2005. Sequence turnover and tandem repeats in *cis*-regulatory modules in *Drosophila*. Mol Biol Evol. 22:874–885.
Stone JR, Wray GA. 2001. Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol Biol Evol. 18:1764–1770.
Swanson CI, Schwimmer DB, Barolo S. 2011. Rapid evolutionary rewiring of a structurally constrained eye enhancer. Curr Biol. 21:1186–1196.
Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619.
Tsurumi A, et al. 2011. STAT is an essential activator of the zygotic genome in the early *Drosophila* embryo. PLoS Genet. 7:e1002086.
Wagner A. 2005. Robustness and evolvability in living systems. Princeton (NJ): Princeton University Press.
Wagner GP, Altenberg L. 1996. Perspective: complex adaptations and the evolution of evolvability. Evolution 50:967–976.
Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.
Zeyl C, Vanderford T, Carter M. 2003. An evolutionary advantage of haploidy in large yeast populations. Science 299:555–558.
Zinzen RP, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the *Drosophila* embryo. Curr Biol. 16:1358–1365.

Associate editor: Ross Hardison