Reliable prediction of regulator targets using 12 *Drosophila* genomes

Alexander Stark*+[1,2], Pouya Kheradpour*[2], Sushmita Roy[3] and Manolis Kellis+[1,2]

1. Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA.
2. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
3. Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA.

Running Title:

Comparative prediction of motif instances

* shared first authors

+ corresponding authors:

Manolis Kellis
32 Vassar Street, 32G-564
Cambridge, MA 02139

Phone: 617-253-2419
Fax: 617-253-7512

manoli@mit.edu

Alexander Stark
32 Vassar Street, 32G-536
Cambridge, MA 02139

Phone: 617-253-3434
Fax: 617-253-7512

alex.stark@mit.edu

Keywords:

Regulatory motifs and networks, comparative genomics, transcription factor targets, microRNA targets, *Drosophila*

Abstract

Gene expression is regulated pre- and post-transcriptionally via *cis*-regulatory DNA and RNA motifs. The identification of individual functional instances of such motifs in genome sequences is a major goal for inferring regulatory networks yet has been hampered due to the motifs' short lengths that lead to many chance matches and poor signal-to-noise ratios. In this paper, we develop a general methodology for the comparative identification of functional motif instances across many related species, using a phylogenetic framework that accounts for the evolutionary relationships between species, allows for motif movements, and is robust against missing data due to artifacts in sequencing, assembly or alignment. We also provide a robust statistical framework for evaluating motif confidence, which enables us to translate evolutionary conservation into a confidence measure for each motif instance, correcting for varying motif length, composition, and background conservation of the target regions.

We predict targets of fly transcription factors and miRNAs in alignments of 12 recently sequenced *Drosophila* species. When compared to extensive genome-wide experimental data, predicted targets are of high quality, matching and surpassing ChIP-chip and recovering miRNA targets with high sensitivity. The resulting regulatory network, suggests significant redundancy between pre- and post-transcriptional regulation of gene expression.

Availability

All data and predicted transcription factor and miRNA targets are freely available at http://compbio.mit.edu/fly/motif-instances/.

Introduction

Understanding gene expression and its regulation in response to developmental and environmental stimuli is one of the greatest challenges of modern biology. Regulatory control of gene expression occurs at many levels, both pre- and post-transcriptionally, generally based on short DNA and RNA signals known as regulatory motifs. These are recognized in a sequence-specific way by diverse protein and RNA regulators to direct transcription initiation, mRNA export, stability, and translation, ultimately leading to diverse gene regulatory programs in organogenesis and development, and in response to environmental stimuli.

The sequence-based nature of regulatory control should in principle enable computational identification of regulator targets, by recognizing individual motif instances that constitute functional binding sites.

However, due to their short lengths, motifs match very frequently to the genome or in fact any (random) nucleotide sequence by chance alone, and the majority of genome-wide motif occurrences do not lead to functional regulator binding, being either occluded by chromatin structure, separated from necessary co-factor motifs, or otherwise non-consequential to transcriptional regulation (Wasserman and Sandelin 2004). To address the large signal-to-noise problem and predict functional regulatory elements, previous computational approaches have sought regions of motif clustering across several co-operating motifs, which are often associated with enhancer function (Berman et al. 2002; Markstein et al. 2004; Philippakis et al. 2006; Schroeder et al. 2004). Although these approaches have been successful in identifying novel enhancers, which are functional when tested *in vivo*, they only identify a small subset of all functional targets of each regulator, and are only applicable when the specific combinations of factors are already known. In particular, they are unable to identify individual motif instances when these act in isolation, or with diverse sets of co-factors.

Comparative genomics provides a general methodology for distinguishing functional regulatory motif instances, as biologically meaningful elements are typically under negative selection during evolution, the type and extent of evolutionary conservation generally reflecting the specific requirements of the selected function (Miller et al. 2004; Ureta-Vidal et al. 2003). As closely related species often share substantial parts of their morphology and developmental programs, the expression of important genes, their regulatory connections, and the underlying regulatory elements are also likely conserved. In fact, some gene regulatory network kernels involved in organogenesis, such as heart specification, are conserved in species as distant as flies and vertebrates (Davidson and Erwin 2006). Thus, although some processes are subject to more rapid divergence or positive selection (e.g. body color and pigmentation (Prud'homme et al. 2006)), this suggests that comparative genomics at a range of evolutionary distances should allow for the identification of many regulatory components of gene expression programs.

Indeed, previous comparative genomics studies have used the conservation of regulatory elements for the *de novo* discovery of regulatory motifs across related species (Chan et al. 2005; Cliften et al. 2003; Ettwiller et al. 2005; Kellis et al. 2003; Xie et al. 2005). These studies have relied on the average conservation of thousands of motif instances for each regulator, leading to a high genome-wide signal for motif discovery. However, it has remained unclear what fraction of conserved motif instances were functional, and what fraction of functional instances were conserved, namely whether in fact comparative genomics is applicable for high-specificity and high-sensitivity identification of individual motif instances. Moreover, the available genomes have been either too few for sufficient neutral divergence, or too distantly related for motif instances to be conserved (e.g. (Cooper et al. 2005; Ettwiller et al. 2005)). Accurate motif instance identification would thus require many closely related species, which also present

novel conceptual and methodological challenges, with respect to sequence coverage, alignment accuracy, and motif movement, gain, and loss (Boffelli et al. 2003; Cooper et al. 2005; Eddy 2005; Margulies et al. 2003; Margulies et al. 2007; Thomas et al. 2003).

Methods such as phylogenetic footprinting, evolutionary rate profiling and phylogenetic HMMs have been successfully used to identify genomic regions under evolutionary selection (Cooper et al. 2005; Margulies et al. 2003; Margulies et al. 2007; Siepel et al. 2005; Wasserman et al. 2000), but they cannot determine the regions' functions that are selected for. Similar to more complex models of motif evolution (e.g. (Moses et al. 2004; Zhou and Wong 2004)), such methods are often restricted to regions that are well aligned, and can be sensitive to motif-movements or errors in sequencing, assembly, or alignment (Margulies et al. 2007; Moses et al. 2004). Further, methods to predict genomic regions with regulatory potential generally do not allow the identification of regulatory targets for individual factors or miRNAs (Elnitski et al. 2003; Taylor et al. 2006). Lastly, the comparative prediction of miRNA binding sites in 3'UTRs proved successful (reviewed in (Lai 2004; Rajewsky 2006)), but has relied on site presence in defined sets of informant species, and a severe loss of sensitivity has been observed when the number of informant species was increased (Grun et al. 2005; Lewis et al. 2003; Stark et al. 2005).

In this paper, we develop a general methodology for identifying functional motif instances based on their evolutionary conservation across many related species, and provide a robust statistical framework for evaluating motif confidence, enabling us to achieve both high sensitivity and high specificity. Our approach uses a phylogenetic framework, which allows for motif-movements and local alignment inaccuracies, and is robust against missing data due to artifacts in sequencing, assembly or alignment. Our statistical framework enables us to translate evolutionary conservation into a confidence measure for each motif instance, correcting for varying motif length, composition, and background conservation of the target regions.

We apply our framework to whole-genome alignments of 12 recently sequenced *Drosophila* species (Drosophila 12 Genomes 2007; Stark et al. 2007), and predict targets of 83 transcription factors (TFs) and 78 miRNAs (57 distinct families), leading to 46,525 regulatory connections. We use genome-wide ChIP-chip experiments and direct tests of TF or miRNA targeting (independently published by us (Stark et al. 2005; Zeitlinger et al. 2007) and others (Abrams and Andrew 2005; Sandmann et al. 2007; Sandmann et al. 2006; Sethupathy et al. 2006)) to show that computationally-predicted regulator targets are of very high quality, matching and surpassing ChIP-chip sensitivity and specificity, and can identify seemingly functional instances even when these are not bound in the conditions experimentally surveyed. Lastly, we

study properties of the resulting network, which suggest significant redundancy between pre- and post-transcriptional regulation.

Assessing motif-instance conservation across many genomes

Unlike protein-coding and RNA genes, which are typically well aligned in the multiple sequence alignments of related species, many regulatory motifs are too short to guide alignment algorithms, and thus may not appear at orthologous positions in multiple sequence alignments (Wasserman and Sandelin 2004; Wray et al. 2003). As motifs can act at a wide range of distances, individual motif instances may move, either by insertions and deletions, or by 'birth' of new motifs and loss of old motifs via compensatory mutational changes (Ludwig et al. 2000). In addition, individual instances of regulatory motifs may actually diverge across different species, and may experience duplication, gain, and loss across the evolutionary tree (Ludwig et al. 2005; McGregor et al. 2007; Prud'homme et al. 2006). Lastly, comparison of many species introduces new artifacts due to sequencing, assembly and alignment, which may affect the alignment of equivalent regulatory motif instances (see **Figure S1** and (Margulies et al. 2007)).

To account for these unique evolutionary and alignment properties of regulatory motifs, we developed a phylogenetic framework for motif instance identification which tolerates motif movement and loss, while recognizing their clear selective pressure across the phylogenetic tree. Briefly, we search for motif instances in each of the aligned genomes and, given the set of species that contain motif instances within tolerable distances of the *D. melanogaster* instance, we evaluate the total evolutionary branch length over which the motif appears conserved. The overall score of a motif instance becomes this total branch length of the phylogenetic tree over which the motif is conserved, which we call the Branch Length Score, or BLS (**Figure 1**). We thus implicitly assume that all motif instances in *D. melanogaster* are potentially ancestral and count instances in the informant species as evidence when they are conserved. We do not interpret presence/absence patterns of motif instances as evolutionary gain- and loss events, as they could arise from artifacts in sequencing or alignment. The BLS value of a given motif instance ranges from BLS=0.0 (non-conserved) to BLS=1.0 (fully-conserved), representing the fraction of the total phylogenetic tree covered by the species containing the motif.

This BLS conservation measure has many attractive properties, which enable us to define the conservation level of motif instances across a complete genome, to select conservation thresholds for defining all genome-wide instances of a regulatory motif, and to assign confidence values to the observed conservation, as we describe below. Moreover, because missing instances in the aligned species are not interpreted as evolutionary loss-events and are not explicitly penalized, the BLS measure is robust against

missing sequence due to low coverage sequencing, assembly errors or alignment artifacts. Lastly, BLS provides a direct estimate of the expected neutral divergence of the species compared (Felsenstein 2004), accounting for different divergence times between species, correcting for redundant contributions of individual species in a complex tree, and their different rates of divergence (**Figure 1**).

Establishing confidence levels for BLS conservation scores

To translate this BLS conservation score to a robust statistic that can be used across different motifs and different types of genomic regions (e.g. promoters, introns, 5' or 3' UTRs, etc.), we mapped each BLS score to a confidence value between 0% and 100%, representing the probability that a given motif instance is functional. This probability reflects the increased conservation of motif instances compared to overall sequence similarity and is estimated using control motifs, similar to the signal-to-noise ratio for miRNA target predictions (Lewis et al. 2003). Evaluated in a motif- and region-specific way, it corrects for differences in motif length and composition, and for different average conservation levels and nucleotide-composition of different genomic regions. Intuitively, longer and highly-specific motifs are very unlikely to be conserved by chance and thus result in high confidence levels, even for modest BLS thresholds. Further, regions of overall high conservation (such as protein-coding exons) are likely to contain many conserved motif instances by chance alone, and thus require more stringent BLS thresholds to achieve a desired confidence level. Lastly, AT-rich motifs are likely to have many conserved occurrences in AT-rich regions due to chance alone (and GC-rich motifs in GC-rich regions), and thus require higher BLS thresholds if the corresponding control motifs show similarly high conservation.

We found that the number of random motif instances generally decreased rapidly for increasing BLS values, while the number of instances for known motifs remained high (**Figure 2a**). For example, at BLS≥0.50, the motif for *Snail* (CAGGTG) has 229 occurrences in promoter regions, compared to 54 motif instances on average for a pool of 10 control motifs. Therefore, we would expect that of these 229 *Snail* instances, 54 are likely due to chance, while 175 of them (76%) are non-random, leading to a confidence of $C_{0.5}$=76% for each of these motif instances, at BLS=0.5. At a more stringent conservation threshold of BLS≥0.70, *Snail* shows 152 instances while the control motifs show 24 instances on average, leading to a confidence of $C_{0.7}$=128/152=84%. Similarly, the miRNA K-Box motif (Lai et al. 1998) (CTGTGAT; 5' seed motif of *Drosophila* miRNAs 2, 6, 11, 13, and 308) reaches confidence values >75% at a BLS=0.4, and >90% at a BLS=0.76 (**Figure 2a**). We note that the confidence measure is conservative by nature: 76% confidence for example means that 76% of conserved instances are conserved above background and are thus likely functional. The remaining 24% of conserved instances

might contain functional instances that cannot be discerned from noise, suggesting a maximum false positive rate of up to 24%.

We found that with increasing confidence levels motifs were predominantly found in regions in which they are known to function. For example, with increasing confidence the percentage of TF motif instances within promoter regions rises from 20 to 90%, that of miRNA motif instances within 3'UTRs from 25% to 100% (**Figure 2b,c**). In addition, the percentage of miRNA motif instances on the transcribed strand of 3'UTRs rises from essentially random (uniform 50%) to exclusively on the transcribed strand (100%), while promoter motifs do not show any strand preference (**Figure 2d**). These results illustrate the effectiveness of region-specific confidence values (which require more stringent BLS thresholds for more highly-conserved regions), as high-confidence motif instances were not simply biased towards regions with overall high conservation, but specifically selected in regions they are known to act.

Effect of allowing motif movements on instance identification.

Using confidence cutoffs also allowed us to assess the influence of tolerating motif movements on the recovery of functional motif instances. Allowing for motif movement permits capturing functionally equivalent instances across genomes, independent of their relative positions in the alignment. However, while this approach will always increase the number of conserved instances recovered for real motifs, it also increases the number of spurious motif instances that appear conserved due to increased background conservation for large tolerated movements.

The number of motif instances recovered at a given confidence value presents a robust measure of overall discovery power, as it evaluates sensitivity at a fixed specificity. If the window of tolerated motif movement is too small, many true motif instances will be missed. Conversely, if the window of tolerated motif movement is too large, we would expect both real and control motifs to show increased conservation, thus reducing the confidence and leading to fewer confidently identified instances. Between these two extremes, we would expect the number of high-confidence motif instances to peak for an optimal window of tolerated motif movement, and decrease for lower or higher values.

Indeed, we found that allowing for motif movements of 10 to 500 nucleotides relative to the *D. melanogaster* instance often increased the number of confident motif instances, while allowing for large movements generally decreased this effect (**Figure 3b**). Different window sizes were optimal for different motifs: longer TF motifs with higher information content peaked for longer windows (Pearson correlation 0.40 for information content, 0.33 for length), while motifs with many matches in *D. melanogaster* showed shorter optimal windows (Correlation -0.27 for TF, -0.26 for miRNA motifs). We also found a

correlation with GC-content for miRNA motifs (0.28), as expected since 3'UTRs are AT-rich, but there was very little correlation for TF motifs (-0.15). These results illustrate the more rapidly increasing noise levels for motifs with low information content, while motifs with higher information content are less likely to appear by chance within the length of the tolerated window.

Overall, the single best window improved the recovery of 56% of TF motifs (20 nucleotides), and of 71% of miRNA motifs (50 nucleotides; both at 60% confidence). For 71% TF motifs, some window between 10 and 500 nucleotides improved sensitivity, and the improvement was substantial for 11% (at 60% confidence; $P \leq 0.05$ after Bonferroni correction to account for testing multiple windows). Similarly, 93% of miRNA motifs showed improved sensitivity, which was substantial for 13%. Improvements were observed over a wide range of confidence cutoffs, showing that tolerating motif movement is important at any desired confidence level for motif instance identification. These results confirm our intuition that indeed, many motif instances are offset considerably in the 12-species alignments, whether due to alignment artifacts or evolutionary plasticity of regulatory motifs.

BLS measure enables increased sensitivity

The confidence measure also enabled us to gauge the sensitivity of the BLS measure, measured as the number of instances recovered at a fixed specificity, compared to different methodological choices. In particular, we asked whether requiring perfect conservation across fewer species (the 9 *Sophophora* subgroup species, the 4 melanogaster subgroup species, and *D. pseudoobscura* as the only informant) would lead to higher sensitivity / specificity levels, perhaps due to many lineage-specific motifs.

We found that the BLS measure across all 12 species recovered most instances for all TF and miRNA motifs, at all confidence levels (**Figure 3a**). For TF motifs, our approach recovers more than 1.4-fold more instances than the second most sensitive of the other approaches at 60%, 1.5-fold more at 70%, and 3-fold more at 80% confidence (for miRNAs motifs, 1.8-fold more at 60%, 2-fold more at 70% and 1.8-fold more at 80% confidence). When comparing the 3 other approaches for confidence thresholds below 65%, we found that perfect conservation was indeed more sensitive across the 4 closely related species in the melanogaster subgroup and in *D. pseudoobscura* compared to perfect conservation across 9 Sophophora species. However, very few motifs reached higher confidence levels, due to the high overall sequence similarity between these species, resulting in an apparent drop in motif recovery. We also found that the discovery power in *D. pseudoobscura* was comparable to the 4 melanogaster species, likely due to its position in the phylogenetic tree.

Lastly, the BLS and confidence measures allow us to gauge the effect of additional species. We found that evaluating motif conservation across all 12 species allowed more motifs reach confidence levels of 60% than was possible with the other species combination and led to higher average signal-to-noise ratios than any other species combination for TFs and miRNAs (**Figure 3c**).

These results show that the discovery power for target gene identification continues to increase even with more distantly related species. The usefulness of distant species only becomes effective by the use of the BLS measure, while the inclusion of distantly related species resulted in lower performance when perfect conservation was required. Overall, the combination of additional species, and a phylogenetic framework for evaluating motif conservation allowed high-sensitivity and high-specificity in motif instance identification.

Conserved motif instances identify functional *in vivo* targets

We then compared our computationally determined conserved motif instances with experimentally determined *in vivo* targets of known regulators. To define *in vivo* targets, we used several large-scale experimental datasets: a set of high-confidence direct *CrebA* targets confirmed with a variety of reporter assays (Abrams and Andrew 2005), three genome-wide Chromatin IP experiments for developmental TFs with known motifs (*Snail*, *Mef-2*, and *Twist*) (Sandmann et al. 2007; Sandmann et al. 2006; Zeitlinger et al. 2007), and a set of experimentally confirmed targets for different miRNAs (Sethupathy et al. 2006; Stark et al. 2005). We note that the experimentally validated miRNA sites were initially predicted based on conservation to *D. pseudoobscura*, and thus are biased towards higher conservation (already showing BLS>0.26). However, the *CrebA* and the three ChIP datasets were determined independently of any comparative information, and thus provide an entirely independent evaluation of our methodology, allowing us to estimate both sensitivity and specificity of our predictions.

For each regulator, we compared motif instances at different confidence cutoffs with the experimentally derived *in vivo* targets. We found that motif instances at increasing confidence thresholds strongly enriched for experimentally derived *in vivo* targets (**Figure 4a**). In absence of any comparative information, *Mef-2* motif instances in *D. melanogaster* showed no enrichment for experimentally derived targets, while conserved instances showed up to 5-fold enrichment (at 60% confidence). Similarly, enrichment rose from 3-fold to 7-fold for *Snail* at increasing confidence levels, from 4-fold to 9-fold for *Twist*, and from 4.5-fold to 12-fold for *CrebA (*$P=4x10^{-11}$, $3x10^{-10}$, $2x10^{-6}$, and $1x10^{-7}$ *at the highest confidence for the four factors)*. This illustrates the ability of evolutionary information to select for functional motif occurrences, experimentally shown to be bound and/or functional *in vivo*. In fact, the enrichment was most pronounced for *CrebA* (12-fold enrichment; $P=1.4x10^{-7}$), for which the targets had

been shown to be direct transcriptional targets, while some of the ChIP-derived targets may reflect indirect binding, or binding that is non-consequential for transcription.

We also found that even stringent confidence thresholds recovered a large fraction of experimentally derived *in vivo* targets, illustrating the high sensitivity of our approach (**Figure 4b**). When ChIP-bound motifs overlapped experimentally defined enhancer elements(Sandmann et al. 2007; Sandmann et al. 2006; Zeitlinger et al. 2007), 65% *Mef-2*, 65% *Snail*, and 25% *Twist* motif instances were recovered at our 60% confidence cutoff. The lower rate for Twist was possibly due to an overly specific Twist motif used (Markstein et al. 2004). Recovery was again highest for *CrebA*, for which 76% of motif instances were conserved, illustrating the high sensitivity of comparative genomics methods for validated direct targets (**Figure 4c**).

Recovery was much lower when all ChIP-bound regions were considered, regardless of enhancer information, suggesting that some of the ChIP-derived targets may be due to noise, and that conservation is able to pinpoint functional enhancers within ChIP-bound regions. Lastly, we recovered 90% of miRNA motif instances in experimentally confirmed targets at 80% confidence (Sethupathy et al. 2006; Stark et al. 2005) (**Figure 4d**), showing that despite the added branch length (from BLS>0.26 for *D. pseudoobscura* to BLS>0.60 at 80% confidence across the 12 genomes on average), our methods maintain high sensitivity.

In contrast to evaluating conservation by the BLS methodology, requiring perfect conservation across all 12 *Drosophila* species or across the 9 *Sophophora* species recovered significantly fewer experimentally validated motif instances for TF and miRNA motifs (see above and **Figure S2**).

Non-conserved binding events show decreased functional enrichment

Although the overlap between conservation derived motif instances and *in vivo* binding was highly significant and we recovered a substantial fraction of ChIP-bound enhancers, *CrebA* targets, and miRNA targets, we noted that numerous motif instances in ChIP-bound regions were not conserved above 60% confidence, especially for regions that had not previously been shown to be enhancers (**Figure 4b**). Non-conserved sites might be functional, but missed due to unusually large motif movements or sequencing and alignment errors. Alternatively, they may play roles with only lineage-specific selection (and thus not meeting our 60% confidence threshold) or represent largely non-consequential binding, without a specific biological role subject to evolutionary selection. To distinguish the two possibilities, we studied the enrichment of conserved and non-conserved motif instances of the mesodermal factors *Mef-2*, *Twist*, and *Snail* in muscle genes.

We found that ChIP-bound motif instances that were evolutionarily conserved showed enrichment or depletion in promoters of muscle genes for all three factors: the transcriptional activators *Mef-2* and *Twist* showed 8-fold and 7-fold enrichment respectively, and *Snail*, a mesodermal repressor, showed 3-fold depletion in muscle genes. In contrast, ChIP-bound motif instances that were not conserved showed only 1 to 2-fold enrichment or depletion for all three factors (**Figure 4e**). This suggests that potential lineage-specific roles corresponding to non-conserved ChIP-bound sites may lie outside the regulators' conserved functions in core development processes (e.g. mesoderm/muscle development). Alternatively, these sites may be of decreased biological significance, perhaps representing non-consequential binding sites with no role in gene-expression regulation, which are known to be recovered in ChIP experiments (Boyer et al. 2005; Lee et al. 2006). In either case, our results show that non-conserved sites are not simply due to low sensitivity of comparative methods, but are functionally distinct from conserved sites.

ChIP-derived and conservation-derived targets show comparable functional significance

Interestingly, evolutionary conservation identified many high-confidence motif instances outside ChIP-bound regions. These may be functional sites reflecting higher coverage for conservation-derived targets, or spurious sites reflecting noise in the methodology. To distinguish the two possibilities, we used the correlation of these additional motif instances with muscle genes, providing an independent assessment of the overall quality of our predictions.

We found that conservation-derived targets outside ChIP regions were enriched in the same categories in which the factors are known to act. In fact, even outside ChIP regions, conserved sites showed comparable or higher enrichment or depletion in muscle genes than those identified by the ChIP methodology (**Figure 4f**), suggesting they may be of similar overall quality. For *Twist*, enrichment was 1.3-fold higher, for *Snail*, depletion was 2.5-fold higher, and for *Mef-2* enrichment was slightly lower (0.9-fold). Overall, when assessing ChIP- and conservation-derived targets independently (i.e. considering all ChIP targets and all conservation-derived targets), our approach showed a consistently higher enrichment or depletion in muscle genes than ChIP- chip (1.4-fold for *Twist*, 2-fold for *Snail*, and 1.01-fold for *Mef-2*; **Figure 4f**).

Our results suggest that the additional sites outside ChIP-bound regions are likely functional and reflect the higher coverage of conservation-derived targets as compared to experimentally derived targets. Indeed, while ChIP-derived targets are constrained by the developmental stages or cell types surveyed, comparative approaches capture all conserved gene targets regardless of their spatial or temporal constraints. Moreover, comparative approaches are not constrained by the abundance of TFs at bound sites, but only by the strength of evolutionary selection; they can thus identify important sites even when

these are bound more rarely (or in few cell types). Lastly, comparative genomics enables us to capture additional functional targets that may be missed due to experimental limitations of ChIP technology, for which reported false negative rates are up to 30% (Boyer et al. 2005; Lee et al. 2006).

A regulatory network of *D. melanogaster* at 60% confidence

We conclude that comparative genomics provides a powerful methodology for identifying functional targets showing high sensitivity and high specificity. For factors with experimentally determined *in vivo* binding sites, we showed that evolutionary conservation provides comparable discover power as ChIP, and importantly reveals additional functional sites that potentially function at stages or tissues not surveyed. More generally, even when ChIP studies are not available, comparative genomics can provide a first overview of the regulatory connections across a complete genome.

We used our comparative approach to present an initial regulatory network of *D. melanogaster* at 60% confidence for both pre- and post-transcriptional regulators (**Figure 5**). Overall, 49 of 57 miRNA motifs (86%) and 67 of 83 TF motifs (81%) had instances with confidence values of 60% or higher and were considered (**Tables S1 and S2**). The remaining motifs may have too few physiologically relevant and conserved target sites to discern them reliably from background, or may not accurately reflect the factors binding properties, potentially being overly specific or degenerate.

We find a total of 46,525 regulatory connections for TF motifs and 3,662 for miRNA motifs, targeting 8,287 genes and 2,003 genes, respectively. The distribution of targets is highly asymmetric: while we find on average 123 targets per TF motif and 41 targets per miRNA motif, some TF motifs have up to 4,129 targets (homeobox factors), and some miRNA motifs more than 150 targets (miR-4, miR-92, and miR-1). We note, that some motifs (e.g. the homeobox TF motif or the K-box miRNA motif) correspond to multiple TFs or miRNAs, and thus the numbers likely represent combined targets for all individual factors. The distribution of target sites per gene (in-degree) is also highly imbalanced: while a typical gene is regulated by 6 different TF motifs and 2 different miRNA motifs on average, some genes have targeted by up to of 33 different TF and up to 14 different miRNA motifs. Genes with high in-degree were enriched in morphogenesis, organogenesis, neurogenesis, and a variety of tissues, while genes with small in-degree were enriched in ubiquitously expressed or maternal genes with functions in DNA, RNA, or protein metabolism for both TF and miRNA motifs (**Table S3**). Many genes with high in-degree were TFs ($P<10^{-9}$ for TF and miRNA motifs) and transcriptional regulators were indeed more densely targeted than other genes, by both TF- (10.1 vs. 5.5, $P<10^{-20}$) and miRNA motifs (2.3 vs. 1.8, $P<5x10^{-5}$). The similarity between the TF and miRNA motif network was further illustrated by mutual enrichment: genes

with high TF in-degree are enriched in genes with high miRNA in-degree ($P=8x10^{-5}$), as are genes with low in-degree for both types of regulators ($P=2x10^{-7}$).

This initial network contained many connections with independent support in the literature (**Figure 5; Table S4**). For example, we identified the direct regulation of *achaete* by *hairy* (Van Doren et al. 1994), several direct targets of *Suppressor of hairless Su(H)* in the *enhancer of split E(spl)* complex (Bailey and Posakony 1995), direct regulation of the gap gene *giant* by *bicoid* (Kraut and Levine 1991). In addition, the network proposed many novel connections supported by experimental evidence, including a direct regulation of *bagpipe* by *tinman*, which both cooperate in mesoderm induction and heart specification (Yin and Frasch 1998). More generally, when tissue-specific expression data was available, we found that on average 46% of all targets were co-expressed with their factor in at least one tissue (**Figure 5**), which is significantly higher than expected by chance ($P=2x10^{-3}$).

Discussion

We showed that comparative analysis of many related genomes allows us to identify functional motif instances with very high confidence. Overall, 86% miRNA motifs and 81% TF motifs had instances with confidence values of 60% or higher. The remaining factors may have too few physiologically relevant and conserved target sites to discern them reliably from background, or may contain inaccuracies in their binding site motifs might be artificially specific or degenerate.

We found that the availability of many genomes allowed for very high signal-to-noise levels for many motifs at the most stringent settings. However, more importantly, we showed that the BLS measure allowed us to use the increased number of species to strongly increase sensitivity at any given specificity compared to requiring perfect motif conservation in arbitrary subsets of species. While requiring perfect conservation across many genomes is of limited use, the increased power enables approaches that account for artifacts in sequencing, assembly and alignment, and tolerate diverged, missing, or moved motif instances. Our BLS measure is more generally applicable to PWMs (Stormo 2000), to more complex models of regulatory motifs that account for dependencies between individual motif positions (Naughton et al. 2006; Yada et al. 1998), and to more advanced rules for miRNA-target recognition that for example score the contribution of the 3'pairing energy (Brennecke et al. 2005; Stark et al. 2003).

We found that comparative genomics and ChIP-chip showed similar power for functional target identification. The two approaches are complementary, each with unique advantages: conservation helps pinpoint evolutionarily selected functional targets across all conditions, while ChIP-chip reveals stage- and tissue-specific binding *in vivo*, as well as species-specific sites which may play important

evolutionary roles in the emergence of new functions. As motifs of additional regulators are derived by experimental (e.g. by SELEX (Tuerk and Gold 1990) or protein-binding micro-arrays (Mukherjee et al. 2004)) or computational approaches (e.g. by motif-overrepresentation (Tompa et al. 2005) or genome-wide motif-instance conservation (Kellis et al. 2003; Xie et al. 2005)), and tissue-specific binding becomes available for dozens of factors (e.g. through the ENCODE and modENCODE projects), comparative studies can help establish and refine their genome-wide targets. Indeed, we found that motif instances identified by both approaches had the highest functional enrichments, suggesting that combined approaches may prove useful in the future. Although the regulatory network we present likely lacks many true regulatory relationships that could not be reliably recovered, our comparison with ChIP-chip data and other validated targets showed that the network is of high overall quality. We anticipate that the network and the predicted regulatory connections prove to be a useful resource for the fly community working on the biology of TFs or miRNAs and their target genes, and their roles in development. The methodology to assess motif conservation across many genomes and predict functional motif-instances with high sensitivity is more generally applicable for the study of any genome.

Methods

Regulatory motifs

We obtained TF motifs from Transfac (Matys et al. 2003), Jaspar (Sandelin et al. 2004), FlyReg (Bergman et al. 2005), and the literature. To remove redundancy for global statements about motif-targets, we clustered TF motifs using centroid-linkage hierarchical-clustering with a Pearson correlation coefficient cutoff of 0.8 (calculated on the columns of the equivalent PWM) at the best alignment offset (Gupta et al. 2007; Pietrokovski 1996; Schones et al. 2005; Xie et al. 2005). To avoid the creation of artificial motifs by averaging, we chose the original motif from each cluster that is closest to the cluster average as the cluster representative. We defined miRNA motifs as the non-redundant set of 7mers reverse complementary to miRNA 5'ends positions 2-8 (seeds after (Lewis et al. 2003)) for all Rfam miRNAs (Griffiths-Jones et al. 2006). We represent all motifs as consensus sequences over an alphabet of 15 characters (IUPAC code, http://www.chem.qmul.ac.uk/iupac/) consisting of the four nucleotides A,C,G,T, the six two-fold degenerate characters S=(CG), W=(AT), Y=(CT), R=(AG), M=(AC), K=(GT), the four three-fold degenerate characters H=(ACT), B=(GCT), V=(G,A,C), D=(G,A,T) and the four-fold degenerate character N=(ACGT). A motif instance (or motif occurrence) is a sequence that matches the motif at each position, i.e. containing one of the allowed characters at that position.

We translate consensus sequences to PWMs given the definition of the degenerate characters. We translate PWMs to consensus sequences by choosing the character with the highest sum of the PWM column entries corresponding to that character minus a correction for character degeneracy (1/2 for ACGT, 2/3 for SYRMK, 5/6 for HBVD, and 1 for N).

Genome alignments and annotation

For all analyses, we used whole genome MULTIZ alignments of 12 *Drosophila* genomes (Stark et al. 2007), available from UCSC (Kent et al. 2002). We used the *Drosophila melanogaster* genome-annotations from FlyBase (Release 4.3), and excluded simple repeats, repeat masked regions obtained from UCSC, and non-coding exons according to FlyBase 4.3.

Motif matching and BLS measure

We searched all motif instances in the *Drosophila melanogaster* genome and evaluated their conservation in the 12 species using the whole-genome alignments. For each motif instance in *D. melanogaster*, we recorded all instances in the other genomes that were aligned, allowing for motif movements (see below). We prevented double-counting of motif instances by assigning each instance in an informant species to the closest instance in *D. melanogaster*. We evaluated the conservation of all motif instances by summing the branch-lengths of the sub-tree of the species with conserved motif-instances (BLS). This procedure implicitly assumes that all instances are potentially ancestral, such that an instance conserved in a remote informant species would score more highly than instances in closely related informants. One disadvantage of this approach is therefore that chance occurrences or gains in distant species may contribute false positives. The phylogenetic tree branch lengths were obtained from a whole genome alignment of all 12 species (Dewey et al. 2006; Stark et al. 2007).

P-values

All P-values are calculated based on the hypergeometric distribution, and correction for multiple-testing was done with the Bonferroni correction.

Allowing for motif movements

When assessing motif conservation, we allowed motif instances in the informant species to be offset relative to the alignment position of the *D. melanogaster* instances within a given window (counted as distance in either direction in characters excluding gaps). We did not use a prior for a cutoff on maximal tolerable motif movement, as we are not aware of a systematic experimental study that assessed typical movements of functionally equivalent motifs in related species, nor a systematic assessment of the

maximum movement tolerable while maintaining function. We consequently used the window that maximized signal over noise.

While it is clear that increasing tolerated windows may capture additional equivalent instances across genomes, thereby increasing sensitivity, they also increase the number of spurious motif instances that are recovered by chance. We account for the increased background conservation by the use of control motifs (see above), and determine the optimal allowable motif movement window (the one that recovered most motif instances) out of 32 windows between 0 to 500 nucleotides (0, 5, 10, 20, 30, ..., 90, 100, 120, 140, …, 480, 500). For figure 4b and for analyzing the correlation of optimal window size with different motif properties, we assessed 119 windows between 0 and 10,000 nucleotides (0, 10, 20, …, 190, 200, 300, …, 9,900, 10,000). Similarly, we allow for strand reversals of TF motif instances in informant species, when they help instance recovery in the respective windows. The significance of sensitivity improvement for individual windows and for allowing windows in general was assessed by hypergeometric P-values compared to motif instances identified with a window of 0 nucleotides, i.e. perfect alignment of instances.

Estimation of confidence levels of motif instances

For each motif and type of genomic region (promoter, 5'UTR, 3'UTR, intron, etc.), we created 100 shuffled control motifs and selected those that had a similar number of matches to the region in the *D. melanogaster* genome (+/- 20%). By requiring the control motifs to have occurrence rate similar to real motifs in the respective genomic regions in *D. melanogaster* (i.e. without conservation), we corrected for biases in di- or tri-nucleotide frequencies (see discussion in (Lewis et al. 2003)). To remove possible redundancy, we clustered the control motifs (cutoff 0.8) and selected only one representative per cluster, limiting to 10 motifs total that were least similar to known motifs. For each real motif and its controls, we computed the conservation rate (the number of conserved instances at a given BLS cutoff divided by the total number of instances in the *D. melanogaster* genome) in each region and at each BLS cutoff. We determined the confidence at each BLS as the fraction of conserved motif instances above background conservation, where the latter was estimated using the conservation ratio of the control motifs. This provided a BLS-to-confidence mapping for each motif and region. The variation between the control motifs lead to an average standard-error of 5% for TF motifs, and 4% for miRNA motifs at 60% confidence, indicating an accurate assessment of background conservation.

Comparison with experimental datasets

We obtained all experimentally validated miRNA target gene pairs from Tarbase (Sethupathy et al. 2006) and our previous study (Stark et al. 2005). We obtained ChIP-chip regions and the subset that overlapped

known enhancers from (Sandmann et al. 2007; Sandmann et al. 2006; Zeitlinger et al. 2007) and *CrebA* target genes from (Abrams and Andrew 2005). We calculated the enrichment of sites at different confidence cutoffs between 3'UTRs of validated miRNA/target pairs and all 3'UTRs, and between ChIP regions within 2kb upstream regions and the union of all 2kb upstream regions. As CrebA targets were originally defined through mostly 5'UTR instances (Abrams and Andrew 2005) and Mef-2 showed considerable overlap with 5'UTR regions, we included the 5'UTR and restricted the upstream region to 500bp instead. We assessed the recovery of motif-instances as the fraction of motif-instances in in the functional regions (with the same restrictions) that reached the indicated confidence. To assess the fraction of these that are expected by putatively increased overall conservation in these regions, we assess the recovery of control motifs at the same BLS (not confidence, as the control motifs – by definition – would not reach high confidence levels).

Evaluation of experimental and motif instances by correlation with muscle genes

We used correlation with expression patterns to independently evaluate ChIP-regions and predicted motif instances. Muscle genes were 616 genes annotated as "muscle system (13-16)" by the manually curated BDGP *in situ* database (ImaGO) (Tomancak et al. 2002). To obtain a unique assignment of regions to genes, we restricted our analysis to the 5' UTR and 500 bases upstream of each gene. We calculated functional enrichments as the fraction of nucleotides covered by motif instances (at 60% confidence) or ChIP regions in muscle genes divided by the corresponding number in all genes present in ImaGO. Hypergeomertic p-values were computed for motif instances using control motifs at the same BLS and window and for ChIP regions using the fraction of muscle genes matched versus the fraction of all genes matched (note that individual nucleotides are correlated, such that nucleotide P-values would overestimate the significance).

Assessing the indegree distribution

We assessed the non-randomness of the indegree distribution against a control Erdos-Renyi random network (Bollobás 2001) with the same number of edges. To construct this network, we added edges by selecting a source and target node with probability 1/m and 1/n, where m and n were the number of source and target nodes in the true network, respectively. We assessed the difference of indegree distributions between the true and control network with a Wilcoxon rank sum test. We also assessed the difference in indegree distribution between all transcription factors (as defined by (Adryan and Teichmann 2006)) and all other genes also with a Wilcoxon rank sum test.

Functional/Imago enrichment of high and low indegree genes

We considered all genes with a GO (Ashburner et al. 2000) and ImaGO (Tomancak et al. 2002) functional annotation (n=7,495 and 5,996, respectively) and computed the indegree (number of incoming edges) for each gene in the transcription factor (TF) and miRNA networks. For both networks we defined high indegree nodes as the 1% with the highest indegree (>=20 for the TF network and >=4 for the miRNA network) and low indegree nodes as miRNA anti-targets (indegree=0) and the same fraction of nodes with lowest indegree in the TF network (80%; <=7 edges). For each GO/ImaGO category, we assessed over-representation and depletion with a hypergeometric P-value.

Mutual Enrichment between high indegree transcriptional and miRNA targets

We considered all genes that were either a target or a regulator in the TF and microRNA networks resulting in a total of 8760 nodes and defined high and low indegree sets as above. We then evaluated if nodes in the miRNA network with high indegree were enriched high indegree nodes of the transcriptional network (or vice versa) using a hypergeometric P-value.

Tissue co-expression

For each TF with available expression information (n=42; ImaGO (Tomancak et al. 2002)), we counted the number of targets that was co-expressed with the TF any of the annotated tissues and the number of targets that was not co-expressed. The statistical significance of co-expression of a TF with its target was estimated using the hypergeometric distribution with p being the probability of a gene being present in one of the tissues in which the TF is known to be expressed, x the total number of co-expressed targets and n the total number of targets of the TF with known tissue expression.

Network figure

The network figure was drawn in Cytoscape (Shannon et al. 2003) to display genes (nodes) and regulatory connections (edges) of the 60% confidence network. We colored edges and nodes if genes were expressed in the same tissue according to ImaGO (Tomancak et al. 2002). For clarity, we only show 20 randomly picked targets per transcription factor, i.e. without influencing the fraction of colored edges.

Availability

All data and predicted transcription factor and miRNA targets are freely available at http://compbio.mit.edu/fly/motif-instances/.

Acknowledgements

We thank Matt Rasmussen, Mike Lin (CSAIL, Broad), and other members of the Kellis lab for helpful discussions and for sharing unpublished data. AS thanks the Human Frontier Science Program Organization (HFSPO) for a postdoctoral fellowship (LT00495/2006-L). PK was supported in part by a National Science Foundation Graduate Research Fellowship. SR thanks Terran Lane and Maggie Werner-Washburne (University of New Mexico) for their support.

Figure Legends

Figure 1: The BLS measure (Brach Length Score) for assessing motif-conservation in many genomes. A. Conservation level and corresponding BLS scores for two *Mef-2* motif instances. The BLS measure scores the total branch length of the subtree connecting the species with motif instances, as a fraction of the total branch length of all twelve species. As shown in these examples (*Mef-2* motif: YTAWWWWTAR), BLS accounts for local alignment inaccuracies, gaps, motif movement, and motif loss. Species abbreviations used for: *Drosophila melanogaster, simulans*, *sechelia*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *pseudoobscura*, *willistonii*, *mojavensis*, *virilis*, and *grimshawii*. **B. BLS scores for different instance conservation scenarios.** Given the pattern of presence (black) and absence (white) within a phylogenetic tree, BLS evaluates the total branch length of the sub-tree connecting the species that contain the motif: when all species are present, BLS is 100% (column A); different sets of species lead to different BLS scores based on their evolutionary distances – distantly-related species lead to higher scores as they span larger evolutionary distances (columns B,C); species that are very closely related to each other lead to only small incremental contributions, due to their phylogenetic redundancy (columns D, E); sequencing, assembly, and alignment artifacts are not penalized, such as those stemming from lower-coverage genomes, as redundancy of branches between close species complements BLS (column F). Information about sequence coverage is from (Drosophila 12 Genomes 2007) and (Richards et al. 2005).

Figure 2: High-confidence recovery of individual motif instances. A. Mapping BLS scores to confidence values. Recovery of conserved motif instances the transcriptional repressor *Snail* (CAGGT) in promoter regions (2 kb regions upstream of transcription start sites), and the K-box miRNA (CTGTGAT) in 3'UTRs, at different BLS cutoffs (x-axis). Instances of shuffled control motifs (grey area) decrease much more rapidly than instances of real motifs (height of black curve), leading to a large fraction of motif instances conserved above background (black area). The motif confidence score (red line) is calculated as the fraction of conserved instances above background. Random motifs are selected to have equal frequency as real motifs at BLS=0. **B. Increasing confidence values select functional motif**

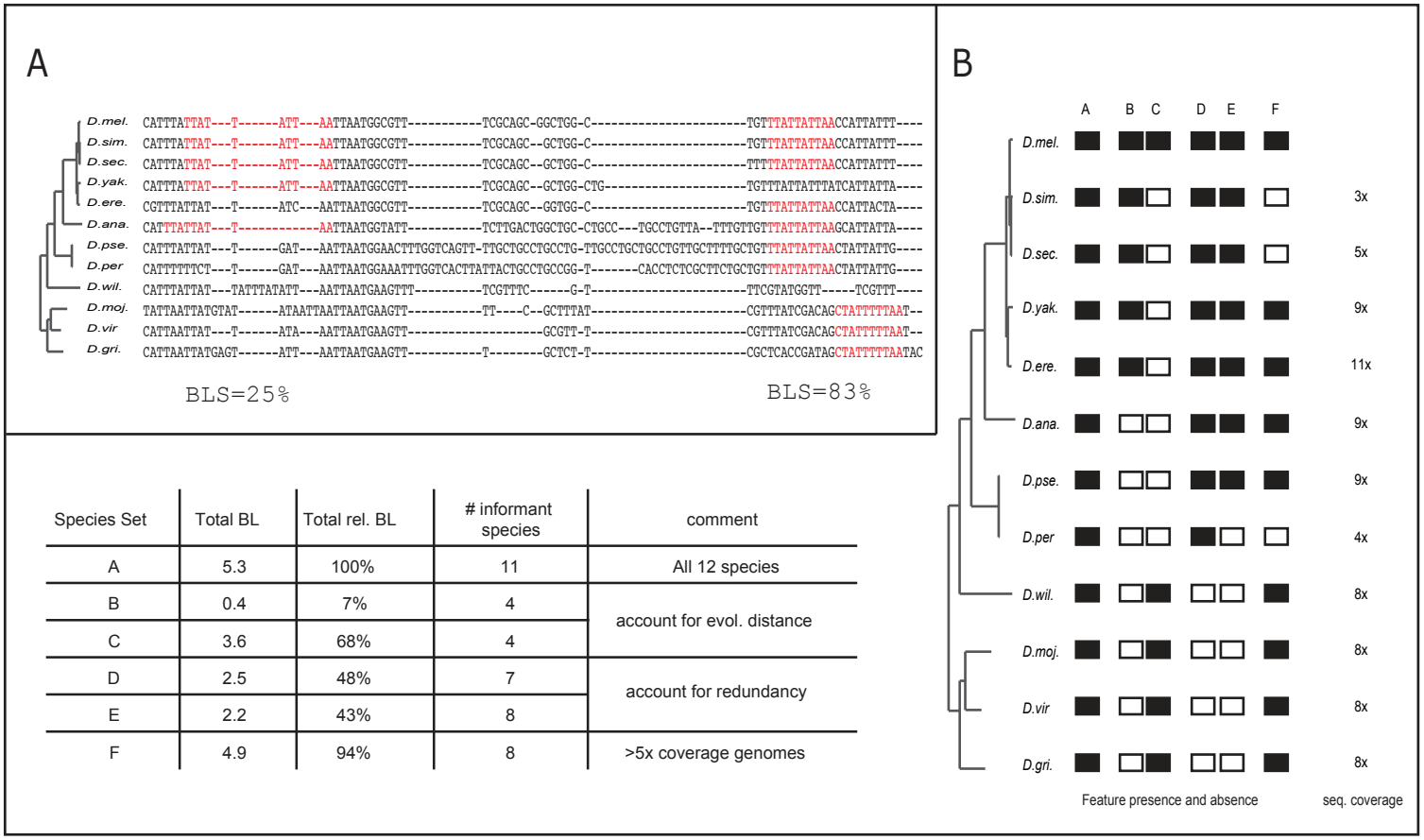
instances. B. With increasing confidence cutoffs (x-axis), transcription factor (TF) motif instances fall increasingly in promoter regions (light blue), 5'UTRs (red), and introns (green), at the exclusion of 3'UTRs (dark blue) and coding regions (yellow). In contrast, miRNA motif instances fall increasingly into 3'UTRs to the exclusion of promoters and other regions. Relative size of regions is normalized at BLS=0. **C.** miRNA motif instances at increasing confidence cutoffs are increasingly on the transcribed strand of 3'UTRs (black curve), while no such trend is seen for TF motifs (grey). Curves are truncated when < 10 instances reach the respective confidence.

Figure 3: Discovery power for motif instance prediction. A. Effect of tolerated motif movement. Number of recovered motif instances at 60% confidence for TF and miRNA motifs. Left panel: For both TF motifs (grey: *bicoid* motif, VVVBTAATCC), and miRNA motifs (black: miR-iab-4 motif, GTATACG), instance recovery increases until an optimal window size (500 and 400 nucleotides, respectively), and then decreases for larger movements, suggesting that tolerating motif movements increases overall discovery power. Right panel: Performance across all TF motifs (black) and all miRNA motifs (grey) shows improved recovery until windows of 300-500 nucleotides (for 60-80% of motifs), but reduced performance for larger window sizes. The performance for individual examples (left panel) shows a sharper peak than the overall performance across all motifs (right panel), as different window sizes are optimal for different motifs. **B. BLS measure leads to increased sensitivity.** Number of motif instances recovered (y-axis) at each confidence value (x-axis) for transcription factor (TF) motifs (left panel) and miRNA motifs (right panel). The BLS measure applied to the 12 fly genomes (blue) recovers more motif instances at each confidence, as compared to approaches requiring motif presence in all compared species ('full' conservation), applied to the 5 melanogaster species (red), the pairwise comparison of *D. melanogaster* and *D. pseudoobscura* (yellow), or the 9 *Sophophora* species (green). **C. Additional species lead to increased specificity.** Two measures of discovery power for the BLS measure applied to the 5 *melanogaster* group species (green), a pairwise comparison of *D. melanogaster* and *D. pseudoobscura* (grey), the 9 *sophophora* species (black), and all 12 *Drosophila* species (red). Left panel: More TF and miRNA motifs reach 60% confidence for increasing number of genomes at larger evolutionary distances. Right panel: Increasing numbers of genomes at larger evolutionary distances also lead to increased signal-to-noise ratio, measured as the conservation level of real motifs vs. control motifs at the most stringent BLS cutoff.

Figure 4: Conserved motif instances identify functional *in vivo* targets. Functional *in vivo* targets were determined for *Mef-2*, *Twist* and *Snail* using ChIP-chip (Sandmann et al. 2007; Sandmann et al. 2006; Zeitlinger et al. 2007), and direct transcriptional targets were determined for *CrebA* using various assays

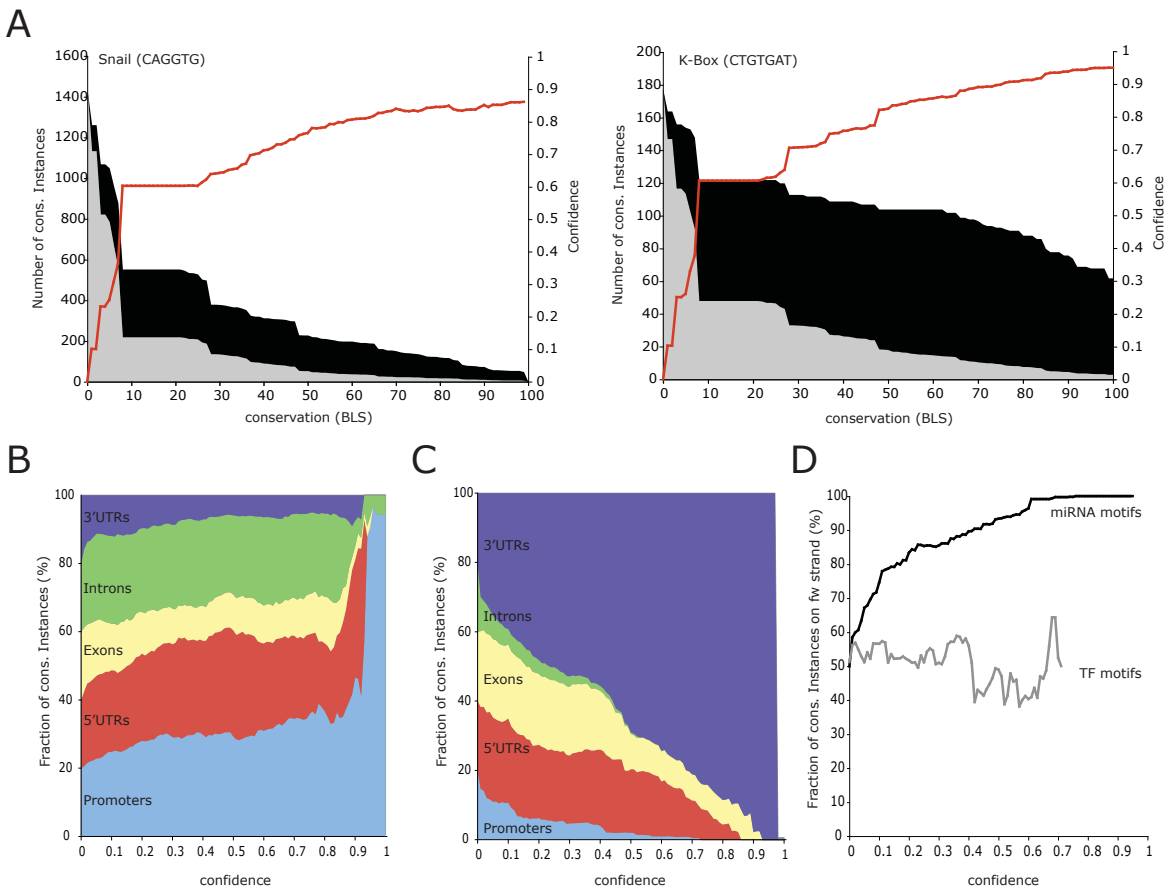
(Abrams and Andrew 2005). **A. Increasing confidence values show increased enrichment for *in vivo* sites.** Fold enrichment in functional *in vivo* sites (y-axis) for conserved motif instances at varying confidence values (x-axis). Hypergeometric P-values for max fold enrichments are $4x10^{-11}$ for *Mef-2*, $2x10^{-6}$ for *Twist*, $3x10^{-10}$ for *Snail*, and $1x10^{-7}$ for *CrebA*. Increasing confidence levels selected functional *in vivo* sites with increased enrichment for all four regulators, showing that high conservation selects for functional motif instances (X=0% shows the enrichment in the absence of comparative information, i.e. without requiring conservation). Curves are truncated when motifs don't reach the respective confidence levels. **B,C,D. High-sensitivity recovery of *in vivo* targets for TF and miRNA regulators.** Fraction of motifs in bound regions recovered at 60% confidence (black bars), compared to the fraction expected given the overall conservation of the respective regions, as assessed by control motifs using the same BLS cutoff (grey; suggesting preferential conservation of the corresponding TF motif instances). **B**. Recovery of ChIP-bound motifs, across all ChIP-bound regions (C), and only those instances overlapping known enhancers (E). Recovery rates show high sensitivity for TF motif instances, especially when these overlap enhancer regions. **C.** Recovery of experimentally validated direct *CrebA* targets shows even higher sensitivity, likely due to the multiple lines of experimental evidence establishing them as direct targets. **D.** miRNA recovery at 80% confidence is very high. **E. Non-conserved ChIP sites show reduced functional enrichments.** Enrichment in promoter regions of muscle genes for motif instances of activators *Twist* and *Mef-2*, and depletion for motif instances of repressor *Snail* are reduced for ChIP-bound regions for which motif instances are not conserved, suggesting they may contain a higher fraction of non-functional sites. The enrichment/depletion is even weaker for ChIP-bound regions without motif instances (all enrichments are significant with P-values between $1.1x10^{-4}$ and $5.1x10^{-13}$ except those for Snail). **F. Conservation-inferred targets and ChIP-inferred targets show comparable functional enrichments.** Conservation-inferred motif targets at 60% confidence (red; all $P<10^{-4}$) show higher muscle-gene enrichment/depletion than ChIP-inferred targets (black). Even outside ChIP-bound regions, conserved motifs show comparable enrichment and depletion (blue; all $P<5x10^{-3}$).

Figure 5: An initial regulatory network in *Drosophila*. Regulatory network with 46,525 connections between 83 TF and 57 miRNA motifs (circles) and their target genes (squares) at 60% confidence. If the regulator and its target are co-expressed in at least one tissue according to ImaGO (Tomancak et al. 2002), the corresponding edges and nodes are colored red, otherwise they are left grey. The high fraction of red colored edges (46%, $P=2x10^{-3}$) highlights the quality of the network. Nodes with gene names and bold edges indicate examples of regulatory connections with evidence in the literature (see **Table S4**).



Species Set	Total BL	Total rel. BL	# informant species	comment
A	5.3	100%	11	All 12 species
B	0.4	7%	4	account for evol. distance
C	3.6	68%	4	
D	2.5	48%	7	account for redundancy
E	2.2	43%	8	
F	4.9	94%	8	>5x coverage genomes

Figure 1


A
Snail (CAGGTG)
Number of cons. Instances
Confidence
conservation (BLS)
K-Box (CTGTGAT)
Number of cons. Instances
Confidence
conservation (BLS)

B
3'UTRs
Introns
Exons
5'UTRs
Promoters
Fraction of cons. Instances (%)
confidence
C
3'UTRs
Introns
Exons
5'UTRs
Promoters
Fraction of cons. Instances (%)
confidence
D
miRNA motifs
TF motifs
Fraction of cons. Instances on fw strand (%)
confidence

Figure 2

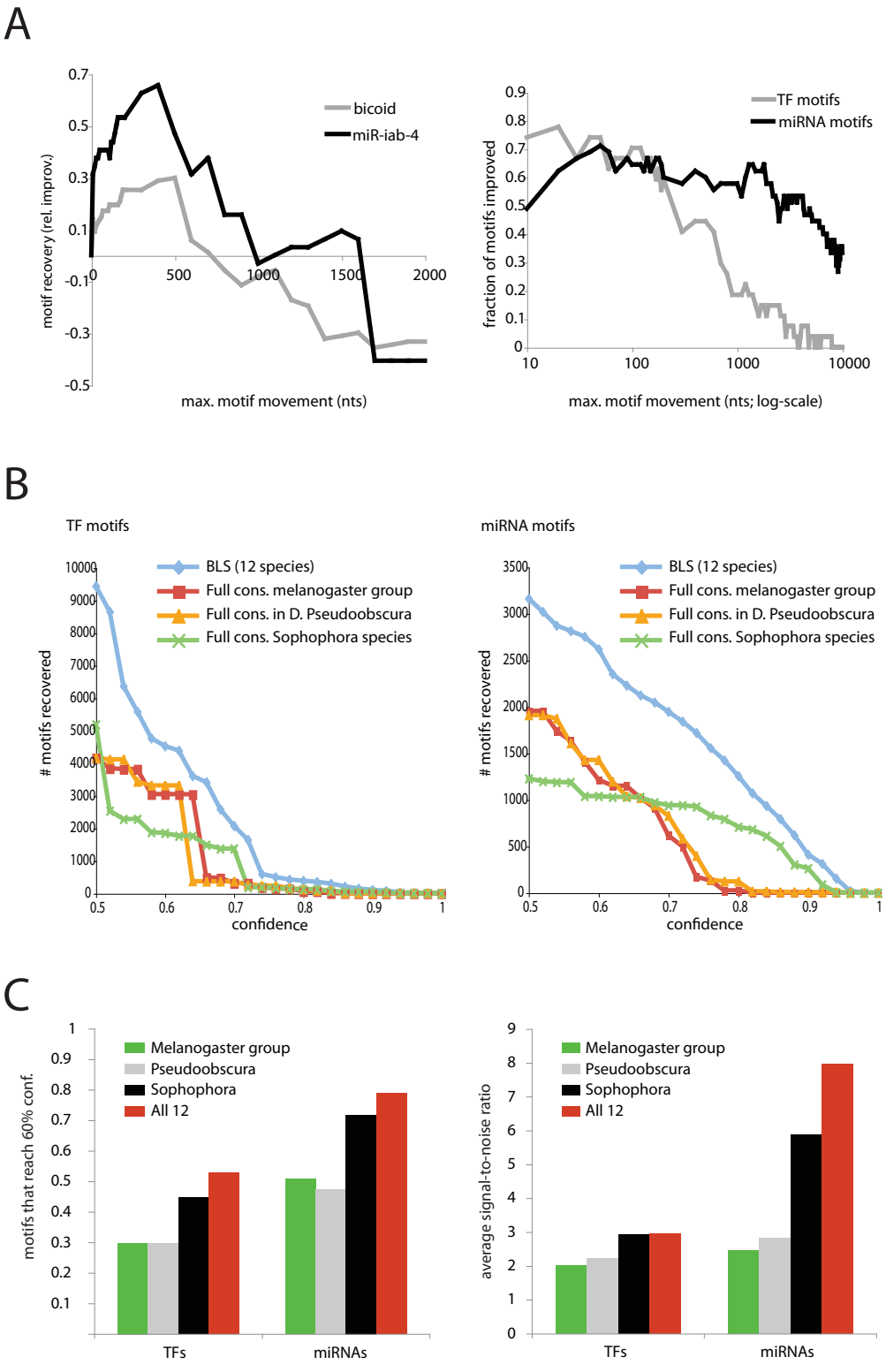

A
bicoid
miR-iab-4
motif recovery (rel. improv.)
max. motif movement (nts)
TF motifs
miRNA motifs
fraction of motifs improved
max. motif movement (nts; log-scale)
B
TF motifs
miRNA motifs
BLS (12 species)
Full cons. melanogaster group
Full cons. in D. Pseudoobscura
Full cons. Sophophora species
motifs recovered
confidence
C
Melanogaster group
Pseudoobscura
Sophophora
All 12
motifs that reach 60% conf.
average signal-to-noise ratio
TFs
miRNAs

Figure 3

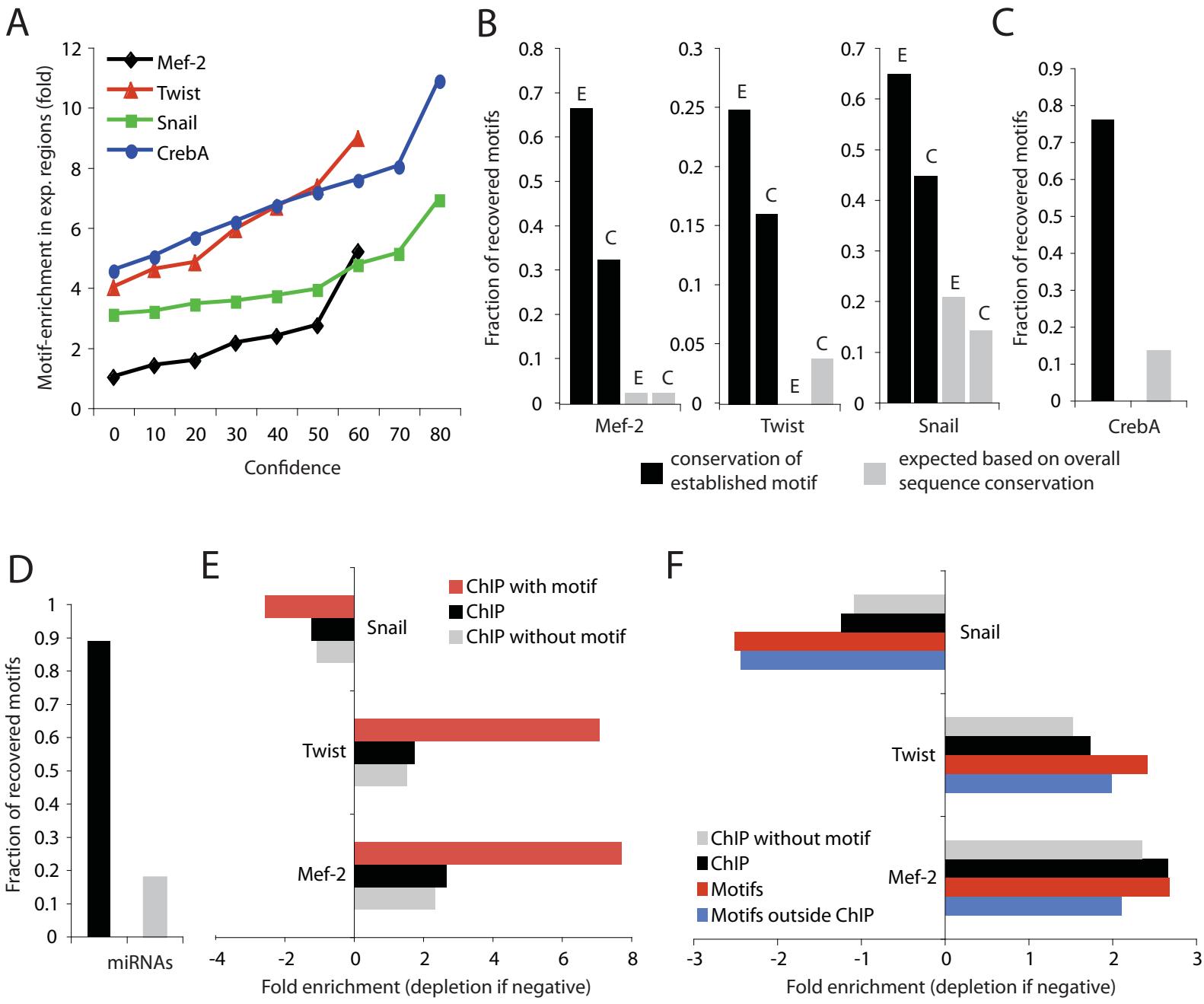

A
Motif-enrichment in exp. regions (fold)
Mef-2
Twist
Snail
CrebA
Confidence
B
Fraction of recovered motifs
E
C
Mef-2
Twist
Snail
CrebA
conservation of established motif
expected based on overall sequence conservation
D
Fraction of recovered motifs
miRNAs
E
ChIP with motif
ChIP
ChIP without motif
Snail
Twist
Mef-2
Fold enrichment (depletion if negative)
F
Snail
Twist
Mef-2
ChIP without motif
ChIP
Motifs
Motifs outside ChIP
Fold enrichment (depletion if negative)

Figure 4

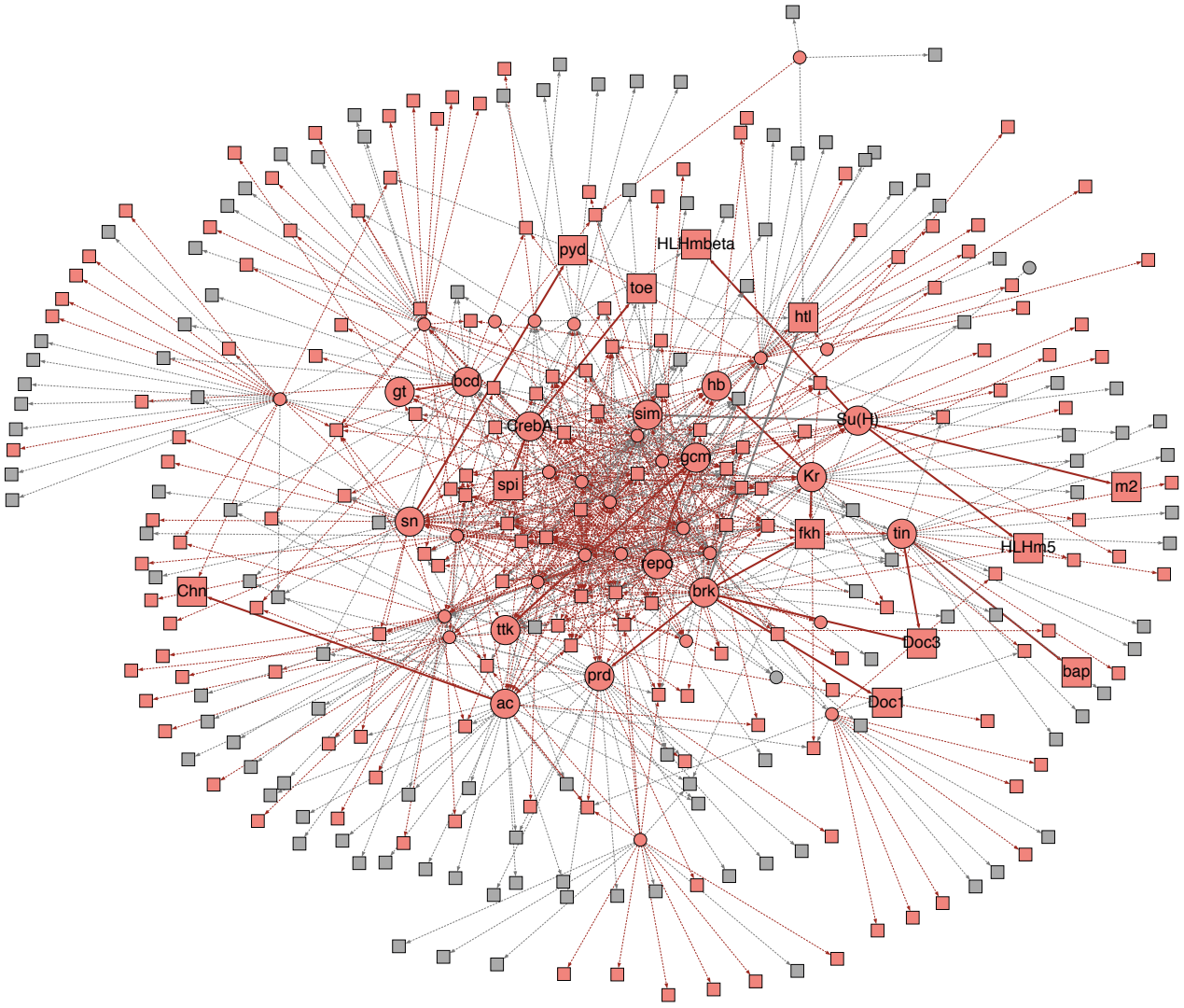

pyd
HLHmbeta
toe
htl
gt
bcd
hb
sim
GrebA
Su(H)
gcm
Kr
m2
spi
sn
fkh
tin
HLHm5
epo
Chn
brk
ttk
Doc3
bap
prd
ac
Doc1

Figure 5

References

Abrams, E.W. and D.J. Andrew. 2005. CrebA regulates secretory activity in the Drosophila salivary gland and epidermis. Development 132: 2743-2758.

Adryan, B. and S.A. Teichmann. 2006. FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. Bioinformatics 22: 1532-1533.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

Bailey, A.M. and J.W. Posakony. 1995. Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. Genes Dev 9: 2609-2622.

Bergman, C.M., J.W. Carlson, and S.E. Celniker. 2005. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747-1749.

Berman, B.P., Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, and M.B. Eisen. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99: 757-762.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K.D. Lewis, I. Ovcharenko, L. Pachter, and E.M. Rubin. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391-1394.

Bollobás, B. 2001. Random graphs. Cambridge University Press, Cambridge ; New York.

Boyer, L.A., T.I. Lee, M.F. Cole, S.E. Johnstone, S.S. Levine, J.P. Zucker, M.G. Guenther, R.M. Kumar, H.L. Murray, R.G. Jenner et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.

Brennecke, J., A. Stark, R.B. Russell, and S.M. Cohen. 2005. Principles of MicroRNA-Target Recognition. PLoS Biol 3: e85.

Chan, C.S., O. Elemento, and S. Tavazoie. 2005. Revealing Posttranscriptional Regulatory Elements Through Network-Level Conservation. PLoS Comput Biol 1: e69.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen, and M. Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71-76.

Cooper, G.M., E.A. Stone, G. Asimenos, E.D. Green, S. Batzoglou, and A. Sidow. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-913.

Davidson, E.H. and D.H. Erwin. 2006. Gene regulatory networks and the evolution of animal body plans. Science 311: 796-800.

Dewey, C.N., P.M. Huggins, K. Woods, B. Sturmfels, and L. Pachter. 2006. Parametric alignment of Drosophila genomes. PLoS Comput Biol 2: e73.

Drosophila 12 Genomes, C. 2007. Evolution of Genes and Genomes on the Drosophila Phylogeny. Nature in press.

Eddy, S.R. 2005. A model of the statistical power of comparative genome sequence analysis. PLoS Biol 3: e10.

Elnitski, L., R.C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M.J. O'Connor, S. Schwartz, W. Miller, and F. Chiaromonte. 2003. Distinguishing regulatory DNA from neutral sites. Genome Res 13: 64-72.

Ettwiller, L., B. Paten, M. Souren, F. Loosli, J. Wittbrodt, and E. Birney. 2005. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. Genome Biol 6: R104.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Griffiths-Jones, S., R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140-144.

Grun, D., Y.L. Wang, D. Langenberger, K.C. Gunsalus, and N. Rajewsky. 2005. microRNA Target Predictions across Seven Drosophila Species and Comparison to Mammalian Targets. PLoS Comput Biol 1: e13.

Gupta, S., J.A. Stamatoyannopoulos, T.L. Bailey, and W.S. Noble. 2007. Quantifying similarity between motifs. Genome Biol 8: R24.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241-254.

Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. Genome Res 12: 996-1006.

Kraut, R. and M. Levine. 1991. Spatial regulation of the gap gene giant during Drosophila development. Development 111: 601-609.

Lai, E.C. 2004. Predicting and validating microRNA targets. Genome Biol 5: 115.

Lai, E.C., C. Burks, and J.W. Posakony. 1998. The K box, a conserved 3' UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts. Development 125: 4077-4088.

Lee, T.I., R.G. Jenner, L.A. Boyer, M.G. Guenther, S.S. Levine, R.M. Kumar, B. Chevalier, S.E. Johnstone, M.F. Cole, K. Isono et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-313.

Lewis, B.P., I.H. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge. 2003. Prediction of mammalian microRNA targets. Cell 115: 787-798.

Ludwig, M.Z., C. Bergman, N.H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564-567.

Ludwig, M.Z., A. Palsson, E. Alekseeva, C.M. Bergman, J. Nathan, and M. Kreitman. 2005. Functional evolution of a cis-regulatory module. PLoS Biol 3: e93.

Margulies, E.H., M. Blanchette, D. Haussler, and E.D. Green. 2003. Identification and characterization of multi-species conserved sequences. Genome Res 13: 2507-2518.

Margulies, E.H., G.M. Cooper, G. Asimenos, D.J. Thomas, C.N. Dewey, A. Siepel, E. Birney, D. Keefe, A.S. Schwartz, M. Hou et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 17: 760-774.

Markstein, M., R. Zinzen, P. Markstein, K.P. Yee, A. Erives, A. Stathopoulos, and M. Levine. 2004. A regulatory code for neurogenic gene expression in the Drosophila embryo. Development 131: 2387-2394.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.

McGregor, A.P., V. Orgogozo, I. Delon, J. Zanet, D.G. Srinivasan, F. Payre, and D.L. Stern. 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448: 587-590.

Miller, W., K.D. Makova, A. Nekrutenko, and R.C. Hardison. 2004. Comparative genomics. Annu Rev Genomics Hum Genet 5: 15-56.

Moses, A.M., D.Y. Chiang, D.A. Pollard, V.N. Iyer, and M.B. Eisen. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5: R98.

Mukherjee, S., M.F. Berger, G. Jona, X.S. Wang, D. Muzzey, M. Snyder, R.A. Young, and M.L. Bulyk. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet 36: 1331-1339.

Naughton, B.T., E. Fratkin, S. Batzoglou, and D.L. Brutlag. 2006. A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. Nucleic Acids Res 34: 5730-5739.

Philippakis, A.A., B.W. Busser, S.S. Gisselbrecht, F.S. He, B. Estrada, A.M. Michelson, and M.L. Bulyk. 2006. Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. PLoS Comput Biol 2: e53.

Pietrokovski, S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 24: 3836-3845.

Prud'homme, B., N. Gompel, A. Rokas, V.A. Kassner, T.M. Williams, S.D. Yeh, J.R. True, and S.B. Carroll. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440: 1050-1053.

Rajewsky, N. 2006. microRNA target predictions in animals. Nat Genet 38 Suppl 1: S8-S13.

Richards, S., Y. Liu, B.R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M.J. Hubisz, R. Chen, R.P. Meisel et al. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res 15: 1-18.

Sandelin, A., W. Alkema, P. Engstrom, W.W. Wasserman, and B. Lenhard. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91-94.

Sandmann, T., C. Girardot, M. Brehme, W. Tongprasit, V. Stolc, and E.E. Furlong. 2007. A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev 21: 436-449.

Sandmann, T., L.J. Jensen, J.S. Jakobsen, M.M. Karzynski, M.P. Eichenlaub, P. Bork, and E.E. Furlong. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. Dev Cell 10: 797-807.

Schones, D.E., P. Sumazin, and M.Q. Zhang. 2005. Similarity of position frequency matrices for transcription factor binding sites. Bioinformatics 21: 307-313.

Schroeder, M.D., M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E.D. Siggia, and U. Gaul. 2004. Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol 2: E271.

Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. Rna 12: 192-197.

Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

Stark, A., J. Brennecke, N. Bushati, R.B. Russell, and S.M. Cohen. 2005. Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution. Cell 123: 1133-1146.

Stark, A., J. Brennecke, R.B. Russell, and S.M. Cohen. 2003. Identification of Drosophila MicroRNA Targets. PLoS Biol 1: E60.

Stark, A., M.F. Lin, P. Kheradpour, J.S. Pedersen, L. Parts, J.W. Carlson, M.A. Crosby, M.D. Rasmussen, S. Roy, A.N. Deoras et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature in press.

Stormo, G.D. 2000. DNA binding sites: representation and discovery. Bioinformatics 16: 16-23.

Taylor, J., S. Tyekucheva, D.C. King, R.C. Hardison, W. Miller, and F. Chiaromonte. 2006. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res 16: 1596-1604.

Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, J.C. McDowell et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788-793.

Tomancak, P., A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S.E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S.E. Celniker et al. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3: RESEARCH0088.

Tompa, M., N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.

Tuerk, C. and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249: 505-510.

Ureta-Vidal, A., L. Ettwiller, and E. Birney. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4: 251-262.

Van Doren, M., A.M. Bailey, J. Esnayra, K. Ede, and J.W. Posakony. 1994. Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. Genes Dev 8: 2729-2742.

Wasserman, W.W., M. Palumbo, W. Thompson, J.W. Fickett, and C.E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26: 225-228.

Wasserman, W.W. and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-287.

Wray, G.A., M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377-1419.

Xie, X., J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434: 338-345.

Yada, T., Y. Totoki, M. Ishikawa, K. Asai, and K. Nakai. 1998. Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. Bioinformatics 14: 317-325.

Yin, Z. and M. Frasch. 1998. Regulation and function of tinman during dorsal mesoderm induction and heart specification in Drosophila. Dev Genet 22: 187-200.

Zeitlinger, J., R.P. Zinzen, A. Stark, M. Kellis, H. Zhang, R.A. Young, and M. Levine. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev 21: 385-390.

Zhou, Q. and W.H. Wong. 2004. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A 101: 12114-12119.