Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances

KIM, Jaebum, et al.

Abstract

We address the problem of finding statistically significant associations between cis-regulatory motifs and functional gene sets, in order to understand the biological roles of transcription factors. We develop a computational framework for this task, whose features include a new statistical score for motif scanning, the use of different scores for predicting targets of different motifs, and new ways to deal with redundancies among significant motif-function associations. This framework is applied to the recently sequenced genome of the jewel wasp, Nasonia vitripennis, making use of the existing knowledge of motifs and gene annotations in another insect genome, that of the fruitfly. The framework uses cross-species comparison to improve the specificity of its predictions, and does so without relying upon non-coding sequence alignment. It is therefore well suited for comparative genomics across large evolutionary divergences, where existing alignment-based methods are not applicable. We also apply the framework to find motifs associated with socially regulated gene sets in the honeybee, Apis mellifera, using comparisons [...]
Functional Characterization of Transcription Factor Motifs Using Cross-species Comparison across Large Evolutionary Distances

Jaebum Kim, Ryan Cunningham, Brian James, Stefan Wyder, Joshua D. Gibson, Oliver Niehuis, Evgeny M. Zdobnov, Hugh M. Robertson, Gene E. Robinson, John H. Werren, Saurabh Sinha

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America,
Department of Genetics Medicine and Development, University of Geneva Medical School, and Swiss Institute of Bioinformatics, Geneva, Switzerland,
School of Life Sciences, Arizona State University, Tempe, Arizona, United States of America,
Department of Biology, University of Osnabrück, Osnabrück, Germany,
Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America,
Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America,
Department of Biology, University of Rochester, Rochester, New York, United States of America

Abstract

We address the problem of finding statistically significant associations between cis-regulatory motifs and functional gene sets, in order to understand the biological roles of transcription factors. We develop a computational framework for this task, whose features include a new statistical score for motif scanning, the use of different scores for predicting targets of different motifs, and new ways to deal with redundancies among significant motif–function associations. This framework is applied to the recently sequenced genome of the jewel wasp, Nasonia vitripennis, making use of the existing knowledge of motifs and gene annotations in another insect genome, that of the fruitfly. The framework uses cross-species comparison to improve the specificity of its predictions, and does so without relying upon non-coding sequence alignment. It is therefore well suited for comparative genomics across large evolutionary divergences, where existing alignment-based methods are not applicable. We also apply the framework to find motifs associated with socially regulated gene sets in the honeybee, Apis mellifera, using comparisons with Nasonia, a solitary species, to identify honeybee-specific associations.

Introduction

Computational discovery and analysis of gene regulatory networks begins with the characterization of transcription factor (TF) motifs, through experimental or computational means. The next task of characterizing the biological functions regulated by these motifs is crucial for gaining broad, systems-level insights about the regulatory network, and has been the subject of several studies in recent years [1–3]. We present a general framework for discovering such motif – function associations through genome sequence analysis, while using evolutionary conservation as a guide. Evolutionary comparisons in this framework are carried out without relying upon alignment of non-coding sequences, making the framework especially well suited for species that are greatly diverged from their nearest sequenced relatives.

Starting with a list of TF motifs, a researcher is often faced with the task of annotating putative binding sites matching those motifs, the so-called “motif scanning” [4] task. The predicted binding sites may then be used to annotate a set of genes (typically genes that are proximal to the sites) as being putative regulatory targets of the motif. Such a set of (predicted) target genes of a TF is called its “motif module” [5]. A motif module is thus a part of the gene regulatory network, representing the direct regulatory targets of a TF. Prediction of motif modules has been the focus of numerous studies in the past [6–8]. In a later section (“Motif scanning methods”), we briefly review existing approaches to this problem, most of which are based on finding sites whose quality of match to the motif exceeds a threshold, or locations where clusters of above-threshold matches are found. Each of these approaches has its merits and problems, and it is not clear which method ought to be used in practice. We examine this issue systematically, while proposing a new statistical score for motif scanning, and find different methods to be most efficacious for predicting the motif module for different TFs.

A motif module may be tested for statistical enrichment for any given gene set, such as genes in a Gene Ontology (GO) functional category [9], a metabolic or signal transduction pathway [10], or genes coordinately expressed in a particular condition [11–12]. Such statistical enrichment can shed light on possible biological roles of the motif. A compendium of statistical associations between motifs and functions is called a “motif function map” [2]. This map is a potential starting point for researchers exploring the cis-regulatory basis of a particular biological process [5]. It may be constructed by straight-forward statistical procedures for signifi-
Author Summary

We develop a computational pipeline for predicting the functions of transcription factor motifs, through DNA sequence analysis. The pipeline is applied to the newly sequenced genome of the jewel wasp, *Nasonia vitripennis*. It exploits the wealth of molecular data available in another insect species, the fruitfly *Drosophila melanogaster*, and uses cross-species comparison to its advantage. Our main contribution is to show how this can be done despite the large evolutionary divergence between the two species. The methodology presented here may be applied more generally to other scenarios (genomes) where comparative regulatory genomics must deal with large evolutionary divergences.

Nasonia vitripennis (Insecta: Hymenoptera), the first of a parasitoid species to be sequenced [21]. Even though the evolutionary divergence of *Nasonia* from its closest sequenced relative, the honeybee *Apis mellifera* (~180 Myrs, [21]), and from the fruitfly, *Drosophila melanogaster* (~300 Myrs, [21]), precludes alignment-based comparison of non-coding sequences, we are able to exploit these two genomes as well as the vast knowledge base in *Drosophila* to make reliable predictions in *Nasonia*.

Results

Overview

We begin with an outline of the major contributions of this work, pointing out the specific challenges that needed to be addressed.

Computational pipeline development and evaluation.

There are two major components here.

1) Motif function prediction in single species: First, we examine the motif scanning problem, i.e., predicting regulatory targets of a TF, given its binding specificity (motif). We propose a new statistical score, based on hidden Markov models, for this problem. We implement this score, as well as two alternative scores that capture the gist of existing statistical approaches to the problem [7,22–24]. Second, the highest scoring target genes of a motif are tested for association with specific functions, i.e., Gene Ontology (GO) [9] categories. Since this step in its conventional form tends to report motif associations with numerous mutually redundant GO categories [2], we propose a new statistical approach, based on an extension of the Hypergeometric test, to trim the list of significant associations to a non-redundant list. Finally, we apply the above two steps on a well studied genome (e.g., *Drosophila melanogaster*) where motifs have been discovered and genes have been annotated with GO categories, to choose one of the three motif scanning scores as the most appropriate one for each motif. The selected motif scanning score, along with tools for the first two steps above, can now be used to predict motif – function associations in any genome, for any given motif.

2) Enabling comparative genomics across large divergences:

a) To apply our pipeline to a relatively less studied species such as *Nasonia vitripennis* (henceforth called the “target” species or genome), we first need to specify the input set of TF motifs. We obtain this from the nearest genome (*D. melanogaster* in this case) where such a collection exists [13,25–26]. (Other taxonomical groups with relatively large collections of experimentally characterized motifs include yeast [14], mouse [13,27] and human [13,28].)

b) We consider the possibility that a motif characterized in one species (*D. melanogaster*) may not be usable in a greatly diverged target species (*N. vitripennis*) due to a significant change in the binding specificity of the TF. We address this potential problem by using an automated pipeline to align DNA-binding domains of orthologous TFs in the two species and exploiting structural information to determine if DNA-contacting residues have changed, thereby obtaining information on the evolutionary conservation of the corresponding motifs.

c) The steps outlined thus far (steps 1, 2a, 2b above) are sufficient to discover motif – function associations in the target genome. However, with the goal of boosting the specificity of such predictions, we apply the pipeline...
separately to the target genome as well as other genomes where the associations are expected to be conserved by and large. We then identify associations that are statistically significant in every species, thus using evolutionary conservation as a "filter". Unlike previous studies [3,15,29] that used the conservation filter to improve binding site prediction (by requiring that sequence alignments reveal the site to be conserved), we use evolutionary conservation at a higher level that does not rely upon non-coding sequence alignment.

Furthermore, we systematically assess the effect of using cross-species comparison on the accuracy of motif function characterization. For this purpose, we design benchmarks comprising highly reliable motif – GO term associations, based on the wealth of chromatin immunoprecipitation (ChiP)-based and genetics-based data on TF – DNA binding in Drosophila. We then show that using the new approach (step 2c above) consistently achieves significantly greater precision than the single-species version of the same pipeline.

Applications of pipeline: We first compile a compendium of highly significant motif associations with function categories in GO, through direct application of the above pipeline to Nasonia. We then present alternative ways in which comparison of motif – function associations across species can be used to gain biological insights: (a) associations with social behavior-related gene sets in the honeybee are compared with the solitary taxa Nasonia and Drosophila, in search of a *cis*-regulatory code of sociality, and (b) motifs with known roles in regulation of oxidative phosphorylation in *Drosophila* are tested for associations with this pathway in *Nasonia*.

A computational pipeline for charting a "motif function map"

A "motif module" [5] is the set of genes computationally predicted as being targets of a given motif. A motif module can be tested for statistical enrichment for any given gene set, typically a Gene Ontology (GO) functional category, and the full complement of statistical associations between motifs and functions is called a "motif function map" [2]. This section describes our new computational pipeline for charting a motif function map. The description follows the outline presented above and is illustrated in Figure 1.

Motif scanning methods. The pipeline implements three different motif scanning scores, where a motif is represented as a position weight matrix (PWM) [30].

"site-LLR". The traditional approach to motif scanning is to find strong matches to the PWM using information theoretic measures of similarity and a high threshold on the similarity measure. The most popularly used binding site prediction programs applicable to single species data belong to this genre (e.g., Patser [http://ural.wustl.edu/software.html] and MATCH [23]). In some cases, a count of such strong matches in a small window (~500 bp) has been used [22]. We refer to this approach as the "site-LLR" method, and implement our own version as described in Methods.

"Stubb": In our previous work [2], we argued for the use of a new probabilistic score, obtained from the "Stubb" program [24] based on a hidden Markov model (HMM), that integrates all potential sites, weak and strong, in a small window (~500 bp), rather than relying only on strong sites defined by *ad hoc* thresholds. The Stubb program computes the likelihood of the sequence under a "two-state HMM" (Figure S1) parameterized by the given motif and then uses its ratio to the likelihood under a null ("background") model that does not include the motif. This approach is similar in spirit to some other available motif scanning methods, such as "Clover" [7], while substantially different from the site-LLR approach outlined above.

"SWAN". In order to address certain limitations of Stubb and other existing HMM-based scores (see Discussion), we defined the following new score for motif scanning:

- In the first step, the two-state HMM (Figure S1) is trained on the background sequences, which may be the entire genome, or some selected portion of it. This step learns (via likelihood maximization) a value for the motif transition probability, also called "motif weight", that captures the frequency of occurrence of the motif in background sequences. Note that "occurrence" here refers implicitly to stochastic transitions to the motif state, rather than to threshold-based matches.

- The second step computes a log likelihood ratio (LLR) score for the target sequence, where (1) the denominator is the likelihood of the target sequence under a new background model – a two-state HMM with motif weight fixed at the value learned in the previous step, and (2) the numerator, as in Stubb, is the likelihood under a two-state HMM with motif weight being a free parameter (constrained to be greater than the motif weight learned above).

We have implemented this new score that we call "SWAN" (Stubb With Another Null) (see Methods for more details, especially with respect to the "background state" in the HMM). Each of the above scores may be used to report the "motif module" for a given motif, as the genes with the highest scoring promoter regions in the genome (Figure 1A and Methods).

Motif – function associations. The next step is to search for statistically significant associations between motifs and GO function categories, based on the overlap between a motif module (reported by any of the three scores described above) and the genes in a GO category, using the Hypergeometric test (Figure 1A). This step reports the associations in ascending order of *p*-values, along with *q*-values [31] for multiple hypothesis correction. However, it is common for the set of significant associations to include multiple GO categories that are highly overlapping/redundant; e.g., a motif may show strong associations with "pattern formation", and "anterior posterior pattern formation", the latter being a strict subset of the former. To identify such redundancies in the list of associations, we also produce a reorganized list where, if an association is "statistically explained" by another association already reported (that has stronger *p*-value), the former is grouped with the latter to distinguish it from a truly distinct association. Figure 1C shows a snapshot of this reorganized format for reporting associations. We quantify the notion of one association "statistically explaining" another association by extending the Hypergeometric test to consider three subsets instead of the usual two, and imposing the observed overlap structure of these sets as a constraint that the computed *p*-value is conditional on (see Methods for details, and also see Discussion for related work, e.g., Grossmann et al. [32]). A similar reorganization is also applied to reports of all motif associations for any particular GO category; this is important since our motif compendium includes multiple motifs for the same TF, and also because in some cases different TFs have very similar binding specificities.

Selection of motif scanning method for each motif. The computational pipeline includes three motif scanning methods, "site-LLR", "Stubb", and "SWAN". One of these methods is to be used to predict the motif module required for detecting motif –
Figure 1. Computational pipeline for charting a motif function map. (A) Each motif is scored against each gene’s promoter (“motif scanning”). The top scoring target genes of a motif (“motif module”) are analyzed for enrichment for GO gene sets using the Hypergeometric test, and statistically significant motif – GO associations (red cells) from the test constitute a “motif function map”. (B) Different motif scanning methods produce different motif function maps by the process in (A). For each motif, the best motif scanning method (score) is selected by evaluating each motif function map based on the number of associations and a suitable control (see Methods). (C) For each motif, redundant GO associations are identified by using an extended Hypergeometric test (see Methods) and the motif function map is reorganized. This panel shows GO associations of the Fushi tarazu (FTZ) motif, with redundant associations being indented. The “cond-pval” column is the conditional p-value of an association given the stronger association it is redundant with (see Methods). For example, the association with “sensory perception of smell” is highly significant (p-value ~ 6E-4), but is “statistically explained” by the association with “odorant binding” (conditional p-value ~ 1); the Venn diagram on the right illustrates why this is the case.

doi:10.1371/journal.pcbi.1000652.g001
function associations. However, it is not clear a priori which method would be most suited to this goal, or whether any one of these methods would be the best choice for all motifs. The next component of the pipeline selects the best motif scanning score for each motif (Figure 1B), by evaluating the motif function map arising out of each score on a genome that is not the target genome and where GO annotations for genes are available (Drosophila in our case). The selected score will now be used in applications of the pipeline to the target genome.

The evaluation is based on the following simple premise: (1) the better motif scanning method should lead to more associations (at the same statistical level of confidence) between motif modules and GO categories, and (2) if we randomize (shuffle) each gene promoter, the recomputed motif modules should not have significant associations with GO categories. While it is clear that the second condition serves as a form of “negative control”, its precise motivation may not be obvious at first. A significant p-value of association between a motif module and a biological gene set is a potentially interesting finding, provided that the motif module consists of sequences specifically associated with the motif (TF). This requirement may not always be met, for example if an unusual nucleotide composition (G/C content) of the promoters of a gene set leads to several false binding site predictions and therefore to a false motif association. This phenomenon was widely observed in our previous analysis of the Apis mellifera genome [33]. The second condition defined above explicitly tests for such false associations that are artifacts of abnormal G/C content rather than reflecting enrichment for the motif pattern. Details of our evaluation scheme are described in Methods (see Figure 2A for two example evaluations).

Based on evaluations in the Drosophila genome, we found that different motif scanning programs perform best for different motifs (Figure 2B). Of the 224 motifs in our compendium, SWAN, Stubb, and site-LLR were the best method (by the “AUC” criterion, see Methods) on 102, 66 and 56 motifs respectively (see Figure S2 for comparisons by other measures). We next asked if certain motif characteristics (e.g., G/C content, length, information content) were correlated with amenability to specific methods (Table S1). The only such correlation observed was that Stubb tended to be especially suited to motifs with low G/C content (p-value <0.01). We also used this evaluation approach to choose important parameters for the methods (see Methods and Figure 2B).

Extending the pipeline to use information from other species

Motifs and GO annotations. In order to apply the computational pipeline to a target genome where motifs and GO annotations are not available, we propose obtaining such data from the nearest genome where they are available. The latter is called the reference genome. (Also see Discussion.) GO annotations are deduced based on a homology map between the
target and reference genomes. Motifs from the reference genome are used in the target genome “as is” (Figure 3). However, since the two genomes may be greatly diverged, the pipeline attempts to determine whether a motif from the reference genome is likely to represent the binding specificity of the orthologous TF in the target genome. We used the software of Morozov and Siggia [34] to compare the relevant DNA binding domain (amino acid sequence) in the target genome with its ortholog in the reference genome, employing a structural template (of protein bound to DNA) to identify DNA-contacting residues as the key residues for site recognition. (For domains of the zinc finger family ZF-C2H2, we focused instead on four key residues known to be involved in binding specificity [35].) We then assigned a “motif conservation score (MCS)” to the motif based on whether these key residues were conserved (either identical or changed to a chemically similar amino acid) or not (see Figure 3 and Methods). The reported motif function map indicates whether a motif is evolutionarily conserved in this sense, thereby increasing the reliability of that motif’s associations. Among 160 Drosophila motifs scored by us, 80% scored highly (MCS ≥3 on a scale of 1 to 4) for conservation in Nasonia. Sixty four (28% of all) motifs could not be evaluated by our pipeline due to reasons explained in Table S2.

**Evolutionarily conserved motif – function associations.** Comparative genomics has played a key role in curbing false positive errors in cis-regulatory analyses [15–16,36]. However, when the target genome’s non-coding part is not alignable with any available genome, most of the existing frameworks for sequence-level comparative genomics are rendered useless. Alignment-free approaches have been proposed to address this problem, in the context of ab initio motif discovery [37–38], as well as motif target prediction [20]. Here, we exploit the power of comparative genomics by looking for conservation at a higher level, i.e., by finding motif – function associations that are statistically significant in multiple species, even though the motif scanning step is performed independently in the different genomes. We propose applying the pipeline not only on the target genome, but also on one or more other genomes (separately), and reporting motif – function associations that are statistically significant across genomes, based on a “combined p-value” (see Methods) computed from the individual p-values in each genome. For instance, we report below the motif associations that are conserved in Nasonia and Drosophila (also see Figure 3).

To assess the advantage of this strategy, we constructed a benchmark of highly reliable motif – function associations that...
were based on chromatin immunoprecipitation (ChIP)-based TF occupancy data. We roughly followed the methodology of Boden and Bailey [39], where a “gold standard” of TF–GO associations was constructed for yeast and human. We started by compiling 13 data sets of ChIP-based binding data in Drosophila, corresponding to 10 distinct TFs. We used the respective author-defined TF target gene sets, and compiled the GO terms enriched in these target sets at three different levels of significance (E-value 0.05, 0.01, 0.001, see Methods). These TF–GO associations were treated as the benchmark of “true” motif–function associations that our pipeline would try to predict, either in its single species version, or by exploiting cross-species comparison. To examine the effect of the species with which comparisons are made, we included the genomes of Apis mellifera, Tribolium castaneum, Anopheles gambiae, and Drosophila virilis, in addition to Drosophila melanogaster, as separate evolutionary filters for the Nasonia motif function map.

The performance of these predictions is shown in Figure 4, as the precision of the top 5 and top 10 predictions per motif (“PrecAt5” and “PrecAt10”, respectively), the precision at a fixed significance threshold (p-value of 0.005) (“PrecAtPval.005”), and as the point where precision equals recall (“PrecEqRecall”). Here, “precision” is the fraction of predicted associations that are “true” and recall is the fraction of “true” associations that were predicted as being significant. We note that the performance (by all measures) improves substantially in going from Nasonia (single species) to pairwise comparison-based predictions, the only (minor) exception being the “PrecAtPval.005” measure for Nasonia – Tribolium comparisons (Figure 4A-C). The improvement is most pronounced for Nasonia – Drosophila comparisons (e.g., “PrecAt5” improves from 0.2 to 0.36), presumably due to the benchmark being from Drosophila. We also note that in these large divergence comparisons, the actual evolutionary distance from Nasonia (e.g., ~180 Myrs for Apis and ~300 Myrs for Anopheles) does not make a significant difference in performance, except for the “PrecAtPval.005” measure that is substantially more improved with Apis comparisons than with Tribolium or Anopheles comparisons. The effect of cross-species comparison on the Drosophila motif function map (Figure 4D–F) shows a slightly different trend. The precision consistently improves in going from Drosophila melanogaster (single species) to Drosophila melanogaster – Drosophila virilis comparison-based predictions, although the recall drops (“PrecEqRecall” remains at 0.30 in either case). However, comparison with largely diverged species such as Anopheles, Tribolium, Nasonia and Apis suffers both in precision and recall, again with the exception of the “PrecAtPval.005” measure which conveys a mixed message. Finally, we observe that single species predictions are substantially better in Drosophila than in Nasonia, which is expected since (a) the benchmark associations are derived from Drosophila and may not be biologically “true” in Nasonia, and (b) the pipeline’s application to Nasonia uses motif and GO data from Drosophila. The above trends, and particularly the improvement in precision through the use of cross-species comparison, were also confirmed with a second benchmark that we constructed based on bona fide TF binding sites from the REDfly database [25]. (See Methods and Figure S3.)

Applications of computational pipeline

Motif function map in Nasonia, Apis, and Drosophila, based on Gene Ontology. The pipeline was run on Drosophila with the score selection component (Figure 1B) activated, and then run on Nasonia and Apis with the scoring scheme selected for each motif. We used a collection of experimentally validated motifs in Drosophila obtained from various sources (see Methods). An online interface to the motif function map in each of the three species is available at http://europa.cs.uiuc.edu:8080/nasonia/. For each species, this provides a “motif-centric view”, i.e., all GO associations for each motif, and a “function-centric view”, i.e., all motif associations for each GO category.

Motif function associations common to Nasonia and Drosophila. We looked for motif–function associations that were statistically significant in both Nasonia and Drosophila based on combined p-values. Overall, 177 such associations were discovered at a q-value of less than 0.05, representing evolutionarily conserved and presumably more reliable associations (Table S3). (All such associations had uncorrected p-value <0.004.) 91 of these 177 associations were non-redundant, 119 (67%) were for motifs that were scored for evolutionary conservation and 99 (53%) of these were highly conserved (motif conservation score MCS ≥3 on a scale of 1 to 4), as reported in Table S3. (MCS of 3 or more implies that every critical residue in the DNA-binding domain is either exactly conserved or substituted by an amino acid with a similar biochemical characterization (see Methods).) The discovered associations included several regulatory interactions that have already been experimentally characterized, chiefly in Drosophila. For instance, the motif for Suppressor of Hairless (SU(H)) is associated with the GO category “Notch signaling pathway” in both species (combined p-value 5E-9, Drosophila p-value 4E-6, Nasonia p-value 9E-5); the role of SU(H) in regulation of this pathway is well known [40] and conserved even in vertebrates [41]. The motif for Abdominal B (ABD-B) (MCS = 4) is associated with “ectodermal development” (p-value 1E-5, supported by [42]) and “salivary gland development” (p-value 7E-5, supported by [43]). The GAGA motif was assigned to several different biological processes, e.g., “tracheal system development” (p-value 4E-6) and “mesoderm development” (p-value 9E-6), consistent with its previous characterization as potentially regulating a broad range of cellular processes [44]. Some of the other most significant motif–function associations that are supported by the literature include Hunchback (HB) with “nervous system development” (p-value 5E-7, supported by [45]), Zerknullt (ZEN) (MCS = 4) with “ectodermal development” (p-value 3E-5, [46–47]), Mitochondrial transcription factor A (MTTFA) with “apoptosis” (p-value 3E-5, [48]), Antennapedia (ANTP) (MCS = 4) with “antennal morphogenesis” (p-value 0.001, [49]) and with “central nervous system development” (p-value 6E-5, [50]), and Heat shock factor (HSF) (MCS = 2) with “response to heat” (p-value 4E-5, [51]), among others. (The low MCS of HSF is due to a single substitution (I→M) at a predicted backbone contact residue, see Table S4.) Figure 5 shows, for four examples of conserved motif association, the motif targets and non-targets in both species.

We also found conserved motif–function associations that have not been previously identified in any species. For example, Reversed polarity (REPO) (MCS = 4) is associated with “transmission of nerve impulse” (p-value 4E-5). REPO is a major player in glial differentiation [52] and may be involved in transmission of signals [53].

In some cases, conserved motif associations could not be unambiguously assigned to a TF, due to similar binding specificities of different TFs. The Sex combs reduced (SCR) motif (MCS = 4) was assigned to the functional category of “protein synthesis” in both species; however this association was statistically explained by associations for similar motifs Empty spiracles (EMS) (in Drosophila) and Buttonless (BTN) (in Nasonia, MCS = 3), and our discovery may be pointing to an enrichment for a homeobox motif T/C/A)TAAT[G/T]A/G in the promoters of proteolysis-related genes, rather than for SCR binding sites in particular.

The GG12361 motif (MCS = 4) was associated with “cyclic nucleotide metabolism” (p-value 4E-5). This motif targets 7 of the
26 genes in this GO category in *Drosophila* and 6 of the 19 genes in the orthologous set in *Nasonia*, but the target sets (of size 7 and 6 respectively) are mutually exclusive. This presents an interesting situation where motif function associations are conserved but our procedure does not find the corresponding motif gene associations to be evolutionarily conserved. In other words, the association is discovered only at a higher level of comparative genomics.

The biological process “ectoderm development” was found to be associated with 15 non-redundant motifs (combined p-value, 0.004, q-value, 0.02), indicating that this is in part a highly conserved transcriptional network. Many of these associations are for motifs of factors with known roles in this process (e.g., ABD-B [MCS = 4] [42], ZEN [MCS = 4] [46–47], Abdominal A [ABD-A] [MCS = 4] [54], GAGA factor [55], SCR [MCS = 4] [56], Odd skipped [ODD] [57], ANTP [MCS = 4] [58]), while others are to our knowledge novel associations not reported in the literature (Extradenticle [EXD] [MCS = 4], HB, CG7056 [MCS = 2], PDHP [MCS = 4], Bric a brac 1 [BAB1], Hairy [H] [MCS = 4], SU(H)). The predicted motif change for CG7056 was due to two substitutions at key residues, and was corroborated by a specialized tool (http://ural.wustl.edu/flyhd/) that predicts the specificity of homeodomains (Figure S4).

The motif for the TF Bicoid (BCD) was found to be associated with “posterior head segmentation” and “trunk segmentation”.
This is interesting because BCD is associated with segmentation in Drosophila [59], but there is no known ortholog of this factor’s gene in Nasonia. The apparent conservation of the BCD – segmentation association may be due to another motif (Orthodenticle (OTD)) that is very similar to the BCD motif and is believed to play an important role in the above biological functions [60]. We discuss later the confounding effect of multiple TFs with similar binding specificity, as in this example, and its implications for our analysis.

Searching for motif associations that are statistically significant in the target genome (Nasonia) and another genome (Drosophila) is not the only manner in which evolutionary comparisons can inform a motif function map. In the following two subsections, we illustrate alternative ways in which cross-species comparisons may lead to new biological insights. A third example analysis is presented in Supplementary Text S1.

Motifs associated with social behavior in honeybee: Nasonia as an evolutionary filter. First, we present an example analysis where evolutionarily conserved associations may be of lesser interest biologically than lineage-specific ones. The honeybee, Apis mellifera, is a model organism for studying social behavior, and prior work has identified gene sets whose expression in the brain responds to social cues during behavioral maturation (e.g., from nurse to forager bees [61]). Nasonia is a member of the Hymenoptera order, to which the honeybee also belongs, but is not a social animal. Therefore, a motif association that is specific to behavioral gene regulation in Apis should be absent in Nasonia, when considering Nasonia orthologs of the same gene sets. Likewise, a conserved motif association undermines the hypothesis of a social behavior-specific role, and is likely an artifact of a more basal (not socially-specific) biological process that these genes are part of in both species. Working with gene sets analyzed in [33], we identified significant motif associations in Apis, and noted also the p-values of association (of the same motifs) from orthologous sets in Nasonia and Drosophila. Genes up-regulated in the Apis brain in response to Manganese treatment showed 67 significant motif associations in Apis (Table S3). However, upon invoking an “evolutionary filter” that requires the p-values in the other two (asocial) species to be above a threshold, only 27 associations remained. Thus, for this gene set, cross-species comparison was able to filter out 40 (59%) of the Apis associations (31 due to the Nasonia filter).

We also found 14 motif associations for other social behavior-related gene sets from Apis, four of which (including the previously predicted role of the GAGA factor [33]) do not pass the evolutionary filter (Table 1). The remaining ten (9 distinct) associations are potentially involved in the social regulation of gene expression in honeybee brains. Particularly interesting are the bee – specific enrichment for Broad (BR), Adh transcription factor 1 (ADF1) and Tramtrack (TTK) motifs in gene sets responding to Methoprene treatment – Methoprene is a Juvenile Hormone analog that causes precocious foraging behavior, the TF BR is known to respond to hormone stimulus [62], ADF1 is known to be involved in memory, learning and certain behaviors in Drosophila [63], and TTK is known to have mutant phenotypes affecting aggressive behavior [64]. (None of these motifs could be scored for evolutionary conservation.) The association between H (MCS = 4), a factor involved in sensory organ development (a Juvenile Hormone dependent process [65]), and a set of genes over-expressed in foraging bees (also a Juvenile Hormone dependent condition) [61,66] is also notable (Table 1). The motifs involved in some of the other bee – specific associations are known to play important roles in nervous system function and development in Drosophila, e.g., Mothers against dpp (MAD) (MCS = 4) is known to regulate synaptic growth [67], and Knirps (KNI) is known to be involved in dendrite morphogenesis [68].

Oxidative phosphorylation genes in Nasonia. In this analysis, we show how predicted motif associations in the target genome may be substantiated by literature-based evidence from another species, rather than by statistical significance of the same association in the other species. The oxidative phosphorylation (OXPHOS) pathway plays a very important role in the production of ATP, the principal source of cellular energy. Evolution of OXPHOS genes, their structure, and regulation, has been studied previously [69], and annotation of genes from this pathway in the

![Figure 5. Example of conserved motif – GO associations between Drosophila and Nasonia. All genes in a GO category in Drosophila are shown as columns, with grey indicating that the ortholog was not found in Nasonia. Dark red indicates motif presence at the threshold used, light red indicates medium strength motif presence, white indicates motif absent in the gene’s promoter. The p-value of association between motif and GO category is shown for each species, at the right end of its row. doi:10.1371/journal.pcbi.1000652.g005](image-url)
**Table 1. Motif associations with gene sets implicated in social behavior in honeybees.**

| Gene set                  | Motif                  | MCS* | Motifsource | A.mel | N.vit | D.mel |
|---------------------------|------------------------|------|-------------|-------|-------|-------|
| Pre-foraging maturation ↑ (top 100) | Trl                    | ?    | F           | 0.001  | 0.066** | 0.254 |
| Methoprene ↓              | 1_ADF1_Q6              | ?    | T           | 0.001  | 0.155  | 0.569 |
| Forager ↑                 | hairy.new.6            | 4    | B           | 0.003  | 0.650  | 0.870 |
| Pre-foraging maturation ↑ (top 100) | I_GAGAFACTOR_Q6         | ?    | T           | 0.004  | 0.399** | 0.654 |
| cGMP ↑                    | Kruppel                | 2    | T           | 0.005  | 0.217  | 0.046** |
| Methoprene ↑              | Ubx.tst                | 4    | F           | 0.008  | 0.767  | 0.855 |
| Methoprene ↓              | Adfl                   | ?    | F           | 0.008  | 0.604  | 0.951 |
| cGMP ↓                    | kni                    | 2    | F           | 0.008  | 0.304  | 0.016** |
| cGMP ↑                    | br-Z3                  | 4    | F           | 0.008* | 0.922  | 0.107 |
| Hive bee to forager transition ↓ (top 100) | CG11085.new.7         | ?    | B           | 0.009  | 0.397  | 0.577 |
| Pre-foraging maturation ↑ (top 100) | CG7056.new.7         | 2    | B           | 0.009  | 0.905  | 0.643 |
| Hive bee to forager transition ↓ (top 100) | Mad                   | 4    | F           | 0.010  | 0.701  | 0.561 |
| Methoprene ↑              | ttk.new.6              | ?    | B           | 0.010  | 0.733  | 0.742 |

Gene sets (column 1) are from [33], associated motifs are listed in column 2, and the p-value of association in *Apis, Nasonia* and *Drosophila* are listed in columns labeled "A.mel," "N.vit" and "D.mel" respectively.

*Motif conservation score.

| p-value < 0.01. | 0.01 ≤ p-value < 0.05. | 0.05 ≤ p-value < 0.1. |
|----------------|-------------------------|-----------------------|
| Motif source: B, B1H; F, flyreg.org data; T, Transfac. |

**Discussion**

Our work explores the following challenging question related to comparative regulatory genomics: how can we "import" the wealth of molecular information in well-studied genomes such as *Drosophila* to the regime of a less studied genome such as *Nasonia*, given that the non-coding regions of the genomes do not align? We address this question in the context of characterizing motif functions. We identify several methodological issues involved here, and present a computational pipeline that incorporates novel solutions to the issues. Our approach is expected to become increasingly relevant as hundreds or even thousands of other metazoan genomes get sequenced in the future.

**Pipeline for motif function map in single genome**

We recognize that there are several alternatives to motif scanning that have not been explored here. Boden and Bailey [39] point out that most motif-scanning software can be characterized by two basic ideas: “maximum-odds and hit count” (finding strong sites and counting them) and “average odds” (summing over all possible sites, weak or strong). We believe that our choice of the “site-LLR” and the Stubb/SWAN methods capture the essence of these two popular ideas, and are therefore representative of existing knowledge on motif scanning. Approaches that use additional information such as phylogenetic profiles [75] are not evaluated here, as it is not clear how evolutionary information may be extracted from a genome whose non-coding part is not alignable with other species. It will be also be interesting to examine if more biophysically inspired methods, like TRAP [76], provide complementary strengths in motif scanning, as suggested in recent work by Roider et al. [1]; however, here we chose to operate within the statistical regime of the HMM that has been studied more extensively in the literature.

The log likelihood ratio (LLR) score computed by Stubb and other HMM-based methods [24,77–78] asks the statistical question: does the motif help "explain" the data (sequence) significantly better than the background model can? In some cases, this may not be the right statistical question to ask. We illustrate this issue with the example of the "HB" motif (consensus: TTTTTTGT). This motif has a high match score to the poly-T string (TTTTTTTTTTT), but this is not only because the motif roughly matches the string, it is also in part because poly-T substrings happen to be more common in the genome than the simple background model can capture. A low order Markov chain that is typically used as
background model may be inadequate to capture frequencies of certain patterns (such as the poly-T substring) in the background sequences; if such patterns happen to be similar to a motif, the inadequacy of the background model will adversely affect the LLR. Sequences; if such patterns happen to be similar to a motif, the certain patterns (such as the poly-T substring) in the background model may be inadequate to capture frequencies of clear the better choice for a sizeable set of motifs. Found that each of the three statistical approaches evaluated is may need to be motif dependent. We examined this issue, and considered the possibility that the choice of motif-scanning method with entire compendia of motifs. Surprisingly, past work has not sites is more efficacious than using strong sites alone, when testing (e.g., [1–2,39]) have each shown that clustering of strong and weak sites and their clustering. Existing work on motif function maps whether some TFs tend to operate mostly through strong binding background sequences significantly better off using a higher motif weight than the value learned from corresponding TFs. Another limitation of our analysis (and one that is unfortunately common to most genomic studies today) is that multiple hypothesis correction, performed here through the use of q-values, does not account for the fact that the underlying association tests are statistically dependent, often being for highly similar motifs or GO sets. Also, our choice of searching only the 5 Kbp upstream regions of genes may lead to missed sites, but we believe that it results in an overall increase in the signal to noise ratio. Other possibilities, such as including all of a gene’s surrounding region up to the neighboring genes on either side, or including intronic regions, provide avenues for future research.

The task of finding statistical associations between gene sets and annotations has itself been a topic of much research, as reviewed in [81]. In particular, several ideas have been presented to deal with redundancies between GO categories. Grossmann et al. [32] deal with “parent-child” relationships present in the GO hierarchy, by modifying the Hypergeometric test. This is closest in spirit to how we handle redundancies among gene sets, except that our approach is designed to work for any pair of gene sets (E and O, following terminology introduced in Methods), and not only for “parent-child” pairs. This is useful not only to deal with redundant pairs of GO categories (such as “odorant binding” and “sensory perception of smell”, with an overlap of 62/69), but also to deal with pairs of motif modules that are largely overlapping.

Extending the pipeline to use information from other species

We have demonstrated, on benchmarks constructed from ChIP-based and genetics-based data, that requiring cross-species conservation of motif – function associations leads to significantly higher specificity. While Nasonia – Drosophila comparison clearly improved specificity compared to single species analysis on Nasonia, we did not see a clear effect of varying evolutionary divergence from the compared species. However, we believe that the reference genome, i.e., where the motif and GO data are “imported” from, is the most judicious choice for cross-species comparison. We also note that our approach is distinct from imposing the conservation requirement at the motif scanning stage, as was done in the alignment-free method of [20]. Our reasoning was that motif – function associations may be “evolutionary robust”; i.e., detectable even though the motif – gene relationships are not detected as being conserved, as illustrated by the discovered association of CG12361 with cyclic nucleotide metabolism.

The choice of the reference genome, for a particular target genome under study, will generally be clear, since the kind of comprehensive molecular data that is required of the reference species is available for only a handful of species. However, we note

| Table 2. Enrichment p-values for the oxidative phosphorylation gene set in Nasonia, shown here for ten motifs implicated in the literature as having a regulatory role in this pathway. |
|---|---|---|---|---|---|---|
| Motif | MCS* | Motif source | p-value | q-value | #commonb | #motif targetsc | #genes in gene setd | #total* |
| btd.new.6 | ? | B | 0.0117 | 0.0776 | 10 | 696 | 58 | 9097 |
| V_APL_C | 4 | T | 0.0322 | 0.0776 | 9 | 701 | 58 | 9097 |
| pho | 4 | F | 0.0362 | 0.0776 | 9 | 716 | 58 | 9097 |
| PR1 | ? | L | 0.0461 | 0.0776 | 9 | 749 | 58 | 9097 |
| nrg | ? | L | 0.2491 | 0.3355 | 6 | 666 | 58 | 9097 |
| dATF2 | ? | L | 0.3744 | 0.4202 | 5 | 630 | 58 | 9097 |
| Ewg | ? | L | 0.4557 | 0.4323 | 5 | 693 | 58 | 9097 |
| I_DREF_Q3 | ? | T | 0.5135 | 0.4323 | 5 | 739 | 58 | 9097 |
| OXBOX | ? | L | 0.6496 | 0.4861 | 3 | 518 | 58 | 9097 |
| Dref | ? | F | 0.8672 | 0.5316 | 3 | 749 | 58 | 9097 |
| E-box | ? | L | 0.8684 | 0.5316 | 3 | 751 | 58 | 9097 |

Motif source: B, B1H; F, flyreg.org data; T, Transfac; L, literature.
*Motif conservation score.
1Number of genes in motif target and oxidative phosphorylation genes.
2Number of motif target genes.
3Number of genes in oxidative phosphorylation gene set.
4Total number of genes in the analysis.
doi:10.1371/journal.pcbi.1000652.t002
that the motifs and GO annotations that are “imported” from the reference to the target genome need to be by and large conserved: if the divergence is too great, (a) most motifs will receive low conservation scores and hence be unreliable, and (b) GO gene sets inferred in the target genome will be highly erroneous, leading to very few significant motif associations.

Our approach to characterizing motif conservation levels is only a first step to solve an important problem in comparative regulatory genomics – to use motifs characterized in one species for analyzing the genome of a highly diverged species. Morozov and Siggia [34] have considered this problem for yeast TFs, and have attempted to model the impact of key residue changes on binding specificity. Similar goals have been pursued by Noyes et al. [82] for homeodomain factors in Drosophila. For now, our pipeline only uses information on conservation (or substitution to a similar amino acid) to roughly estimate the impact on binding specificity, but future versions will attempt to do this in a more quantitative and sensitive manner.

We also note that functional characterization of a transcription factor may be undertaken in a more direct manner through ChIP-chip or ChIP-seq assays for the factor’s binding locations, and may even be coupled with cross species comparison to achieve high specificity. Given the current technology, this approach is clearly more expensive than computational frameworks such as ours, although it can serve as a follow-up to specific motif associations identified computationally.

Methods

Sequence data

5 Kbp promoters of D. melanogaster (Release 5) and D. virilis (Release 1.2) were obtained from FlyBase [83]. A. gambiae (Feb. 2003) promoters were downloaded from UCSC Genome Browser Database [84]. Promoters of A. mellifera (Amel_2.0), N. vitripennis (Nvit_1.0), RefSeq set only – 9163 genes), and T. castaneum (Tcas_1.0) were taken from HGSC (http://www.hgsc.bcm.tmc.edu/).

Motif compendium

A total of 224 motifs were obtained from Transfac [13] (40 motifs), FlyReg [85] (52 motifs), the literature [69,72,74] (7 motifs), and from [26] (125 motifs) (Supplementary Text S2).

Details of SWAN and Stubb motif scanning methods

(a) Raw score of each window was computed as the LLR described in Results. (b) P-value of the window’s raw score was computed empirically based on 1000 genomic windows with the same G/C content as the original window. This is referred to as the “PGC” technique below. (c) A gene was declared as a motif target if any window in its 5 Kbp promoter had a p-value below 0.005. “Stubb” scores were computed using the SWAN program and a motif weight of 0. Other details are identical to SWAN. The “background state” in the two-state HMM used by Stubb and SWAN (to score a sequence as well as in learning the motif weight) was set to emit according to single nucleotide frequencies in the sequence under consideration (i.e., a ‘local background”).

Details of “site-LLR” motif scanning method

The log-likelihood ratio (LLR) score of a string \( \delta \), given a motif \( W \), is defined as \( \log \frac{P(\delta | W)}{P(\delta | Bkg)} \), where “Bkg” refers to the background model (same as for SWAN above). Given a motif, we computed the maximum possible LLR score of a site (over all possible sites), and using a threshold equal to 0.9 times this maximum LLR, we marked all sites that were above the threshold.

Empirical p-value of a sequence window was computed as per the following ordering: (i) a window with more marked sites scores higher; (ii) if two windows have the same number of sites, the window with the stronger individual site scores higher.

Defining GO gene sets

The homology map among Anopheles,Apis, Drosophila, Nasonia, and Tribolium was obtained from http://cegg.unige.ch/. For each Drosophila gene in a GO category (from association files published at http://genemerge.cbcb.umd.edu/associationfiles/ in November 2003), all its orthologs in the second species were included in the GO category definition for that species. Thus, a GO category’s cardinality may be different in different species.

Evaluation of methods

For a given motif, its target sequences were determined in the set of real Drosophila promoters and in the set of shuffled promoters. Motif modules thus determined were tested for association with GO categories, and a “true positive versus false positive curve” was drawn to plot the number of associations in the real set and in the shuffled set, at different thresholds of significance (see Figure 2A for two example plots). Any two motif scanning methods were compared in the following three ways: (a) “Strong criterion”: the curve of one method completely dominates the plot for the other method. (b) “AUC”: the area under one curve is greater than that under the other curve. (c) “NP”: the number of associations in the real set at a significance threshold where the number of associations in the shuffled set is 0. The following methods and techniques were evaluated:

(a) Scanning method (site-LLR, Stubb, SWAN).

(b) “PGC” technique: as mentioned above, the raw score of a particular method on a given sequence window was converted into an empirical p-value of motif occurrence by comparing with scores of 1000 randomly selected windows.

We evaluated two different ways to choose these windows: (i) from all non-coding sequence windows with the same G/C-content (“PGC” technique), and (ii) from all non-coding sequence (non-PGC technique). Figure 2B shows that the former was superior on the majority of motifs.

Multiple hypothesis corrections

Q-values [31] are calculated for each motif, correcting for all GO association tests, as well as for each GO category, correcting for all motif association tests.

Dealing with redundant associations

Let \( M \) be a set of interest in the universe \( U \), and let \( E \) and \( O \) be two other subsets of \( U \), with the cardinality of \( U, M, E \) and \( O \) being \( N, m, n_1 \) and \( n_2 \) respectively. Let \( M \cap E \cap O \) be the “background state” in the two-state HMM used by Stubb and SWAN (to score a sequence as well as in learning the motif weight) was set to emit according to single nucleotide frequencies in the sequence under consideration (i.e., a ‘local background”).

The unconditional (traditional) p-value of association between \( M \) and \( O \) is given by the probability that a random set of size \( | O \| = n_2 \), has an overlap of size greater than or equal to \( | M \cap O \| = \lambda \), with \( M \). We answer the above question by calculating the probability of this event conditional on the observed overlap cardinality between \( M \) and \( E \) and that between \( E \) and \( O \). In other words, if \( R \) is a random subset of \( U \), with cardinality \( n_2 \), we calculate the probability \( \Pr([R \cap M \geq \lambda] \text{ conditional on } |R| = n_2 \text{ and } |R \cap E| = \alpha, \text{ where } E \text{ is a fixed subset of cardinality } n_1 \text{ and } |M \cap \)
E | = γ. This is computed as:

\[
\min \left( \frac{m}{n}, \frac{k}{\beta} \right) \left( \gamma \right) \left( m - \gamma \right) \left( n_1 - \gamma \right) \left( N - m - n_1 + \gamma \right) \left( \frac{1}{\beta} \right) \left( \frac{1}{\gamma} \right) \left( n_2 - z - k + \beta \right) \left( \frac{1}{\gamma} \right) \left( \frac{1}{\beta} \right) \left( \frac{1}{\gamma} \right) \left( n_2 - z \right)
\]

We note that this is an example of the multivariate hypergeometric distribution.

**Combined p-values**

A motif was tested for consistent association with a gene set in multiple species as follows. Let \( \pi_1, \pi_2, \ldots, \pi_k \) be the p-values of a motif–function association in \( k \) different species. We first compute the combined statistic:

\[
p_c = 1 - \left( \prod_{j=1}^{k} (1 - p_j) \right)
\]

We denote this random variable as \( p_c \) and its observed value as \( p_c^{(0)} \). Under the null hypothesis that each \( p_i \) is uniformly distributed, we computed the probability \( \pi_c \) that the combined statistic \( p_c \) has a value less than or equal to the observed value \( p_c^{(0)} \), i.e., p-value for the combined statistic \( p_c \), as follows:

\[
\pi_c(x) = Pr(p_c < x) = Pr(\Pi (1 - p_j) > 1 - x) = 1 - Pr(\Pi (1 - p_j) < 1 - x)
\]

\[
= 1 - \left( 1 - x \right) \sum_{i=0}^{k-1} (-1)^i \frac{(\ln(1 - x))^i}{i!}
\]

where the last step is due to the fact that \( 1 - p_j \) is uniformly distributed in [0,1] under the null hypothesis. The random variable \( p_c \) has the desirable property that it is low only if each \( p_i \) is low, and thus captures consistent motif association (low p-value) in all species. (Contrast this with the product of the \( p_i \)'s, which may be low even if one or more of the \( p_i \)'s is close to 1.) Note however, the p-value \( \pi_c \) computed above corresponds to the strong null hypothesis that every individual \( p_i \) is uniformly distributed.

**Assessing motif conservation between Drosophila and Nasonia**

We used an offline version of the tool described in [34]. Starting with the full complement of protein sequences in either genome, this tool first uses the HMMER software version 2.3.2 (http://hmmer.janelia.org/) to scan for matches to DNA-binding domains catalogued in the PFam database [86]. For each TF in *Drosophila*, it then aligns each of its DNA-binding domains to the most similar domain match in *Nasonia*. It then adds to this pairwise alignment a third domain that belongs to the same family and corresponds to a protein whose structure (in DNA-bound state) is available from PDB [87]. By using this (aligned) domain with structural information, the tool identifies DNA-contacting residues (that make either backbone or side-chain contacts) based on a distance threshold. We consider these DNA-contacting residues as the "key" residues. For matches to the zinc finger family (ZF-C2H2), we define key residues to be the four residues identified previously as imparting DNA-recognition capability to this family [35]. Limiting our attention to the key residues only, we then determine if the orthologs from the two species have undergone an amino-acid substitution, and if so, whether the substitution has been to a similar amino acid, as defined by grouping amino acids into one of the following seven classes [90]: (i) amino acids (a,a)'s with aliphatic R-groups (G,A,V,L,I), (ii) non-aromatic a's with hydroxyl R-groups (S,T), (iii) a's with sulfur-containing R-groups (C,M), (iv) acidic a's and their amides (D,N,E,Q), (v) basic a's (R,K,H), (vi) a's with aromatic rings (F,Y,W), and (vii) amino acids (P). Finally, each motif was assigned a conservation score that could take the value \( \frac{1}{\pi} \) or an integer between 1 and 4 (4 for strongest conservation), as per criteria defined in Table S2. The alignments of DNA-binding domains are available in Table S4. Of the 160 motifs that received a motif conservation score (not \( \frac{1}{\pi} \)), 114 (71%) had the highest score of 4, 8% were scored at 3, and 20% received the low score of 2, indicating greater potential for evolutionary change.

**Assessing cross-species comparison**

To construct benchmarks based on ChIP data, we collected published target genes for 10 TFs: BCD, Caudal (CAD), Giant (GT), HB, and Kruppel (KR) from [89], Dorsal (DL), Snail (SNA), and Twist (TWI) from [90], gaga factor (GAF) from http://intermine.modencode.org/, and PHO from [91]. Statistically significant motif–GO function associations were identified using the Hypergeometric test and E-value cutoffs (0.001, 0.01, and 0.05). (E-value here is the product of the p-value from the Hypergeometric test and the number of GO terms tested for.) These were treated as the "true" associations, and the associations predicted by the motif function map were evaluated against this benchmark. The following four measures were calculated for each TF and an average over all TFs was computed: (a) "PrecAt5": precision (number of correct predictions, divided by total number of predictions) when considering the top 5 predicted associations, (b) "PrecAt10": precision in the top 10 predictions, (c) "PrecAtPval0.005": precision in the associations with p-value less than 0.005, and (d) "PrecEqRecall": precision when the number of predicted associations is equal to the number of true associations.

The benchmark based on genetics data was constructed similarly with published target genes for the REDfly database [25] except that we used less stringent E-value cutoffs (1, 10, and 50) since the target gene sets here are smaller but more reliable than in the ChIP-based benchmark.

**Supplementary website**

5 Kbp promoter sequences, promoter and gene mapping information, motifs, GO gene sets, source code for SWAN, and a link to web interface for a motif function map are available at our site http://europa.cs.uiuc.edu/CompGenomics09/.

**Supporting Information**

**Figure S1** Two-state HMM of the Stubb program. Found at: doi:10.1371/journal.pcbi.1000652.s001 (0.20 MB DOC)

**Figure S2** Comparison of different motif scanning methods. Found at: doi:10.1371/journal.pcbi.1000652.s002 (0.09 MB DOC)

**Figure S3** Performance of predicted motif - GO associations using cross-species comparison, evaluated based on genetics-based binding data. Found at: doi:10.1371/journal.pcbi.1000652.s003 (0.27 MB DOC)

**Figure S4** Predicted motifs for *Drosophila* protein CG7056-PA and its ortholog in *Nasonia* using the online tool at http://ural.wustl.edu/flyhd.
Found at: doi:10.1371/journal.pcbi.1000652.s004 (0.17 MB DOC)

Table S1 Correlation between motif characteristics and amenability to specific methods. Found at: doi:10.1371/journal.pcbi.1000652.s005 (0.05 MB DOC)

Table S2 Possible motif conservation scores and their semantics. See text for explanation of how. Found at: doi:10.1371/journal.pcbi.1000652.s006 (0.03 MB DOC)

Table S3 Motif functional associations common to *Nasonia* and *Drosophila*. Found at: doi:10.1371/journal.pcbi.1000652.s007 (0.31 MB DOC)

Table S4 Alignments of DNA-binding domains in orthologous protein sequences from *Drosophila* and *Nasonia*. Found at: doi:10.1371/journal.pcbi.1000652.s008 (1.66 MB TXT)

Table S5 Motif associations with genes up-regulated in response to Manganese treatment in honeybees. Found at: doi:10.1371/journal.pcbi.1000652.s009 (0.11 MB DOC)

References

1. Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics 25: 433–442.

2. Sinha S, Adler AS, Field Y, Chang HY, Segal E (2008) Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res 18: 477–488.

3. Warner JB, Philppakis AA, Jaeger SA, He FN, Liu J, et al. (2008) Systematic identification of mammalian regulatory motifs’ target genes and functions. Nat Methods 5: 347–355.

4. Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, et al. (2008) Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Comput Biol 4: e1000021.

5. Adler AS, Sinha S, Kawaihara TL, Zhang JY, Segal E, et al. (2007) Motif module conservation scores and their semantics. J Comput Biol 14: 224 motifs used in this study.

6. Bailey TL, Gribskov M (1998) Methods and statistics for combining motif match scores. J Comput Biol 5: 211–221.

7. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, et al. (2004) Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res 32: 1372–1381.

8. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276–287.

9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

10. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and assembly of Drosophila melanogaster chromosome 2R. Nature 426: 219–232.

11. Halfon MS, Gallo SM, Bergman CM (2008) REDfly 2.0: an integrated database of regulatory elements. Bioinformatics 25: 435–442.

12. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098.

13. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.

14. Zhu J, Zhang MQ (1999) SC福德: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics 15: 607–611.

15. Kheradpour P, Stark A, Roy S, Kellin M (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17: 1919–1931.

16. Harbison CT, Gordon DR, Lee TI, Rinaldi NJ, Mains AE, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.

17. Pennacchio LA, Loots GG, Nohrega MA, Ovcharenko I (2007) Predicting tissue-specific enhancers in the human genome. Genome Res 17: 201–211.

18. Berman BP, Pfeiffer BD, Lavery TR, Salberg SL, Rubin GM, et al. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol 3: R61.

Table S6 Enrichment p-values for the oxidative phosphorylation gene set in *Nasonia*. Found at: doi:10.1371/journal.pcbi.1000652.s010 (0.05 MB DOC)

Text S1 Chemoreceptor genes in *Nasonia*. Found at: doi:10.1371/journal.pcbi.1000652.s011 (0.20 MB TXT)

Text S2 224 motifs used in this study. Found at: doi:10.1371/journal.pcbi.1000652.s012 (0.02 MB TXT)

Acknowledgments

We are grateful to Alexandre Morozov for help with the evolutionary analysis of DNA-binding domains. We also thank our anonymous referees for their suggestions.

Author Contributions

Conceived and designed the experiments: JK SS. Performed the experiments: JK RC SS. Analyzed the data: JK SS. Contributed reagents/materials/analysis tools: JK RC BJ SW JDG ON EMZ HMR GER JHW SS. Wrote the paper: JK SS.
37. Blanchette M, Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res 12: 739–748.

38. Girdhan R, Narlikar L, Hartemink A (2008) A Fast, Alignment-free, Conservation-based Method for Transcription Factor Binding Site Discovery. Research in Computational Molecular Biology 2008(RECOMB08).

39. Boden M, Bailey TL (2002) Associating transcription factor-binding site motifs with target GO terms and target genes. Nucleic Acids Res 36: 4106–4117.

40. Fortini ME, Artavanis-Tsakonas S (1994) The suppressor of hairy protein participates in notch receptor signaling. Cell 79: 273–282.

41. Louvi A, Artavanis-Tsakonas S (2006) Notch signalling in vertebrate neural development. Nat Rev Neurosci 7: 93–102.

42. Delorenzi M, Bienz M (1990) Expression of Abdominal-B homeoproteins in Drosophila embryos. Development 108: 923–929.

43. Nelson WJ (2003) Tube morphogenesis: closure, but many openings remain. Trends Cell Biol 13: 613–621.

44. van Stember B, Delrow J, Bussmeker HJ (2005) Genome-wide analysis of Dorsal, Twist, and Snail suggests integration of Dorsal, Twist, and Snail binding site motifs with target GO terms and target genes. Proc Natl Acad Sci U S A 100: 2580–2585.

45. Pinnell J, Lindeman P, Colavito S, Lowe C, Savage RM (2006) The divergent roles of the segmentation gene hunchback. Integr Comp Biol 46: 519–532.

46. Rushlow C, Doyle H, Hoyer T, Levine M (1987) Molecular characterization of the zerknullt region of the Antennapedia gene complex in Drosophila. Genes Dev 1: 1268–1279.

47. Rushlow C, Levine M (1990) Role of the zerknullt gene in dorsal-ventral pattern formation in Drosophila. Adv Genet 27: 277–307.

48. Wang J, Silva JP, Gustafsson CM, Rustin P, Larson NG (2001) Increased in vivo apoptosis in cells lacking mitochondrial DNA gene expression. Proc Natl Acad Sci U S A 98: 4039–4045.

49. Casares F, Mann RS (1998) Control of anterior versus leg development in Drosophila. Nature 392: 723–726.

50. Rogulj-Ottaa A, Renner S, Trzelna GM (2000) Antagonistic roles for Ultrabithorax and Antennapedia in regulating segment-specific apotoposis of differentiated motoneurons in the Drosophila embryonic central nervous system. Development 135: 3435–3445.

51. Sorger PK (1993) Heat shock factor and the heat shock response. Cell 65: 363–366.

52. Xiong WC, Montell C (1995) Defective glia induce neuronal apoptosis in the repo visual system of Drosophila. Neuron 14: 501–509.

53. Huang Z, Kunes S (1998) Signals transmitted along retinal axons in Drosophila: Hedgehog signal reception and the cell circuitry of lamina cartridge assembly. Development 125: 3753–3764.

54. Kurant E, Pai CY, Sharf R, Halachmi N, Sun YH, et al. (1998) Dorsotonals/Samshara, the Drosophila homologue of meis1, interacts with extradenticle in neuron dendrites. Genes Dev 20: 820–835.

55. Bhat KM, Farkas G, Karch F, Gyurkovics H, Gausz J, et al. (1996) The GAGA factor is required in the early Drosophila embryo not only for transcriptional regulation but also for nuclear division. Development 122: 1113–1124.

56. Gindhart JG Jr, King AN, Kaufman TC (1995) Characterization of the cis-regulatory region of the Drosophila homeotic gene Sex combs reduced. Genetics 140: 781–795.

57. Ward EJ, Skeath JB (2000) Characterization of a novel subset of cardiac cells and their progenitors in the Drosophila embryo. Development 127: 4959–4969.

58. Roy S, Shashikhara LS, VijayRaghavan K (1997) Muscles in the Drosophila embryo. Dev Biol 203: 65–78.

59. Gindhart Jr, Keich U, Garretson JS, Bhaskar A, Liachko I, et al. (2008) Computational detection of significant variation in binding affinity across two sets of sequences with application to the analysis of replication origins in yeast. BMC Bioinformatics 9: 372.

60. Robin S, Schbath S, Vandewalle V (2007) Statistical tests to compare motif count exceptionalities. BMC Bioinformatics 8: 84.

61. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucl Acids Res 37: 1–13.

62. Noyes MB, Christensen RG, Wakayabashi A, Stormord GD, Brodsky MH, et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133: 1277–1289.

63. Crosby MA, Goodman JL, Strelkus VB, Zhang P, Gelbart WM (2007) FlyBase: genomes by the dozen. Nucl Acids Res 35: D496–491.

64. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, et al. (2008) The UCSC Genome Browser Database: 2008 update. Nucl Acids Res 36: D73–779.

65. Bergman CM, Carlson JW, Celniker SE (2005) Drosophila DNAse I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747–1749.

66. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucl Acids Res 36: D211–218.

67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucl Acids Res 28: 235–242.

68. Creighton TE (1992) Proteins: Structures and Molecular Properties W. H. Freeman.

69. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol 6: e27.

70. Zeitlinger J, Zinnen RP, Stark A, Kells M, Zhang H, et al. (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.

71. Kwong C, Adryan B, Bell I, Meadows L, Russell S, et al. (2008) Stability and dynamics of polycomb target sites in Drosophila development. PLoS Genet 4: e1000178.