BMC Bioinformatics

BioMed Central

Research article

Open Access

Conservation of regulatory elements between two species of Drosophila

Eldon Emberly†[1], Nikolaus Rajewsky†[2] and Eric D Siggia*[1]

Address: [1]Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY, USA and [2]Department of Biology, New York University, 1009 Main Building, 100 Washington Square East, New York, NY, USA

Email: Eldon Emberly - eemberly@edsc.rockefeller.edu; Nikolaus Rajewsky - nikolaus.rajewsky@nyu.edu;
Eric D Siggia* - siggia@eds1.rockefeller.edu

* Corresponding author †Equal contributors

Published: 20 November 2003

BMC Bioinformatics 2003, **4**:57

Received: 06 May 2003
Accepted: 20 November 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/57

Abstract

Background: One of the important goals in the post-genomic era is to determine the regulatory elements within the non-coding DNA of a given organism's genome. The identification of functional *cis*-regulatory modules has proven difficult since the component factor binding sites are small and the rules governing their arrangement are poorly understood. However, the genomes of suitably diverged species help to predict regulatory elements based on the generally accepted assumption that conserved blocks of genomic sequence are likely to be functional. To judge the efficacy of strategies that prefilter by sequence conservation it is important to know to what extent the converse assumption holds, namely that functional elements common to both species will fall within these conserved blocks. The recently completed sequence of a second Drosophila species provides an opportunity to test this assumption for one of the experimentally best studied regulatory networks in multicellular organisms, the body patterning of the fly embryo.

Results: We find that 50%–70% of known binding sites reside in conserved sequence blocks, but these percentages are not greatly enriched over what is expected by chance. Finally, a computational genome-wide search in both species for regulatory modules based on clusters of binding sites suggests that genes central to the regulatory network are consistently recovered.

Conclusions: Our results indicate that binding sites remain clustered for these "core modules" while not necessarily residing in conserved blocks. This is an important clue as to how regulatory information is encoded in the genome and how modules evolve.

Background

Changes in the body plans of metazoans are thought to be largely a consequence of changes in the spatiotemporal pattern of the expression of developmental genes and thus a consequence of changes in their transcriptional regulation ([1,2] and references therein). Hence, genomic *cis*-regulatory sequences that control developmental gene expression can be thought of as the genomic "source code" for development [3]. Multiple transcription factors recognize and bind regulatory sites in these genomic regulatory sequences ("modules") and together define the rate of transcription of the target gene. This combinatorial mode of gene regulation is thought to be prevalent in all multicellular organisms. Modules are often separable and define space and time specific aspects of gene expression. They integrate inputs from several genes and regulate another gene to form regulatory networks. Modules are typically a few hundred nucleotides long and receive multiple inputs from roughly 4–5 different transcription factors. Recently, it has been demonstrated in *D. melanogaster* that modules may be identified by searching the genome for regions which are dense in binding sites [4-8]. These computational methods are important practically, since the experimental detection of modules is very time consuming. Computational techniques applied to whole genomes may also prove informative about how modules evolve, a question that has been addressed many times in the fly literature [17-22] for specific genes. The development of compuational methods for predicting regulatory modules, and a better understanding of module evolution should proceed in tandem, since both require elucidating what aspects of the sequence matter for function.

Although it is well accepted that blocks of sequence conserved between species are functional in some respect, it is not clear to what extent functional binding sites in either species will reside in conserved blocks; they are logically different assertions. We only consider blocks that are prominent enough to be recognized as statistically signficant by an alignment program without other prior information, and thus can potentially be used to prefilter the noncoding DNA for putative regulatory modules and factor binding sites, a strategy which is used in many studies [9,14,15,21,22,34]. It is not a priori evident how much of the regulatory information in the genome is missed if one only examines the conserved blocks. (The secondary structure of tRNA's is a reminder that the primary sequence can vary greatly yet not affect function.) Fortunately, the sequencing of two different *Drosophila* species has provided a large data set to address these fundamental questions [26]. *D. melanogaster* and *D. pseudoobscura* are roughly 25 Myr apart [24] and have roughly the same degree of conservation in their noncoding DNA as human-mouse which have proven to be amenable to meaningful cross-species comparisons of regulatory elements. Additionally, we will use less extensive sequence data for *D. virilis* [25], roughly 40 Myr [24] removed from *D. melanogaster*.

We will work within one of the experimentally best studied developmental paradigms, the body patterning of the *Drosophila* embryo [23]. Development proceeds by a transcriptional cascade that refines the patterns laid down maternally. Many of the regulatory events have been tied to specific modules. We have collected 30 modules with 315 binding sites for the segmentation gene hierachy, for the anterior posterior patterning, the dorsal-ventral system, and several genes that initiate mesoderm patterning. The modules have all been tested by their ability to drive a reporter gene *in vivo* that recapitulates a portion of the native expression pattern; and all the binding sites have experimental support mostly *in vitro*.

We have correlated this data set with the sequence conserved between two species of *Drosophila* using different multiple alignment programs and exploring extensively their parameter space. Our conclusion that the correlation betweeen binding sites and conserved sequence is not much greater than chance (but still statistically significant since the data set is large) seems robust against all plausible variation in parameters and technique. We show how data from multiple species can be used to enhance the prediction of gene regulatory interactions, if one goes beyond sequence alignment, and rather predicts regulated genes by proximity to clusters of computationally derived binding sites. Multiple species then serve to reduce the rate of false positive predictions. We conclude by considering the implications of our findings for module evolution.

Results

Conservation of known binding sites

Reference [4] collected from the Drosophila literature 21 regulatory modules, with 230 annotated binding sites for 11 factors relevant to early body plan. To these we added modules regulating the fushi tarazu (ftz) gene [27], the engrailed intron enhancer[28], single-minded enhancer[29], even-skipped mesoderm enhancer[7], and heart-broken mesoderm enhancer[7]. Included were binding sites for the well studied maternal factors, (bicoid (bcd), caudal (cad)), zygotic gap genes (hunchback (hb), Kruppel (Kr), knirps (kni), and tailless (tll), pair-rule genes (even-skipped (eve), paired (prd), tramtrak (ttk) and fushi tarazu (ftz)), dorsal-ventral patterned gene factors (dorsal (dl), deadringer (dri), brinker (bri), snail (sna) and twist (twi)) and known factors in mesoderm patterning (mothers against decapentaplegic (Mad), Tinman (Tin), Pointed (Pnt)). It total, 3500 base pairs of sequence are covered by one or more binding site, all based on experiment.

There was no difficulty in finding orthologs for all the genes in our data set in the *D. pseudoobscura* contigs [26]. We found the best syntenous (order preserving) alignment of the noncoding sequence around these genes using two programs LAGAN [30] and SMASH [31], which make different scoring assumptions and produce different spectrums of conserved block sizes (the median is 20 bp for LAGAN and 40–130 bp for SMASH see Fig. 3, Methods). There was then no ambiguity in finding clear homologues for each module in *D. pseudoobscura* since there were prominent syntenous blocks every few hundred bases on average, except for relatively rare gaps. We pulled out the entire up/down stream noncoding sequence of any gene with a module 5' or 3' to it and aligned a total of 200 kb of sequence in each species inorder to provide a context for the annotated regulatory modules, which themselves only totaled 21426 bp. Only a small portion of the *D. virilis* genome is sequenced, but it was targeted at

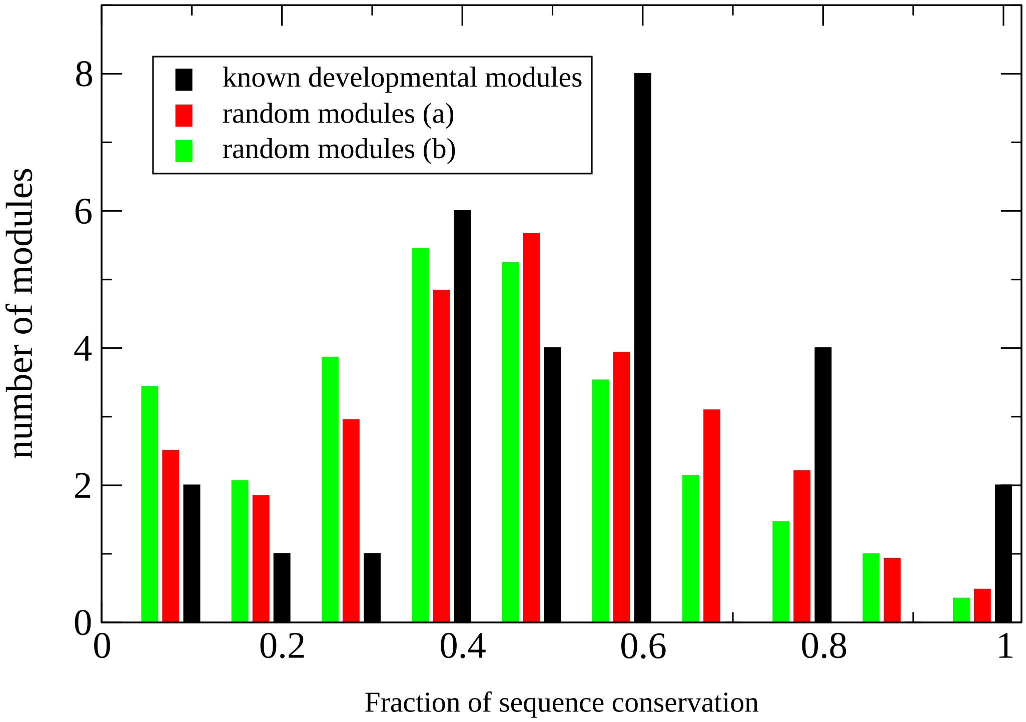

Figure 1
The fraction of sequence conserved within the modules of our test set vs random. The distribution in conserved sequence for the 30 test modules compared with two randomized sets. The red histogram distributes the modules randomly within the 200 kb of sequence we aligned around the regulated genes, while the green data is sampled from all noncoding sequence. The conserved blocks were computed from the SMASH alignments with parameters corresponding method (a) in Table 1. There are 10 equally spaced bins for each histogram.

well studied genes and regulatory regions [25]. The intersection with the 200 kb region for which we have annotated regulatory modules in *D. melanogaster* comprises 32 kb of sequence of which 16048 bp falls within known modules. A total of 2200 bp of *D. melanogaster* binding sites fall within these 16 kb of modules.

Our two alignment programs, work by finding the best syntenous assembly of the high scoring pairs (HSP) generated by a nucleotide BLAST comparison between a *D. pseudoobscura* contig and the *D. melanogaster* chromosomes. They differ in how they score the regions between the prominent 'anchor' regions in the alignments, and how they refine the anchors (see Methods). All aligned blocks in the optimal syntenous assembly, we consider 'conserved' between the two species. The alignments depend on the scoring parameters, defined in conformity with those for BLAST. The parameter dependence was eliminated by searching over the entire plausible parameter space (other definitions being fixed) for those parameters that optimized the significance of the correlation, under a null model that the sites were placed at random. The statistical significance of the correlation was assessed by randomizing 100 times the placement of the binding sites within the interval defined by the modules to compute the mean and standard deviation expected by chance (see Methods). The randomized mean was subtracted from the observed correlation and the difference normalized by the standard deviation to define a Z-score, which is a measure of the probability that the observed correlation was obtained by chance. Thus parameters which score the entire module as conserved, would have zero Z-score. Alternatively one could optimize the fraction of conserved binding sites defined as sites in conserved blocks divided by total sequence conserved, with similar

Table 1: The statistics of conserved sequence between pairs of species for a range of parameters that maximized the correlation between the experimental binding sites and the conserved sequence. Two alignment programs were used along with two measures of intersection: (a) site counted only if entirely within a conserved block, (b) number of conserved bases belonging to site. Columns indicate the fraction of the total sequence aligned in conserved blocks (the percent identity of the conserved blocks), the fraction of sites in blocks, the fraction when sites are placed randomly. (see Methods), and the statistical significance. The percentage intervals in the three columns correspond.

D. melanogaster versus D. pseudoobscura				
method	conserved seq. (%)	sites in conserved seq. (%)	random sites in conserved seq. (%)	Z score
SMASH (a)	41–51 (82–92%)	51–71	37–54	5.0
SMASH (b)	33–51 (82–92%)	48–74	42–57	5.0
LAGAN (a)	33–71 (82–92%)	36–53	25–39	5.0
LAGAN (b)	31–54 (82–92%)	50–80	41–58	6.0–8.0
D. melanogaster versus D. virilis				
method	conserved seq. (%) (PID)	sites in conserved seq. (%)	random sites in conserved seq. (%)	Z score
SMASH (a)	25–31 (83–92%)	18–32	13–25	2.0
SMASH (b)	25–26 (83–92%)	36–41	27	4.5
LAGAN (a)	29–56 (81–92%)	23–36	18–32	2.25
LAGAN (b)	30–48 (81–92%)	45–61	30–50	5.0

results. For the optimal aligment parameters, we also tabulate the percent identity in the conserved blocks.

We implemented two definitions to count when a binding site hit a conserved block. The strictest definition gave a count of 1 if the site was entirely within a block and zero otherwise. This could miss sites which were determined to be longer than the minimal size to bind a protein, and also sites near the edge of blocks where the alignment programs can loose bases. The other definition, went to the opposite extreme, and simply counted all bases of overlap between sites and conserved blocks, no matter how few. Overlapping sites were either merged or considered separate and randomized in a consistent way in each case (see Methods). For each definition (ie 4 cases in all) the parameters were optimized separately.

Table 1 shows the optimal correlation between binding sites and the conserved blocks as a function of the species pair and alignment program. The ranges reflect the various definitions used for intersection. Thus while there is always a positive correlation between binding sites and conserved sequence ($Z > 0$), the number of sites in conserved sequence is never much greater than random for any parameter values, any alignment code, or any reasonable definition of intersection. The agreement between the alignment methods is significant since LAGAN tends to find smaller conserved blocks than SMASH. For this reason, when intersection is counted only when the entire site is contained within a block, the fraction of conserved sequence can be larger than the fraction of the conserved sites, but the fraction found for random data is smaller still. While individual binding sites are often preserved between species, few papers actually compute the odds of this occuring by chance, given either a single base mutation rate, or the overall level of sequence conservation in a region surrounding the site, or module.

It is well known that binding sites within modules appear sometimes to be redundant, ie mutation of a single site may not necessarily result in an observable phenotype. Thus, it could be argued that our observed weak correlation between the position of binding sites and conserved chunks of sequence is biased because we perform our randomization tests always within the modules. However, we checked that this is not likely to be true by also randomizing the position of DNA motifs within randomly chosen non-coding genomic sequence. More precisely, we selected 30 kilobases of non-coding genomic sequence in the vicinity of genes which are not known to be regulated by the body patterning factors. For example, for the factor "bicoid", we found 68 occurrences of the core biocid DNA motif of length 8 (compare [4]). 33 of these sites were in conserved chunks of DNA (using our procedure SMASH (b), Table 1), 32 were expected to be conserved by chance. These numbers correspond well to Table 1 and argue against effects which originate from randomizing within modules.

Moreover, the weak correlation we observed between binding sites and conserved blocks does not appear to arise from a subpopulation of our data set (nb all sites have experimental support). Our factors can all be classified genetically within the segmentation gene hierachy as 'maternal-gap', 'pair rule', and further downstream, for which we have respectively 215, 37, and 59 binding sites. For representative parameters the fraction conserved is 47%, 59%, and 56% for the three types, which we consider close, given the small numbers involved. We also asked whether factors with only a few binding sites in a given module (which are generally more specific) are more likely to be conserved. There are classes of sites that do behave this way (eg the 4 copies of torso response element), but their numbers are small. Another meaningful decomposition of the data is between sites with only vitro evidence, and those for which the site was mutated and a change observed when the new module was expressed in a transgenic construct. We checked how many sites with *in vivo* evidence reside in conserved blocks according to our SMASH alignments (Table 1b). Two hunchback sites (out of 3), four bicoid sites (5), and three Kruppel sites (5) in the even-skipped stripe 2 element [12,13] and all 5 ftz sites in the ftz autoregulatory element[10,11,27] were conserved. However, the total number of these sites is unfortunately too small to allow a meaningful statistical comparison to the conservation of sites with *in vitro* evidence.

Module detection

Interspecies comparisons promise to be a powerful way to locate regulatory modules as regions of enhanced conservation [9,15], and for *D. melanogaster* a new module for the apterous gene was found this way [34]. However the number of species and their evolutionary distance required to do this with a quantifiable reliability (which could be gene dependent) is still a matter for debate.

The difficulties in using just a locally averaged percent identity or for us the fraction of sequence in conserved blocks to detect modules are illustrated in Fig 1. For each of our known modules we computed the fraction of conserved sequence; thereby assuming the correct starting position and length, and thus biasing the data to look more conserved that it would if it were scanned with a single fixed window length. Intervals of the same length were then randomized to sample either the 200 kb environment of the genes in our set (a), or the genome at large (b). In both cases the differences in the means of the distributions were smaller than the combined variances, though only (a) is relevant to whether modules can be distinguished from their environment. In the most favorable case, chosing modules over 60% identity as real, would hit 50% of the known cases, but 39% of the regulatory regions around our genes would score as well. There can be considerable variation in the amount of conservation even between genes in the pair rule group eg *eve* vs *run*. For instance the 6 kb '7-stripe' region upstream of runt conserves only 13–25% of its sequence in *D. virilis* (or 20–35% in *D. pseudoobscura*) under our alignment codes [21], ie less than in Table 1.

The extent to which a statistically significant set of regulatory modules can be inferred from direct sequence comparison may depend on the genes and species being compared. The alignment codes we use, do agree with the modules predicted upstream of the *otx* gene in Fig. 3 of [15] most of which were shown to be functional. Perhaps the two fly species we are comparing are not at the optimal evolutionary distance, or the early patterning genes are not indicative of the modules for genes involved in terminal differentiation.

Recently a number of methods have been proposed to discover regulatory modules by looking for clusters of binding sites [4-8], and it is interesting to inquire how the sequence of *D. pseudoobscura* can be used to improve these predictions. The segmentation genes are a very apt test set, since most of the essential targets were discovered in the screens of Wieschaus and Nusslein-Volhard [23], and about a third of all targets are available from the Berkeley insitu data base (see Methods). We took the top 381 module predictions from the Ahab code [4,46] and scored the 674 genes that bracketed one of them as our patterned set. A similar calculation was then done for the *D. pseudoobscura* sequence. From either single genome calculation, we found 45–50 genes that matched the known patterned set. When we intersected the gene predictions from the two species, we were left with 74 genes of which 22 were in the known set; a substantially higher percentage than from a single species. The number of genes known NOT to be blastoderm expressed comprises 4% 674 gene set and 10% of the 74 genes in the intersected sets. To better quantify how the second species improves our predictions, we compared the actual modules predicted from the two species. The intersected sets of modules, however, was not significantly enriched for known modules (a list roughly 10% the size of our gene comparison list). Gene prediction appears to work better than module prediction since blastoderm patterned genes frequently have several modules, which provide multiple targets for module prediction, with a different module being hit in each species. Whether this phenomena generalizes to other classes of genes remains to be seen, but a limited number of signaling pathways are reused in many contexts, so perhaps key targets will be accessible to our approach. For biological applications, the ability to preform an 'insilico' genetic screen is not without interest.

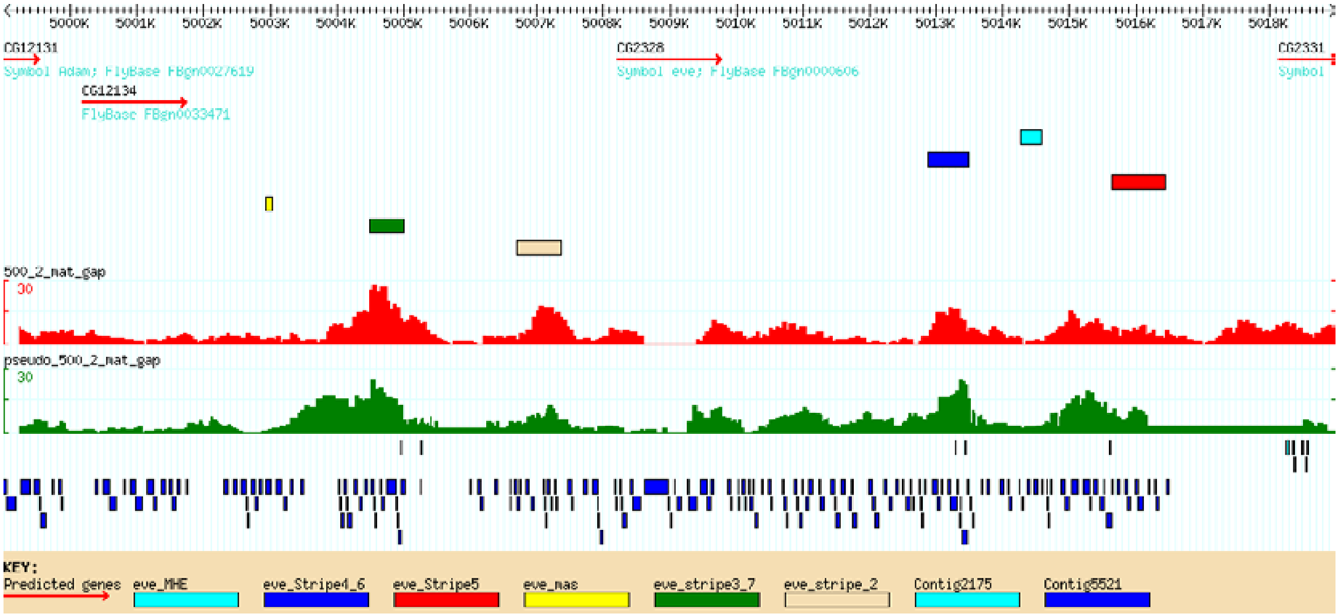

Figure 2
Computational module detection for the *eve* locus in *melanogaster* and *pseudoobscura*. The Ahab score plotted for a 20 kb neighborhood of *eve* as displayed by Gbrowse [44]. The genes are marked as red arrows, and known modules by the colored bars below. The Ahab score for *D. melanogaster* is plotted as the red filled graph below the module bars. The *D. pseudoobscura* score in green is aligned to the *D. melanogaster* sequence by the conserved sequence blocks and linearly interpolated between the blocks. The lowest panel is the conserved blocks as defined by LAGAN, offset vertically for clarity. The right edges of the stripe 3–7 and stripe 5 modules are delimited by gaps in the alignment, but most other boundaries are not. Certain modules score poorly since Ahab was run with the AP gap gene factors only.

Discussion

For an extensive collection of binding sites for factors that pattern the blastoderm, we have quantified their overlap with the blocks of sequence conserved between two *Drosophila* species and found it to be not much greater than chance (but still statistically significant). The fraction of conserved sequence and the size of the conserved blocks do vary with parameters and the algorithm, but our conclusions are unaffected. Since some of the modules we use have been intensively probed for binding sites, there is certainly conserved sequence without obvious function. To the extent that the blastoderm patterning in the fly is a good model for transcription control, our results will generalize beyond the specific case we treated. Phylogenetic footprinting is frequently used to justify current sequencing projects and at least for the factors and species we examined has limited utility in delimiting binding sites. We also found that there was insufficient contrast in the fraction of conserved sequence to distinguish most of the known modules from the rest of the noncoding sequence bordering the genes in our set.

There is a considerable literature which compares the noncoding DNA of other fly species to *D. melanogaster*, but there has not been a systematic test of whether sequence conservation is a good prefilter for functional binding sites. In [22] the regulation of engrailed was examined. A very clumped pattern of conserved and non-conserved sequence was found and some homeodomain binding sites (tested *in vitro*) resided in the conserved blocks. In [20], a clear syntenous alignment of the upstream region of hairy with *D. virilis* and multiple binding sites among the conserved sequence was found. Regulatory modules were moved between the species and it was demonstrated that they correctly drove reporter constructs. The 6 kb regulatory region upstream of runt was compared with *D. virilis* in [21]. They noted a much lower level of conservation than in modules with mostly gap gene input (which our alignment codes reproduce) and loss of some functional regions. Several papers from Dover et al eg [18,19] demonstrate coevolution of bicoid binding sites (regulating hb) and bicoid protein in higher diptera. They identify bcd sites in the homologous modules by consensus sequence and vitro footprinting and do not rely on general sequence alignments. Reference [35] deleted a block of regulatory sequence conserved in *D. virilis* for a vestigial enhancer and observed no change in function. Kim and coworkers [36,37] studied the evolution of the hairy stripes 1,5 enhancer and suggest a role for sequence variation in timing the appearence of these stripes.

Several recent studies in vertebrates were formulated to address the correlation between sequence conservation and functional binding sites. In [14] it was shown that a very high proportion of muscle specific regulatory sites are conserved between human and mouse, and that restricting a Gibbs sampler search to the conserved regions greatly facilitated the recovery of the known regulatory sites. In [16] a 1.6 kb module for the apo(a) gene in primates was analyzed, and a preponderance of functional protein binding to the conserved sequence blocks therein was demonstrated. In other cases (eg [9,39]) the entire module is so well conserved, that individual binding sites should be also, but the interspecies comparisons do not permit binding sites to be delineated.

Other studies examine the conservation of individual binding sites of proven functionality, for situations where the sites can be mapped unambiguously. In [42] 51 regulatory regions were analyzed for which there is data for human and another primate or rodent and it was estimated that perhaps a third of the sites have changed enough so they are not functional in both species. A similar analysis was applied to a number of fly regulatory modules in [43]. More surprising are the results in Ref [40], which documents a high degree of variablity in functional binding sites within the human population alone. The implications of these articles are similar to ours, but the question is formulated differently. We ask whether sequence alignment programs will score known binding sites as conserved, and thus have to contend with the artifacts of these codes. The authors of [40,42] can compare variation within sites against say a neutral rate, provided the site and some of its environment are conserved well enough to be mapped. However, they do not assess what fraction of functional sites can be found by interspecies comparisons (without knowledge of the site in one species), and sites that can not be aligned do not enter their data set.

One limitation of our study should be noted. Our list of known sites, is compiled from many sources, some of which are *in vitro* only. A possible explanation of our results is that the nonconserved sites have less influence in vivo than the conserved ones. This would lessen the importance attached to *in vitro* footprinting and gel shift experiments.

The limited correlation we have observed between known binding sites and conserved sequence has interesting implications for the evolution of regulatory modules which are amenable to immediate experimental test. The binding sites which do not reside in conserved sequence blocks may still be present and even in the same order in both species but either have mutated enough or have moved sufficiently due to insertions/deletions between sites that our alignment program does not recognize them. Another alternative is that the number and type of sites is invariant, but their positions are shuffed. It may be precisely the combinatorial nature of transcriptional regulation that allows many different choices and placements of binding sites inside modules while preserving the function of the module. Reference [38] showed that the *eve* stripe2 enhancer could be moved in toto between *D. melanogaster* and *D. pseudoobscura* with no change in pattern, while a chimeric enhancer made up of a piece from each species generated discernable differences. Thus it is seems that selection can act on the module as a unit and subtle compensations can occur within it as a consequence of genetic drift.

Another interesting possibility is that the expression pattern of the gene is conserved between two species, but the module governing this expression has changed its location or diversified its function, eg by fusing with another module [41]. Our Ahab results are very interesting in this context, since they allow a screen for modules that are very strong in one species and weak in the other. The strong module could be a 'pseudo module' defined in analogy with a pseudo gene, one with an appropriate collection of binding sites, but not operational because of site spacing, nearby insulators, silencers etc. If it is functional, then how is its function implemented in the other species?

Finally, since the function of many blastoderm expressed genes is unknown, and their expression in *insitu* patterns weak, the possbility exists that their expression in the blastoderm provides little selective advantage and is really absent in the related species. In that regard, our list of Ahab predictions present in both species is enriched in genes of known and important function.

There have been very few systematic studies (the globin genes in mammals being perhaps the best studied case [45]) of how well modules can be predicted from interspecies comparisons alone. A number of papers have validated single modules discovered from sequence comparisons, but we do not yet know whether at the same level of significance most functional modules will emerge or just a few. Drosophila is an interesting system in this regard since genetics and promoter bashing have furnished a dense set of tested modules for several genes. The very density of regulation and the absence of 'junk' DNA around the known developmental regulatory genes, may make this system a difficult one to delineate modules without the use of auxilliary information such as we employed with Ahab.

Material and Methods

Alignment algorithms

We used two different algorithms, LAGAN [30] and SMASH [31], to align pairs of noncoding sequences. Both algorithms start by finding local alignments between the sequences and then construct the best syntenous (order preserving) chain of these prominently conserved 'anchor' regions. They differ in how the anchors are found and how the regions between the anchors are treated. LAGAN uses the CHAOS algorithm [32] to find anchors. After the chaining is done a global Needleman-Wunsch algorithm is run to both refine the anchor and align the interanchor sequence. SMASH, works from the local alignments produced by BLAST, and then finds the optimal synte-nous path through them, resolving overlaps with the Needleman-Wunsch algorithm (using the same parameters as BLAST) applied to the anchors only. Thus the interanchor regions are not scored and do not influence the alignment.

The codes have complementary strengths. LAGAN will recover smaller conserved regions that were not found in the initial set of local alignments yet are syntenous with the best chain. However, one set of alignment parameters is used everywhere. SMASH ignores any sequence that was not among the initial set of BLAST, high scoring pairs, which is appropriate if the sequence between the anchors really follows different statistics (ie is single base random). The two codes produce a different spectrum of block sizes, the SMASH ones tend to be longer and may be gapped, while the LAGAN ones are shorter, more numerous and ungapped. The histograms for typical prameters are shown in Figure 3. The density of indels is about 1% in the SMASH blocks for the more restrictive gap parameters, and thus not material to how we score intersection with binding sites. It is around 10% for the more tolerant gap parameters in Fig 3b, so many of the binding sites have gaps, but these are often at the ends and thus the alignment would be nearly as good with the site ungapped. Well conserved regions are found by either method.

Given a list of conserved interspecies blocks represented as intervals, intersections with binding sites can either be scored as binary, ie one only if the site was entirely within the block (ignoring any possible gaps for the SMASH blocks), or by counting the number of bases of overlap. The sites were randomized by placing them randomly within the interval defined by the modules (either up/down stream of the gene). That is for *eve* the binding sites in the stripe 3–7 and stripe 2 modules (both upstream of the gene) were randomized within the entire interval from stripe 3–7 to stripe 2. When a gene had a single module, randomization was done only within its limits. Of the total of 200 kb aligned between *D. pseudoobscura* and *D. melanogaster*, the shuffed sites were distributed among only 30 kb. This restriction was of marginal significance, since the density of conserved sequence in modules was only 3–15% higher than in the rest of the aligned sequence. A number of our experimental sites overlapped, but its not generally known whether this overlap is important for function or parasitic. We therefore treated these cases in two ways: (a) fuse the overlapping sites, randomize them as a unit (but exclude configurations with overlapping sites), and score intersections with the composite site as defined above; or (b) treat the sites as independent in the randomization, allow intersections in the result, and score intersection with each site separately.

Both algorithms assume a common set of alignment parameters, and we explored all combinations within the following sets; match = 1, mismatch = -1,-2,-3, gap start = -2,-4,-6,-8,-10, and gap continue = 0, -1, -2, -3. For each parameter set, a total of 4 calculations were done to cover our definitions of intersection and site overlap, and parameters selected to optimize the significance of the intersection. The percentage intervals in Table 1 encompass the 4 definitions checked, as well as all parameters with scores within 10% of optimal. For Lagan we found that the parameters, gap start = -6, gap continue = 0, and mismatch = -1, -2 gave the best Z score, with the mismatch of -2 giving smaller ungapped blocks than a mismatch of -1 (this was true for both species). For SMASH, good Z scores could be achieved with paramters that generated large blocks and parameters that generated smaller blocks. Large SMASH blocks were achieved using gap start = -4, gap continue = -1 and mismatch of -1. Smaller blocks with a good Z score were achieved using gap start = -4, gap continue = 2, mismatch -2.

Extraction of patterned melanogaster *genes*

There are approximately 1500 genes whose embryonic expression patterns are available from http://www.fruitfly.org/cgi-bin/ex/insitu.pl. Of these 229 have been annotated by Schroeder and Gaul (private communication) as blastoderm patterned, half as strong. An independent literature survey generated 74 patterned genes of which a third where among those chosed for the *in situ* analysis. Thus we have a set of 279 known patterned genes against which to compare predicted modules, and can estimate that the true number is about 700 (ie by scaling the fraction of the literature set recovered).

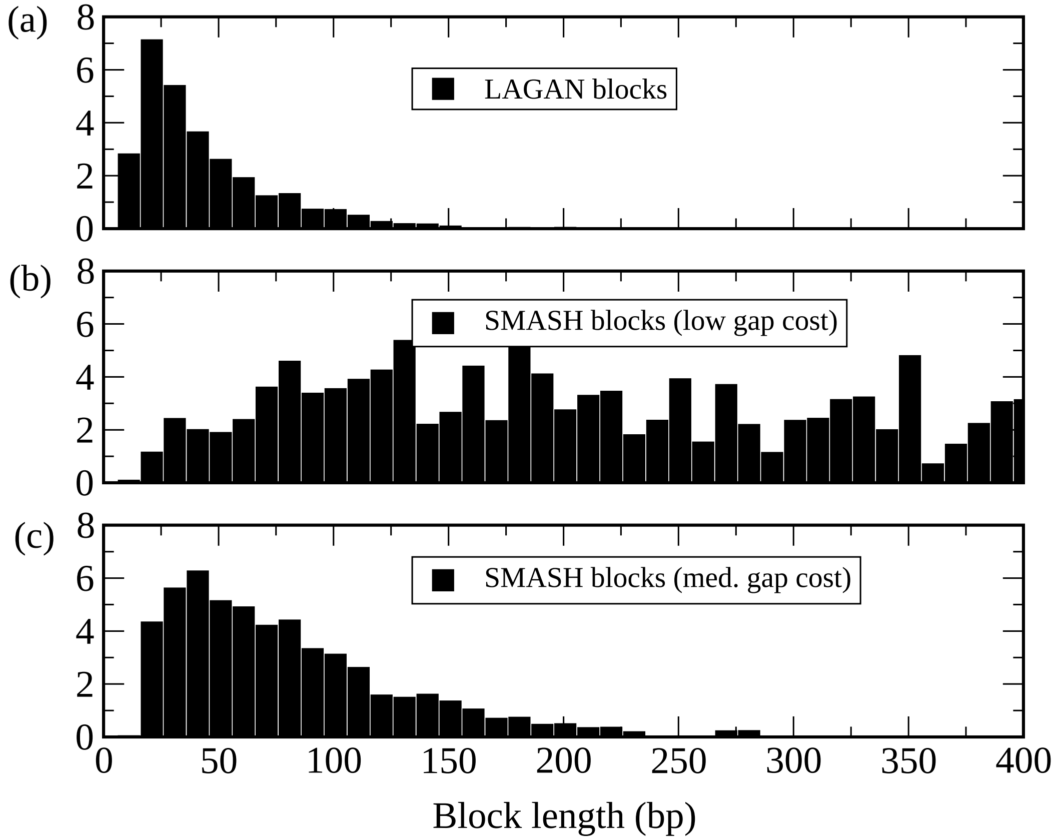

Figure 3
Histograms of the amount of conserved sequence in blocks of various sizes. The data is from the comparison with *D. pseudoobscura* and the parameters are match 1, mismatch -1, gap start -6, gap continue 0 for LAGAN and (1, -2, -2, -1), (1, -2, -4, -2) for the two SMASH runs respectively.

Computational identification of putative modules

We ran the Ahab algorithm [4] independently on the melanogaster (BDGP release 2) and pseudoobscura [26] sequences. Non-ACGT letters in the pseudoobscura sequences were replaced by N's. Tandem repeats in all sequences were masked as described in [4]. We used weight matrices for the factors Bcd, Cad, Hb, Kr, Kni, and Tll from [4]. The window length was set to 500 bp, the Markov model for background sequence was based on 2-mer counts, and the cutoff for the score was set to 17.0. Local maxima in the Ahab score constituted our predictions. We used the BDGP release 2 annotation to extract genes proximal to Ahab predictions.

Given a module prediction on *D. pseudoobscura* we needed to determine the two closest genes or one, if the module fell interior to a gene. We used BLAST to position all the *D. melanogaster* proteins on the contigs and then used SMASH to align a 20 kb region of the *D. melanogaster* chromosome to the contig. The hits were then ranked by SMASH score and the best hit to a given contig determined the chromosome it mapped to. All other genes from that chromosome with a reasonable SMASH score to the contig were placed, and those nearest to the module recorded.

Authors contributions

Authors 1, 2 and 3 carried out this study. All authors read and approved the final manuscript.

Acknowledgements

We thank C. Bergman for sharing the *D. virilis* data used in ref [25]. Mark Schroeder and Ul-rike Gaul kindly supplied us with the annotated list of *in situ* experiments from the BDGP, http://www.fruitfly.org/cgi-bin/ex/insitu.pl. U. Gaul and S. Small provided useful comments on the manuscript and references. Support was provided by the NFS (DMR 0129848) and the NIH (GM66434).

References

1. Davidson EH: **Genomic regulatory systems.** *Academic Press San Diego* 2001.
2. Carroll SB, Grenier JK, Weatherbee SD: **From DNA to Diversity.** *Blackwell* 2001.
3. Davidson EH, McClay DR, Hood L: **Regulatory gene networks and the properties of the developmental process.** *Proc Natl Acad Sci USA* 2003, **100:**1475-80.
4. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules, applied to body patterning in the early *Drosophila* embryo.** *BMC Bioinformatics* 2002, **3:**30.
5. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci USA* 2002, **99:**757-762.
6. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci USA* 2002, **99:**763-768.
7. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-28.
8. Rebeiz M, Reeves NL, Posakony JW: **SCORE: a computational approach to the iden-tification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation.** *Proc Natl Acad Sci* 2002, **99:**9888-93.
9. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W *et al.*: **Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.** *Science* 2000, **288:**136-140.
10. Schier AF, Gehring WJ: **Analysis of a fushi tarazu autoregulatory element: multiple sequence elements contribute to enhancer activity.** *EMBO J* 1993, **12:**1111-9.
11. Han W, Yu Y, Su K, Kohanski RA, Pick L: **A binding site for multiple transcriptional activators in the fushi tarazu proximal enhancer is essential for gene expression in vivo.** *Mol Cell Biol* 1998, **18:**3384-94.
12. Arnosti DN, Barolo S, Levine M, Small S: **The eve stripe 2 enhancer employs multiple modes of transcriptional synergy.** *Development* 1996, **122:**205-14.
13. Small S, Blair A, Levine M: **Abstract Regulation of even-skipped stripe 2 in the Drosophila embryo.** *EMBO J* 1992, **11:**4047-57.
14. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26:**225-228.
15. Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH: **Patchy interspecific sequence similarities efficiently identify positive cis-regulatory el ements in the sea urchin.** *Dev Biol* 2002, **246:**148-161.
16. Boffelli D, McAuliffe J, Ovcharenko D *et al.*: **Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome.** *Science* 2003, **299:**1391-1394.
17. Tautz D: **Evolution of transcriptional regulation.** *Current Opinion in Genetics and Development* 2000, **10:**575-579.
18. Shaw PJ, Wratten NS, McGregor AP, Dover GA: **Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera.** *Evol Dev* 2002, **4:**265-77.
19. Hancock JM, Shaw PJ, Bonneton F, Dover GA: **High sequence turnover in the regulatory regions of the developmental gene hunchback in insects.** *Mol Biol Evol* 1999, **16:**253-65.
20. Langeland JA, Carroll SB: **Conservation of regulatory elements controlling hairy pair-rule stripe formation.** *Development* 1993, **117:**585-96.
21. Wolff C, Pepling M, Gergen P, Klingler M: **Structure and evolution of a pair-rule interaction element: runt regulatory sequences in D. melanogaster and D. virilis.** *Mech Dev* 1999, **80:**87-99.
22. Kassis JA, Desplan C, Wright DK, O'Farrell PH: **volutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the Drosophila segmentation gene engrailed.** *Mol Cell Biol* 1989, **9:**4304-11.
23. Wieschaus E: **Embryonic transcription and the control of developmental pathways.** *Genetics* 1996, **142:**5-10.
24. Russo CA, Takezaki N, Nei M: **Molecular phylogeny and divergence times of drosophilid species.** *Mol Biol Evol* 1995, **12:**391-404.
25. Bergman CM, Kreitman M: **Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences.** *Genome Res* 2001, **11:**1335-45.
26. **Baylor website, 2003** [ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Dpseudoobscura/]. Jan. 13 2003 release
27. Pick L, Schier A, Affolter M, Schmidt-Glenewinkel T, Gehring WJ: **Analysis of the ftz upstream element: germ layer-specific enhancers are independently autoregulated.** *Genes Dev* 1990, **4:**1224-39.
28. Florence B, Guichet A, Ephrussi A, Laughon A: **Ftz-F1 is a cofactor in Ftz activation of the Drosophila engrailed gene.** *Development* 1997, **124:**839-47.
29. Kasai Y, Stahl S, Crews S: **Specification of the Drosophila CNS midline cell lineage: direct control of single-minded transcription by dorsal/ventral patterning genes.** *Gene Expr* 1998, **7:**171-89.
30. Brudno M, Do C, Cooper G, Kim MF, Davydov E: **LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA.** *Genome Research* 2003, **13:**721-731.
31. Zavolan M, Rajewsky N, Socci ND, Gaasterland T: **SMASHing regulatory sites in DNA by human-mouse sequence comparisons.** *Proceedings of the 2003 IEEE Bioinformatics Conferene (CSB2003)* :277-286.
32. Brudno M, Morgenstern B: **Fast and sensitive alignment of large genomic sequences.** *Proceeding of the IEEE Computer Society Bioinformatics Conference (CSB2002)* .
33. Papatsenko DA, Makeev VJ, Lifanov AP, Regnier M, Nazina AG, Desplan C: **Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers.** *Genome Res* 2002, **12:**470-81.
34. Bergman CM, Pfeiffer BD, Rincon-Limas DE *et al.*: **Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome.** *Genome Biol* 2002, **3:**.
35. Certel K, Hudson A, Carroll SB, Johnson WA: **Restricted patterning of vestigial expression in Drosophila wing imaginal discs requires synergistic activation by both Mad and the drifter POU domain transcription factor.** *Development* 2000, **127:**3173-83.
36. Kim J, Kerr JQ, Min G: **Molecular heterochrony in the early develoment of Drosophila.** *Proc Natl Acad Sci USA* 2000, **97:**212-216.
37. Kim J: **Macro Evolution of the hairy Enhancer in Drosophila Species.** *J Exp Zoo (Mol Dev Evol)* 2001, **291:**175-185.
38. Ludwig MZ, Bergman C, Patel NH, Kreitman M: **Evidence for stabilizing selection in a eukaryotic enhancer element.** *Nature* 2000, **403:**564.
39. Leung JY, McKenzie FE, Uglialoro AM *et al.*: **Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters.** *Proc Natl Acad Sci USA* 2000, **97:**6614-8.
40. Rockman MV, Wray GA: **Abundant raw material for cis-regulatory evolution in humans.** *Mol Biol Evol* 2002, **19(11):**1991-2004.
41. Rothenberg EV: **Mapping of complex regulatory elements by pufferfish/zebrafish transgenesis.** *PNAS* 2001, **98:**6540-6542.

42. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.** *Mol Biol Evol* 2002, **19:**1114-21.
43. Dermitzakis ET, Bergman CM, Clark AG: **Tracing the Evolutionary History of Drosophila Regulatory Regions with Models That Identify Transcription Factor Binding Sites.** *Mol Biol Evol* 2002, **20:**703-14.
44. [http://www.gmod.org/].
45. [http://globin.cse.psu.edu].
46. [http://gaspard.bio.nyu.edu/Ahab.html].