TargetOrtho: A Phylogenetic Footprinting Tool to Identify Transcription Factor Targets

Lori Glenwinkel,[1] Di Wu,[2] Gregory Minevich, and Oliver Hobert[1]
Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, New York, New York 10032

ABSTRACT The identification of the regulatory targets of transcription factors is central to our understanding of how transcription factors fulfill their many key roles in development and homeostasis. DNA-binding sites have been uncovered for many transcription factors through a number of experimental approaches, but it has proven difficult to use this binding site information to reliably predict transcription factor target genes in genomic sequence space. Using the nematode *Caenorhabditis elegans* and other related nematode species as a starting point, we describe here a bioinformatic pipeline that identifies potential transcription factor target genes from genomic sequences. Among the key features of this pipeline is the use of sequence conservation of transcription-factor-binding sites in related species. Rather than using aligned genomic DNA sequences from the genomes of multiple species as a starting point, TargetOrtho scans related genome sequences independently for matches to user-provided transcription-factor-binding motifs, assigns motif matches to adjacent genes, and then determines whether orthologous genes in different species also contain motif matches. We validate TargetOrtho by identifying previously characterized targets of three different types of transcription factors in *C. elegans*, and we use TargetOrtho to identify novel target genes of the Collier/Olf/EBF transcription factor UNC-3 in *C. elegans* ventral nerve cord motor neurons. We have also implemented the use of TargetOrtho in *Drosophila melanogaster* using conservation among five species in the *D. melanogaster* species subgroup for target gene discovery.

TRANSCRIPTION factors (TFs) and small RNAs represent the largest families of gene regulatory molecules in eukaryotes. Identifying target genes for these regulatory factors is a key challenge that remains to be solved. While targets of regulatory RNAs can often be inferred by sequence complementarity, there are no clearly delineated rules to *de novo* predict DNA sequence targets of DNA-binding domains of transcription factors.

In vitro techniques such as CASTing (cyclic amplification and selection of targets) (Wright *et al.* 1991), EMSA (electrophoretic mobility shift assay) (Hellman and Fried 2007), and multiple sequence comparisons between small sets of hand picked *cis*-regulatory sequences, as well as *in vivo* techniques such as DNase-seq (Song and Crawford 2010) and ChIP-seq (Carey *et al.* 2009) or mutational analysis of transcription factor-regulated reporter genes, have allowed the derivation of high-information-content consensus-binding motifs for many transcription factors. While ChIP-seq allows for the genome-wide identification of transcription-factor-binding sites (TFBSs), in cases where the signal-to-noise ratio of TF binding is small, a certain level of nonfunctional TF binding is expected to occur, rendering it difficult to predict true regulatory targets with high confidence without utilizing additional predictive strategies.

Using a set of experimentally verified binding sequences, it is possible to build a representative position weight matrix (PWM) and to perform a purely bioinformatic genome-wide search for TF consensus sites. This approach provides a cost- and time-efficient alternative to *in vivo* experiments, and, with the accessibility of whole-genome sequence data, multiple species genomes are available for a comparative genomic analysis that utilizes conservation of binding sites between species. Strong purifying selection is expected to maintain binding elements in functional regions so that conservation of TFBS between species is predictive of function.

doi: 10.1534/genetics.113.160721
Manuscript received January 23, 2014; accepted for publication February 9, 2014; published Early Online February 20, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160721/-/DC1.
[1]Corresponding authors: Columbia University Medical Center, 701 W. 168th St., HHSC 724, New York, NY 10032. E-mail: or38@columbia.edu; lag2175@columbia.edu
[2]Present address: Department of Endocrinology, Harvard Medical School, Boston, MA 02135.

While sequence conservation may suggest function, additional predictive criteria, including binding-site enrichment among orthologous regulatory regions together with expression profiling data or chromatin immunoprecipitation (ChIP) data, especially tissue-specific data, provide a multi-faceted approach for confident regulatory target gene prediction.

Existing tools such as the MEME suite (Bailey and Elkan 1994; Bailey *et al.* 2009), PhyloCon (Wang 2007), PhyME (Sinha *et al.* 2005), PhyloGibbs (Siddharthan *et al.* 2005), and EvoPrinter (Odenwald *et al.* 2005) that utilize sequence conservation for motif discovery as well as programs like EEL (Hallikas *et al.* 2006), which evaluate regulatory modules genome-wide without incorporating sequence conservation, are excellent resources for identifying a TFBS to build a PWM or for identifying novel target genes without considering conservation. These programs do not provide a way to assess the novel regulatory targets of a given TF or do not include sequence conservation for functional prediction, however. TargetOrtho fills this gap by providing an alignment-free conservation assignment of orthologous motifs that is independent of motif orientation and outperforms pairwise alignment methods (Elemento and Tavazoie 2005; Gordân *et al.* 2010). This more relaxed definition of conservation accounts for the inherent degeneracy and orientation independence of TFBS so that variant nucleotides within a motif do not prevent conservation calls between species. Such strategies for target gene prediction have been implemented for specific TF regulatory target gene discovery (Aerts *et al.* 2006; Ward and Bussemaker 2008; Herrmann *et al.* 2012), but these approaches have not been applied to the automated prediction of TF regulatory target genes from user-defined PWMs together with a target gene ranking system that accounts for the degree of motif match conservation, quality, and frequency for target gene prediction. For an overview of target gene prediction strategies, see Aerts *et al.* 2012.

We have previously described one framework for the application of an exhaustive *in silico* approach for the identification of transcription factor target genes using experimentally derived consensus-binding sites together with an alignment-free assignment of conservation across multiple species genomes (Bigelow *et al.* 2004). This application, called CisOrtho, compared genome scans of two distinct nematode genomes. We describe here a number of significant expansions to this original pipeline. The new pipeline, TargetOrtho, includes (1) an expansion from the PWM search of two genomes to that of the genomes of five species (see "genomes" in Supporting Information, File S1); (2) region-specific and alignment-independent conservation assignments controlled by user-defined positional conservation constraints between orthologous motif matches; (3) display of binding-site frequency by gene region and cross-species motif match score-based filtering by gene region; (4) the option to restrict motif location relative to the first or last exon of a gene; and (5) the ability to display predicted binding sites on standard genome browsers including the Wormbase and FlyBase Gbrowse tools in the form of bed-formatted genome browser track files where sites are shaded according to predicted binding-site strength as derived from the binding-site log-likelihood ratio score. The new ranking scheme used by TargetOrtho can be finely tuned by the user by scaling the weight of a given filtering criteria. Moreover, we have expanded TargetOrtho to include an option to search each genome against up to five co-occurrences of TFBSs using up to five predetermined PWMs for the discovery of conserved, enriched *cis*-regulatory modules (CRMs). The CRM option allows the user to restrict the nucleotide distance between TFBSs in the same gene region as well as the order of the TFBSs by using the order from the user's uploaded input motifs. Further filtering may be applied through user-selected query lists that restrict the results or report specifically on a subset of genes such as putative target genes determined through expression-profiling experiments, ChIP-ChIP/ChIP-seq data, or gene ontology associations. Finally, TargetOrtho can now be used for target gene discovery in both *Caenorhabditis* and *Drosophila* species.

Materials and Methods

Ortholog assignments

Nematode ortholog assignments based on Ensembl COMPARA (Vilella *et al.* 2009), which predicts orthology of the longest isoform based on homology as well as on conserved gene order, were downloaded using BioMart WS220 datasets (Smedley *et al.* 2009) .The *melanogaster* subgroup ortholog assignments were downloaded from FlyBase precomputed data files (http://flybase.org/static_pages/downloads/bulkdata7.html, version: gene_orthologs_fb_2013_03.tsv.gz).

Gene coordinates

Exon and gene coordinates for nematode genomes were parsed from gff3 annotations files (current versions: *C. elegans*—WS220; *Caenorhabditis briggsae*—WS234; *Caenorhabditis brenneri*—WS234; *Caenorhabditis remanei*—WS234; *Caenorhabditis japonica*—WS234) downloaded from wormbase's FTP site (ftp://ftp.wormbase.org/pub/wormbase/). Exon and gene coordinates for fly genomes were parsed from exon sequence files (fasta) downloaded from FlyBase precomputed data files (http://flybase.org/static_pages/downloads/bulkdata7.html). Current genome versions include the following: *Drosophila melanogaster*—r-5.1; *Drosophila yakuba*: r-1.3; *Drosophila erecta*—r-1.4; *Drosophila simulans*—r-1.4; and *Drosophila sechelia*—r-1.3.

Source code

TargetOrtho employs the FIMO (Grant *et al.* 2011) tool from the MEME suite (Bailey *et al.* 2009) for genome-wide motif scanning. Motif matches are associated with genes using an ANSI C++ script written by Henry Bigelow. All other Target-Ortho scripts were written in python (or XML for the Galaxy interface scripts) by L. A. Glenwinkel.

Data analysis

Outcomes of comparison tests were determined using the Mann–Whitney–Wilcoxon test using python's scipy.stats module. *q*-values for multiple testing corrections were calculated as in Storey and Tibshirani (2003). *P*-values were accepted as significant if the corresponding q-value was <0.05, which is representative of the minimum false discovery rate that is incurred when calling that test significant.

For each test, motif matches in the set of previously validated transcription factor target genes were compared to a set of 1000 random coding genes for each ranking criteria in each gene region. Six unique gene regions were analyzed (upstream, intron, exon, downstream, best site of any region, and upstream plus intron) for each of eight ranking criteria (*C. elegans* site score, *C. elegans* averaged region score, *C. elegans* site frequency, averaged species site score, averaged species region score, averaged species site frequency, site conservation, and site offset variance measured as the coefficient of variation). In addition, four total gene-ranking criteria (*C. elegans* averaged gene score, *C. elegans* total site frequency per gene, averaged species averaged gene score, and averaged species site frequency per gene) and the cumulative site score derived from all criteria per region were analyzed (see Figure S2 for an overview of the results of all tests in all regions).

For each ranking criteria in each gene region, the best motif match value was considered between comparison groups when several values were present. For example, the best upstream motif-match log-likelihood score per gene region was compared with transcription factor-dependent genes and 1000 random coding genes. Additionally, cumulative site scores derived from upstream and intronic data were compared in previously validated target genes and random genes.

Wilcoxon rank-sum tests were used to compare ventral nerve cord neuron counts in wild-type or *unc-3(e151)* worms (Table S1). See Table S5 for all input parameters used for motif analysis with TargetOrtho.

Gene Ontology term analysis

Gene ontology (GO) enrichment analysis was done using the web-based GOrilla tool (Eden *et al.* 2007, 2009) using the single list of ranked genes option with a *P*-value threshold of 10_e^{-3} using slow mode. See Table S12 for full GO term analysis results. Genes in each ontology category were binned according to the best TargetOrtho upstream or intronic site rank per gene and plotted showing the number of genes in each TargetOrtho ranking bin for selected ontology terms.

Reporter constructs

GFP fusions were generated as in Hobert (2002). The VL6 and BC14284 strains were provided by the Caenorhabditis Genetics Center, which is funded by the National Institutes of Health Office of Research Infrastructure Programs (P40 OD010440). See Table S1 for strain details.

Availability

The TargetOrtho package is available as a command line tool or for installation as a Galaxy tool (Goecks *et al.* 2010). The Galaxy option offers an accessible way to use TargetOrtho on any platform via Galaxy's web hosting option (http://wiki.galaxyproject.org/Admin/Get%20Galaxy). See http://hobertlab.org/targetortho/ for general usage and availability.

Results

To expand the known repertoire of TF target genes for a better understanding of diverse biological processes, we have engineered a bioinformatic pipeline allowing for robust target gene prediction. We first describe the program architecture for the discovery of novel TF target genes as well as target genes regulated by CRMs whereby multiple TFBSs work in concert. In the following sections, we then examine individual criteria for ranking TFBSs across entire genomes and show that, for three motifs with extensive *in vivo*-validated target genes, these criteria are robust predictors of real target genes. Because the regulatory logic of *in vivo* TF binding is not well understood, we implement user-defined adjustments for each of the ranking criteria chosen. We show that the strategy of combining binding-site data from the genomes of multiple species is justified as it drastically improves target gene prediction. Finally, we show that our pipeline further improves target gene prediction by combining the averaged species ranking data into one final cumulative site score for each predicted binding site in the genome.

Features of TargetOrtho

General overview of the pipeline: TargetOrtho provides a comparative genomic approach for the identification of transcription factor target genes for which a collection of binding sites, represented as the PWM, has been experimentally identified. The pipeline is executed in four steps (or five if multiple input PWMs are used). Briefly, genomes of five species are searched for motif matches against a PWM in MEME plain text format (see MEME documentation at http://meme.nbcr.net/meme/doc/meme-format.html and http://meme.nbcr.net/meme/doc/examples/meme_example_output_files/meme.html) derived from experimentally validated binding sites using the FIMO (Grant *et al.* 2011) motif scanner. Sites from each species are then associated with the nearest exon in the upstream and downstream direction and matched to orthologous regions in the reference genome (currently, *C. elegans* or *D. melanogaster*). Finally, filtering and ranking criteria are applied to each reference genome motif match, resulting in a ranked list of sites and their associated target genes. TargetOrtho output consists of browsable HTML tables, tab-delimited text files, and bed-formatted genome browser track files along with a compressed folder containing all results for download (Figure 1 and Table S2). The execution of TargetOrtho is facilitated by Galaxy (Goecks *et al.* 2010),

a general bioinformatics workflow management system in which results are automatically browsable and available for download and sharing from any platform (Figure 2). TargetOrtho can also be installed locally and executed via the command line as a stand-alone program or added as a tool to a locally hosted Galaxy instance (see http://galaxyproject.org). See File S1 for a detailed program overview.

Adjustable program features: TargetOrtho includes several adjustable features (Figure 3 and Table S3): (1) two reference genomes are available for target gene discovery. The *C. elegans* option includes searches across five species of the *Caenorhabditis* genus, while a *D. melanogaster* option includes genome-wide comparative searches across five *melanogaster* subgroup species. A reference genome is defined

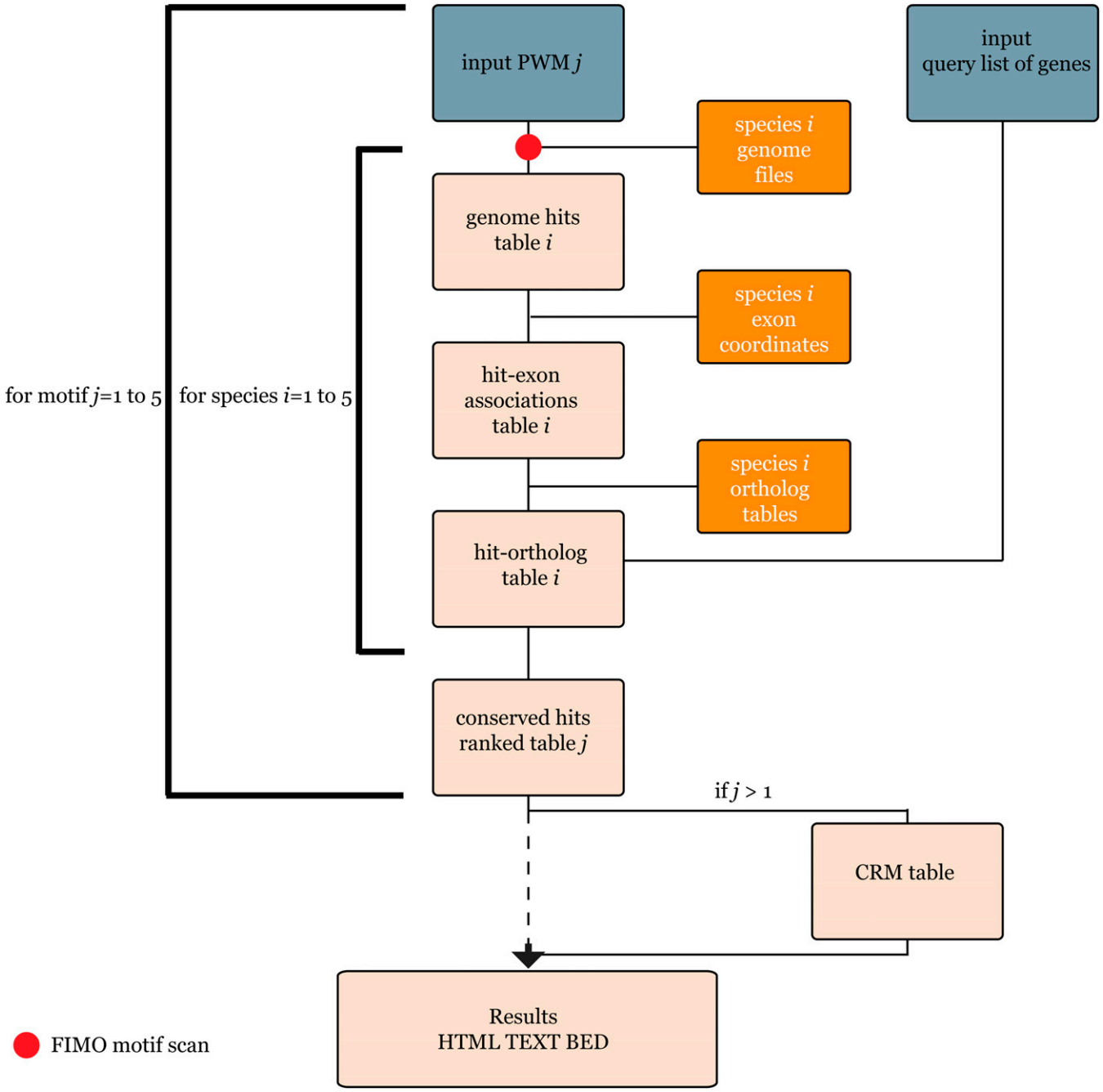


Figure 1 Overview of TargetOrtho pipeline. Beginning with one to five input position weight matrices (PWM_j = 1–5 in meme plain text format) and an optional query list with genes of interest, five species genomes (top orange box) are scanned with the motif scanner FIMO, resulting in one motif match hit table per genome *i* (*i* = 1–5). Each site is then associated with an exon, followed by ortholog pairing between the reference species and each species associated site. Orthologous sites are then ranked according to the TargetOrtho ranking criteria. If more than one input PWM is specified, promoters having at least one motif match for each PWM are filtered to a *cis*-regulatory module table. All results are output as tab-delimited text files, html browsable files, and bed format genome browser files.

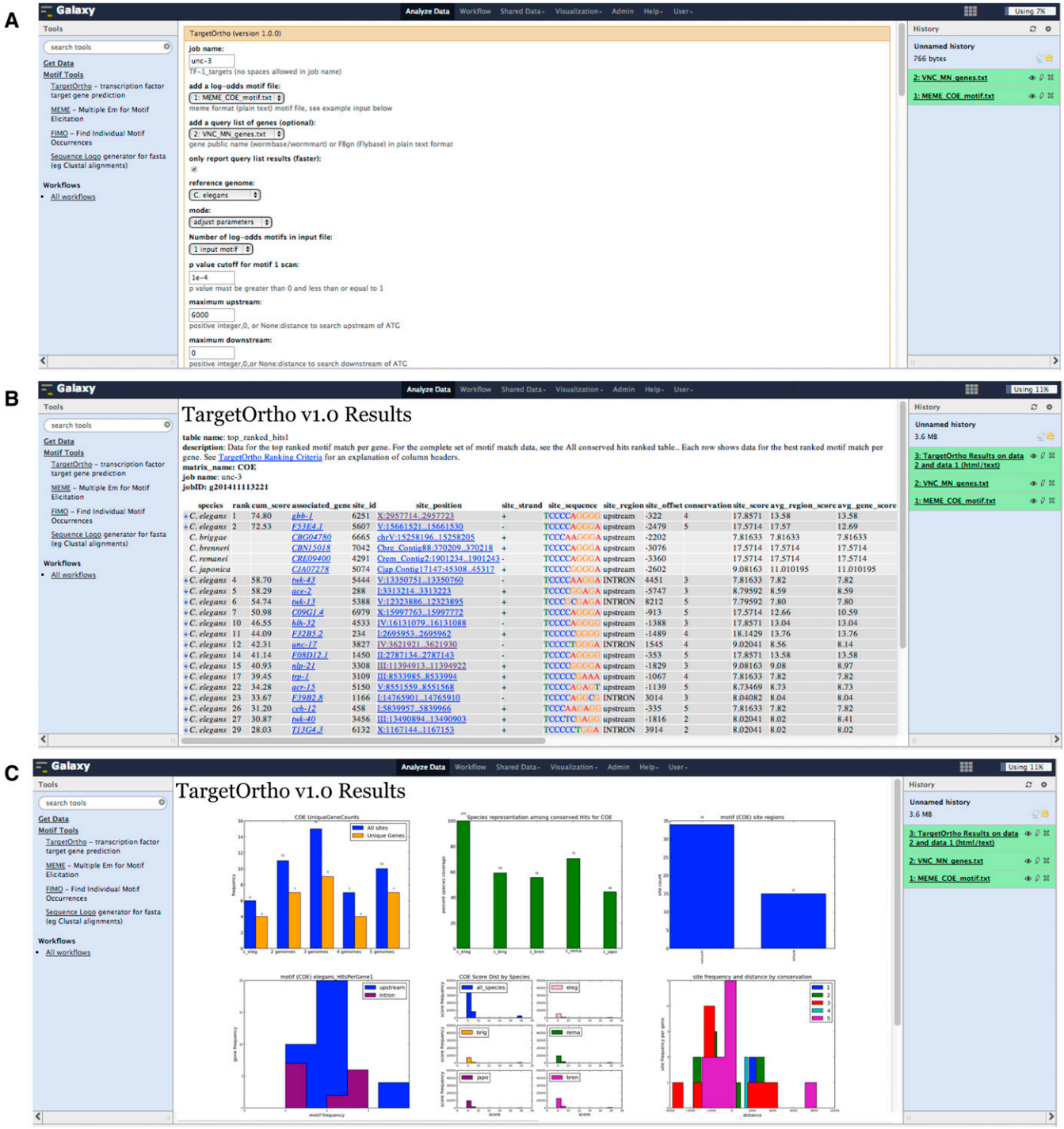


Figure 2 Galaxy screenshots. (A) TargetOrtho user interface hosted by Galaxy. The TargetOrtho tool is shown in the Galaxy tool (left). Two TargetOrtho input files are shown in the History (right). (1) A motif file in Meme version 4 format and (2) a user-defined list of genes in plain text format. These files are uploaded using the "get data" tool built into the Galaxy platform. Adjustable TargetOrtho parameters are shown (middle). (B) TargetOrtho Results screenshot. Upon job completion, two TargetOrtho output files appear in the History (right): TargetOrtho browse results (html/text) is selected and shown in the middle. The top-ranked site per gene table (html version) is displayed along with a link to browse all TargetOrtho output files. A second result file in the History allows for a single-click local download of all results as a compressed directory. (C) TargetOrtho Summary statistics plots are included in the results directory as html files and may be viewed from the Galaxy interface or locally from the downloaded results files. (Top left) Site distribution by conservation. Blue shows all unique motif matches; yellow shows the number of candidate target genes. (Top middle) Species representation among all motif matches. (Top right) Site count by gene region. (Bottom left) Target gene frequency by gene region. (Bottom middle) Log-likelihood motif score distribution by species. (Bottom right) Site positional distribution by species conservation. See Table S2 for additional TargetOrtho results descriptions.

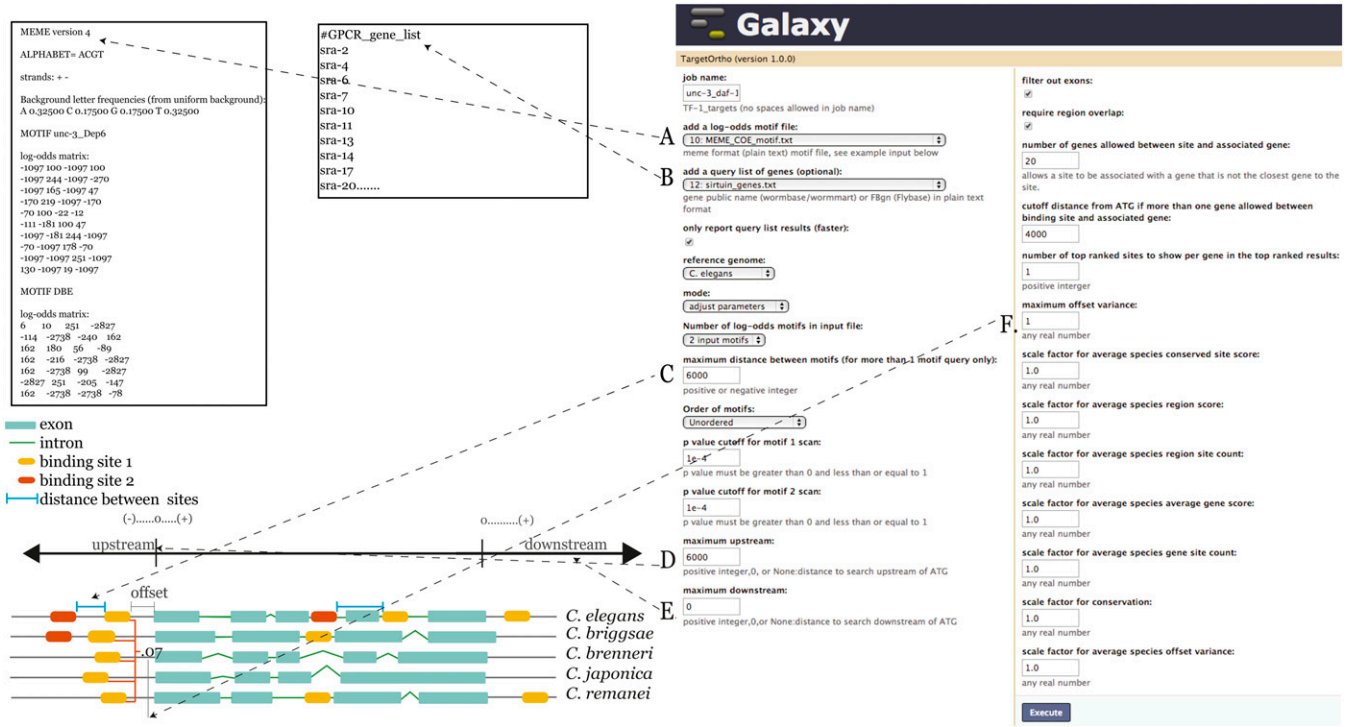
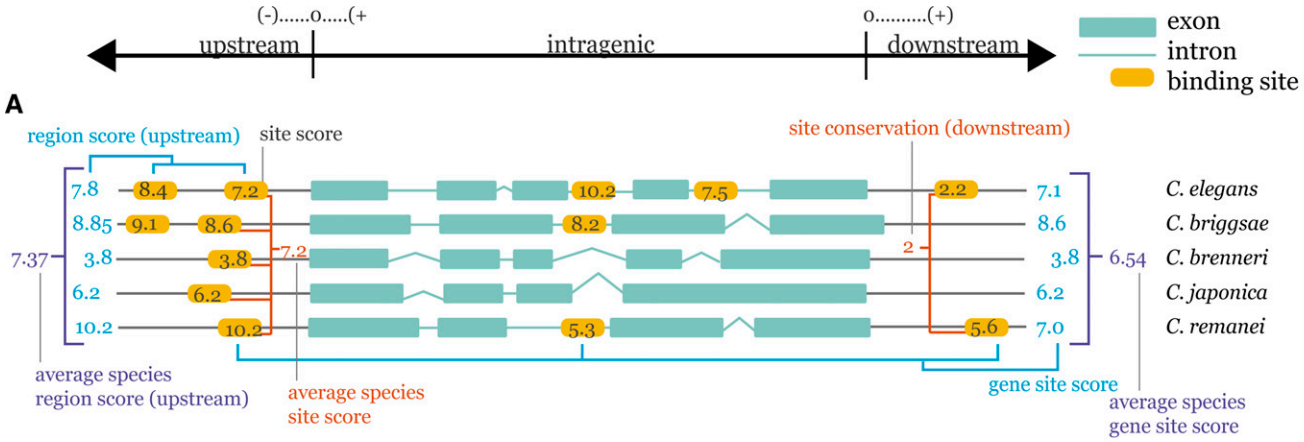


Figure 3 TargetOrtho input parameters. (Right) TargetOrtho user interface hosted on the Galaxy platform. Select TargetOrtho input parameters are shown. (Left) Graphical representation of select input parameters from right panel. The tool interface on Galaxy shows all default values and may be changed by the user. Default values may also be viewed from the command line tool by using the command "python TargetOrtho.py–h ." Each input parameter should be adjusted for the individual input motif/s by the user. See Table S3 for a description of all adjustable input parameters and default values. (A) Example TargetOrtho input motifs for target gene discovery of genes with co-occurrences of motif A and motif B. TargetOrtho takes this input as a Meme version 4 input motif file (http://meme.nbcr.net/meme/doc/meme-format.html) with up to five input motifs. (B) Example of gene query list input file in plain text format showing a subset of user-defined genes for TargetOrtho to specifically report on. Gene names must be in gene public name format (*unc-3*, *ttx-3*) when available; otherwise, transcript names (C09G1.4, F08D12.1) may be used for *C. elegans*. FlyBase gene IDs in the form FBgn must be used for *D. melanogaster* gene names. These may include suspected transcription factor target genes of interest or serve as a negative control list of genes that are expected to be transcription factor independent. An option to report only data for these genes is available (right: "only report query list results"); otherwise, whole-genome results are reported with additional reporting on the query list gene results. (C) The maximum distance between motifs (for more than one motif query only). This option constrains the allowed distance between any two motifs from the motif input file. If five motifs are used as input, this distance limits the distance between any two adjacent motifs where the adjacent motifs are from separate entries in the motif input file. This does not preclude the user from specifying a search for identical motifs in the input file. For example, one may choose to search for target genes having at least two occurrences of motif A in the upstream region. To accomplish this, the user would include motif A two times in the input file. The order of motifs in a given gene region may also be constrained by selecting "ordered' or "unordered" for the "Order of motifs" parameter. If ordered is chosen, co-occurrences of motifs must be positioned in the order given in the motif input file. For example, if motif A, motif B, and motif C are included in the input file with the ordered option, all target gene candidates must have these three motifs in the order motif A, motif B, motif C or in the order motif C, motif B, motif A among all orthologous gene regions for a candidate target gene to be included in TargetOrtho output. (D) The maximum upstream distance that a motif may be positioned for target gene association. (E) The maximum downstream distance that a motif may be positioned for target gene association. In addition to the maximum upstream and downstream distance, the number of intervening genes allowed between any motif match and associated gene, as well as the cutoff distance from the first ATG allowed if any intervening genes are positioned between a motif match and the associated gene, may be specified by the user (right). (F) The maximum offset variance constrains the positional variance allowed between orthologous motif matches between species. The offset variance is calculated by taking the absolute value of the coefficient of variation of each motif match offset (shown as the distance from the first annotated exon of the associated gene) in each species. See *Orthology Matching* section in File S1 for detailed explanation. See Table S3 for explanations of all adjustable parameters.

as the genome from which candidate TF target genes are reported (see "genomes" in Supporting Information, File S1); (2) the distance between distinct motif matches (Figure 3C) and linear motif order for CRM searches (see *CRM searches for multiple motifs*); (3) the offset variance (Figure 3F) of orthologous motif matches to constrain the positional conservation of a motif match (see "orthology matching" in Supporting Information, File S1); (4) the upstream (Figure 3D) and downstream (Figure 3E) motif match distance from the first or last adjacent annotated exon; (5) the number of intervening genes between a motif match and an associated gene may be constrained as well as the intervening distance from the associated gene allowed if annotated genes are positioned between a motif match and its associated gene; and (6) the cumulative site score is constrained by scaling options to weight each site ranking criteria (Figure 4C and Table S3). For example, if the motif frequency among orthologous gene regions is important, the user may up-weight this



TargetOrtho ranking criteria per site	raw score (c_i)	normalized score	average normalized score	site rank
site conservation**	5	100.00 (a)	(cumulative site score*)	45th of 29829
averaged species site score	7.2	94.23 (b)		
averaged species region score	10.7	97.22 (c)		
averaged species region site count	1.4	13.33 (d)	73.84	
averaged species average gene score	6.54	98.85 (e)		
averaged species gene site count	2.6	40.00 (f)		
cross species site offset variance	0.07	96.62 (g)		

(scale factor) default =1 or any user specified real number

*cumulative site score = $[\sum_{i=1}^{n}(c_i-b_i)/(a_i-b_i)100\omega_i] /j$

i=individual ranking criteria
c_i=raw value of ranking criteria i.
a_i =maximum c_i
b_i =minimum c_i
ω_i=weight for ranking criteria i (scale factor shown in red)
j=number of criteria with weight (ω_i)>0

**only sites conserved in 2 or more species are ranked.

Figure 4 TargetOrtho ranking criteria. Each orthologous gene region per species is divided into upstream, intragenic [intron (green line) and exon (green box)], and downstream regions. (A) Log-likelihood score ranking criteria. Individual predicted binding sites (orange bubbles) are overlaid with the site score. "Site score" (black numerals): the log-likelihood ratio score of an individual motif match. "Average species site score" (orange numerals): the averaged site score across orthologous regions between species where each reference species site is matched to the positionally best-matched orthologous-species-region motif match. Best matches are determined by grouping sites across species and filtering for the best offset (site position relative to exon 1 for upstream sites or the last exon for downstream sites). See "offset variance." "Region score" (blue numerals): the average site score within each species across a given region. "Average species region score" (purple numerals): the region score averaged across species for a given region.

factor to inflate the effect of the motif frequency on the cumulative site score. See Table S3 for a description of all adjustable features.

CRM searches for multiple motifs: TargetOrtho includes an option to search each genome against up to five co-occurrences of transcription-factor-binding sites using up to five predetermined PWMs for the discovery of conserved, enriched CRMs. In addition to the filtering applied to individual genome-wide searches, the CRM option allows the user to restrict the nucleotide distance between TFBSs in the same gene region as well as the order of the TFBS by using the order from the user's uploaded motifs (Figure 3C and Table S3). CRM target genes are scored by averaging the adjustable cumulative site score of each component motif (see *Adjustable cumulative site score*).

Binding-site ranking criteria for prediction of regulatory target genes: After conservation assignment, additional criteria were assessed for each site in each genome for eventual cumulative score calculations and final site ranking. Generally, the cumulative site score used for site ranking is determined for each site in a reference genome (*C. elegans* or *D. melanogaster*) according to its binding strength as represented by the log-likelihood ratio score and binding-site frequency associated with the target gene. Each site score and site count is averaged across species for use in the cumulative site-score calculation. Specifically, each site is ranked by the averaged species site score (Figure 4A), the averaged species region score (Figure 4A), the averaged species gene score (Figure 4A), the site conservation (Figure 4A), the offset variance (Figure 4B), the averaged species region site count (Figure 4B), and the averaged species gene site count (Figure 4B). Each site in the reference genome is ranked individually using these ranking criteria.

For example, as shown in Figure 4, consider the site Y with a log-likelihood site score of 7.2 found at −500 nucleotides upstream of a gene and conserved in five species. The averaged species site score of 7.2 is determined by grouping site Y with one orthologous site in each genome and then averaging the site scores across species where site grouping is determined using the minimum positional offset variance (0.07) from the first exon of gene X. The offset variance is also used for site ranking. The averaged species region score of 7.37 is determined by first averaging the site score across the upstream region of gene X as well as the orthologous upstream regions in each species and then averaging this value across species, where the upstream distance is constrained by the user. The averaged species gene score (6.54) is determined by averaging the site score across all gene regions—in this case, the upstream, intron, exon, and downstream regions—for each orthologous gene and then averaging this value across species. An analogous strategy is applied for the site frequency; in this case, the averaged species upstream site count (1.4) and averaged species gene site count (2.6) of gene X. Finally, these criteria are used to generate a final cumulative site score (see *Adjustable cumulative site score*) of 73.84 for TargetOrtho site ranking.

Adjustable cumulative site score: Individual ranking criteria are combined into a single cumulative site score for each site in the reference genome, providing a list of target gene candidates. The cumulative site score is generated as:

$$\text{cumulative site score} = \frac{\sum_{i=1}^{n}(c_i - b_i)(a_i - b_i)100\omega_i}{j},$$

where c_i is the raw ranking criteria value out of n total ranking criteria, a_i is the maximum value from all c_i in a given TargetOrtho search, b_i is the minimum value from all c_i in a given TargetOrtho search, ω_i is an optional scaling factor applied to each ranking criteria (default $\omega_i = 1$), and j is the number of ranking criteria where $\omega_i > 0$. Sites that are found only in the reference genome, and hence are unconserved, were assigned a cumulative site score of zero so that they are automatically ranked last but are still displayed in the TargetOrtho results.

In detail, each motif match is ranked by first determining the average species site counts and averaged species site scores across the associated gene; then each site in the

"Gene site score" (blue numerals): the averaged site score across all regions searched for each species. "Average species gene site score" (purple numerals): the gene site score averaged across species. Conservation (orange numerals): Alignment-independent site conservation is determined by the number of species with at least one predicted binding site in an orthologous region to the reference species motif match. (B) Motif match frequency and position ranking criteria. Individual predicted binding sites (orange) are overlaid with the site offset. "Offset" (black numbers) refers to the site position relative to exon 1 for upstream sites or the last exon for downstream sites. "Offset variance" (orange numerals): the absolute value of the coefficient of variation of the offsets for each matched orthologous motif match between species. Smaller values indicate increased positional constraint compared to motif matches that are differentially positioned between species. "Site count" (blue): The number of predicted binding sites in a given region per species. Averaged species region site count (purple): The site count averaged across orthologous species regions where the region shown is upstream. "Gene site count" (blue): The total site count across all regions of a gene including upstream, intragenic, and downstream (when included) for each species. "Average species gene site count" (purple numerals): the gene site count averaged across all orthologous regions of an associated gene between species. (C) Ranking criteria and cumulative site score per predicted binding site. Column 1, "TargetOrtho ranking criteria per site" indicates the ranking criteria used to calculate the final cumulative score for each predicted binding site (orange) in the reference genome. Column 2, "Raw score": raw values for each ranking criteria described in A and B. Column 3, "Normalized score": Each raw value from column 2 is normalized between 0 and 100 using the minimum and maximum value unique to each motif across the genomes. Column 4, "Average normalized score": The final cumulative score assigned to each predicted binding site in the reference genome is calculated by averaging the normalized scores in column 3. Column 5, "Site rank": The rank order of each predicted binding site taken by ordering each predicted site in the reference genome by the cumulative score.

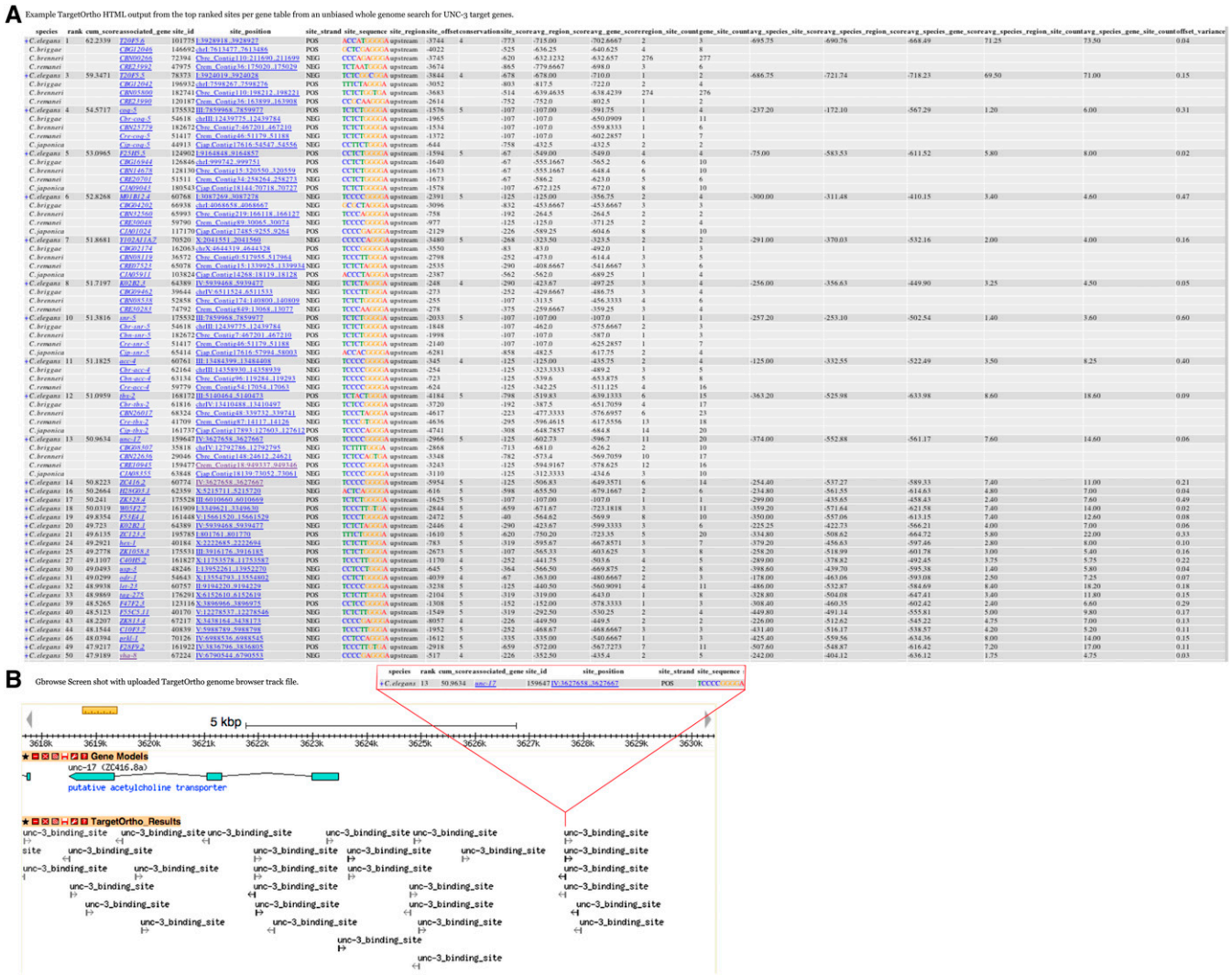
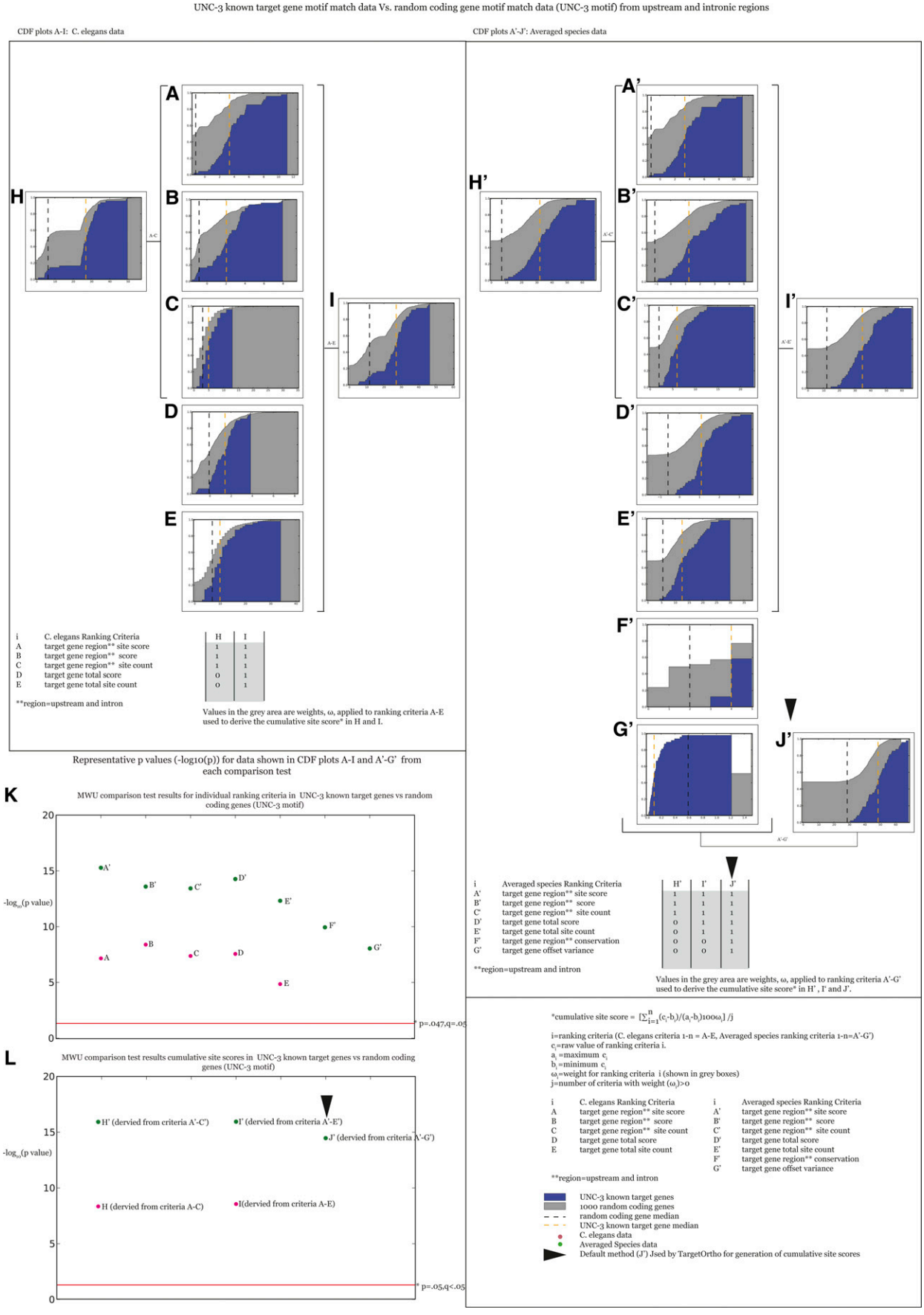


Figure 5 TargetOrtho output example. (A) TargetOrtho top-ranked site per gene table in HTML format. This table is a subset of the "All-conserved-hits-ranked" table (see Table S2 for descriptions of all TargetOrtho output files) showing only the best-ranked site in each candidate target gene as opposed to showing data for every site in every candidate target gene. Each table row shows motif match data for one motif match in the reference genome (*C. elegans* or *D. melanogaster*) with an option to expand the row to show data for other species data. Each top-ranked site shown in the table also includes information about overall site count for the corresponding region and total site count across the entire putative target gene. Additionally, average site scores per region and per gene are shown for each table entry. To see all sites in a gene, consult the "All conserved hits ranked" table. See Figure 4C legend for explanations of column values. (B) Wormbase Gbrowse screenshot of TargetOrtho results. Genome browser track files are output in bed format for viewing predicted binding sites in standard genome browsers. Higher scoring binding sites are shaded darker grey than lower scoring sites. See Table S2 for a description of all TargetOrtho results files.

reference genome is ranked by normalizing each ranking criteria value between 0 and 100 and by averaging the normalized values for each site to obtain the final cumulative site score for sites that are present in at least two orthologous genome regions. Each normalized criteria score may be weighted to affect the cumulative site score according to user preferences (option -A, -B, -C, -D, -E, -F, -G for average species site score, average species region score, average species gene score, average species gene site count, conservation, and offset variance, respectively, where options A–G may be any real number) (Figure 4C; TargetOrtho ranking criteria). Weighting specific ranking criteria may be of interest when prior information is available as to the nature of each ranking criteria in experimentally validated TF target genes. The default strategy of evenly weighting each ranking criteria in the computation of the cumulative site score results in significantly better cumulative site scores in validated target genes compared to random genes in our analysis of three well-characterized *C. elegans* TFBSs.

Program output: TargetOrtho results include a top-ranked-per-gene table for showing the best-ranked site per associated gene as well as an all-conserved-hits-ranked table showing all ranked sites where all motif matches are shown for all candidate target genes. Each site is assigned a rank order corresponding to the cumulative site score where the


UNC-3 known target gene motif match data Vs. random coding gene motif match data (UNC-3 motif) from upstream and intronic regions
CDF plots A-I: C. elegans data
CDF plots A'-J': Averaged species data
A
B
C
D
E
H
I
A'
B'
C'
D'
E'
F'
G'
H'
I'
J'
A-C
A-E
A'-C'
A'-E'
A'-G'
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
**region=upstream and intron
H I
1 1
1 1
1 1
0 1
0 1
Values in the grey area are weights, ω, applied to ranking criteria A-E used to derive the cumulative site score* in H and I.
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
**region=upstream and intron
H' I' J'
1 1 1
1 1 1
1 1 1
0 1 1
0 1 1
0 0 1
0 0 1
Values in the grey area are weights, ω, applied to ranking criteria A'-G' used to derive the cumulative site score* in H' , I' and J'.
Representative p values (-log10(p)) for data shown in CDF plots A-I and A'-G' from each comparison test
K
MWU comparison test results for individual ranking criteria in UNC-3 known target genes vs random coding genes (UNC-3 motif)
-log10(p value)
20
15
10
5
0
A'
B'
C'
D'
E'
F'
G'
A
B
C
D
E
p=.047,q=.05
L
MWU comparison test results cumulative site scores in UNC-3 known target genes vs random coding genes (UNC-3 motif)
-log10(p value)
H' (dervied from criteria A'-C')
I' (dervied from criteria A'-E')
J' (dervied from criteria A'-G')
H (dervied from criteria A-C)
I(dervied from criteria A-E)
* p=.05,q<.05
*cumulative site score = [∑i=1n (ci-bi)/(ai-bi)100ωi] /j
i=ranking criteria (C. elegans criteria 1-n = A-E, Averaged species ranking criteria 1-n=A'-G')
ci=raw value of ranking criteria i.
ai =maximum ci
bi =minimum ci
ωi=weight for ranking criteria i (shown in grey boxes)
j=number of criteria with weight (ωi)>0
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
**region=upstream and intron
UNC-3 known target genes
1000 random coding genes
random coding gene median
UNC-3 known target gene median
C. elegans data
Averaged Species data
Default method (J') Jsed by TargetOrtho for generation of cumulative site scores

best cumulative site score is assigned a rank of 1. Additionally, results tables with all hit-gene associations are included for each species and each motif as well as genome browser track files in bed format. All TargetOrtho outputs are described in Table S2 and Figure 5.

Validation of TargetOrtho using experimentally identified target genes

Strategy: Using three well-characterized TFBSs from *C. elegans* and *in vivo* validation of TargetOrtho predicted target genes, we find that the interspecies motif match score (log-likelihood ratio score), motif conservation, and frequency of TFBSs among orthologous gene regions are successful predictors of TF regulatory target genes. The three TFBSs used for validation of TargetOrtho include the UNC-3-binding site (UNC-3 motif), bound by the terminal selector for cholinergic motor neuron fate in the ventral nerve cord, UNC-3 (Kratsios *et al.* 2012); the TTX-3/CEH-10 heterodimer-binding site (AIY motif), the terminal selector motif for the AIY interneuron (Wenick and Hobert 2004); and the CHE-1-binding site (ASE motif) required for terminal specification of the chemosensory ASE gustatory neurons (Etchberger *et al.* 2007). Several dozen experimentally validated targets genes that contain binding sites for the respective transcription factors have previously been identified. TargetOrtho ranking criteria were compared between TF-dependent genes and 1000 random coding genes for each motif (Figure S1). For a detailed explanation of the data sets and motif construction as well as data set verification bias corrections, see File S1.

Cumulative site scores in upstream and intronic regions better predict regulatory targets of TFs than sites in other gene regions: To assess the predictive value of different gene regions, cumulative site scores derived from data from upstream, upstream + intron, exon, downstream, or the best cumulative site score from any gene region were compared in TF-dependent genes and random coding genes. We find that TF-dependent gene motif matches perform best when upstream and intronic regions are combined to generate the cumulative site score for all analyses performed. Cumulative site scores derived from upstream or upstream + intronic regions resulted in greater differences between TF-dependent genes and random coding genes compared to cumulative site scores derived from other gene regions (Figure S2).

Individual ranking criteria as well as cumulative site scores derived from averaged species data better predict verified TF target genes compared to ranking criteria from a single genome: To assess the predictive value of individual binding-site ranking criteria derived from multiple species as opposed to using a single genome for target gene prediction, we compared individual ranking criteria derived from *C. elegans* data alone or data derived from multiple species. We find that each individual criterion averaged across species shows greater discrimination between TF-dependent gene sites and random gene sites. Comparison tests for individual TargetOrtho site ranking criteria (Figure 6, Figure S3, Figure S4, Figure S5, Figure S6, and Figure S7) suggest that averaging multiple species data (Figure 6, A′–G′) results in more significant differences between criteria in TF-dependent genes and random coding genes compared to ranking criteria data from the reference genome alone (Figure 6, A–E and K). Also see Table S6, Table S7, Table S8, Table S9, Table S10, and Table S11, and corresponding Figure S3, Figure S4, Figure S5, Figure S6, and Figure S7.

To assess the predictive value of cumulative site scores derived from averaged species data compared to scores derived from a single species, we compared cumulative site scores in TF-dependent genes to scores in random genes for both cases. Generating the cumulative site score from combined averaged species data (Figure 6, H′–J′; Figure S3, Figure S4, Figure S5, Figure S6, and Figure S7) increases the significance of the difference between TF target gene sites and random gene sites compared to building cumulative site scores from the upstream and intronic site information in the reference genome alone (Figure 6, H, I, and L). Also see corresponding Figure S3, Figure S4, Figure S5, Figure S6, and Figure S7.

Figure 6 UNC-3 motif analysis. *unc-3*-dependent target gene data (blue) compared to random coding gene data (gray). The set of previously characterized *unc-3*-dependent genes and 1000 random coding genes were submitted to TargetOrtho using the UNC-3 motif as input (Figure S1A). Data distributions for each TargetOrtho ranking criterion were compared between known target genes and random coding genes. CDF plots of individual ranking criteria (plots A–E and plots A′–G′): CDF plots are shown for individual ranking criteria A–E and A′–G′. TargetOrtho ranking criteria derived from averaged species data (A′–G′) better distinguish previously validated TF target genes from random genes compared to using *C. elegans* (reference genome) data alone (A–E). CDF plots A–E show only ranking criteria derived from *C. elegans* genome data while CDF plots A′–E′ show the corresponding ranking criteria derived from averaged species data. CDF plots F′ and G′ show averaged species data having no reference genome counterpart, including the conservation and offset variance data distributions. CDF plots of cumulative site scores (plots H and I and plots H′–J′): Data distributions for cumulative site scores derived from unique combinations of TargetOrtho ranking criteria are shown in CDF plots H, I, H′, I′, and J′. CDF plot H shows the cumulative site score distributions derived from *C. elegans* upstream and intronic data only calculated from A–C. (Left) Plots A′-C′ show the cumulative site score CDF plots calculated from the corresponding averaged species upstream and intronic data. CDF plot I shows cumulative site scores derived from criteria shown in CDF plots A–E where CDF plots D and E represent total gene ranking criteria in *C. elegans* only. (D) *C. elegans* averaged upstream and intronic site scores. (E) *C. elegans* averaged site score across all gene regions. CDF plot I′ (left) shows the data distribution of cumulative site scores derived from A′–E′ where CDF plots D′ and E′ represent the corresponding total gene ranking criteria averaged across species. CDF plot J′ shows cumulative site scores derived from all averaged species ranking criteria (A′–G′). (K) $-\text{Log}_{10}$ (*P*-value) for each ranking criteria comparison test where transcription-factor-dependent genes were compared to 1000 random coding genes. Compare *C. elegans* data A–E to average species data A′–E′ plus F′ and G′. (L) $-\text{Log}_{10}$ (*P*-values) for each comparison test where cumulative site scores in transcription-factor-dependent genes are compared to scores in random coding genes. Compare *C. elegans*-derived cumulative site score (H and I) to averaged-species-derived cumulative sites scores (H′, I′, and J′).

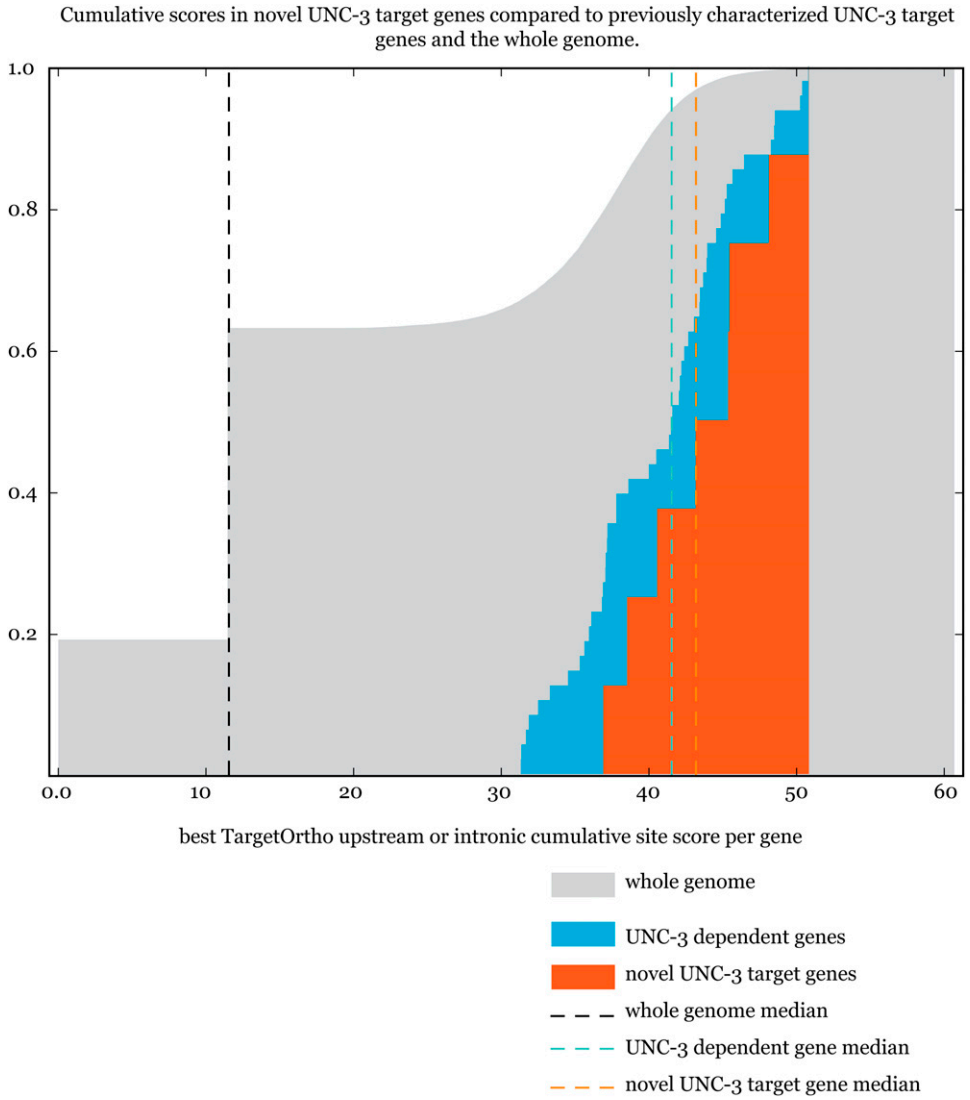


Figure 7 Cumulative site scores of novel *unc-3* target genes. CDF plot of best upstream or intronic cumulative site score per gene in novel *unc-3*-predicted target genes (blue) compared to the whole-genome distribution of upstream cumulative site scores (gray). The range of newly validated UNC-3 target gene cumulative site scores (orange) overlaps previously characterized *unc-3* target genes (blue). Sites for experimental validation were chosen before the final cumulative site score ranking scheme was finalized so that many putative target gene scores from the whole-genome sampling are higher than those from the validation set. While these results suggest that picking novel target genes that rank similarly to previously characterized TF target genes is a valid strategy, choosing candidates from the higher scoring end of the distribution may result in even better predictions.

GO enrichments include relevant TF target genes for further investigation: To demonstrate the utility of TargetOrtho predictions in finding biologically relevant target genes, GO enrichments among top-ranked target genes were assessed. GO analysis was performed on TargetOrtho's top-ranked sites per gene for whole-genome runs using upstream and intronic gene regions with the UNC-3 motif, AIY motif, and ASE motif using the GOrilla tool (Eden *et al.* 2007, 2009). The resulting ontologies among highly ranked predicted TF target genes show enrichments in neurogenesis pathway genes for all three terminal selector genes, providing ample candidates for further *in vivo* experimentation (Figure S8 and Table S12).

Validation of TargetOrtho through identification of novel UNC-3 target genes

For *in vivo* validation of TargetOrtho, 13 highly ranked potential UNC-3 target genes (Figure 7 and Table S4, gene list 7) were further investigated. Eight of these genes are completely uncharacterized while 5 have published expression patterns in the ventral nerve cord (VNC) where UNC-3 exerts its regulation as a terminal selector of cholinergic motor neurons. To examine whether these reporters are expressed in *unc-3*-expressing cells and are regulated by *unc-3*, we generated GFP promoter fusions for the 8 candidate target genes with no reported anatomical expression patterns (Figure 8). Transgenic lines expressing each of these reporters indeed show expression in VNC motor neurons (MNs), where UNC-3 is known to be expressed. Six of these reporter transgenes (C09G1.4, F08D12.1, F32B5.2, F47D12.3, C04E6.13, F57B7.2) were crossed into the *unc-3(e151)* mutant background, and each one of them showed significant loss ($P < 0.001$) of VNC neuron expression in the unc-3(e151) mutant, suggesting UNC-3 dependence (Figure 9; Table S1). We also crossed two (*hlh-32*, *F53E4*.1) of the five transgenes with previously described VNC MN expression into an *unc-3* mutant background and also found significant loss ($P < 0.001$) of VNC neuron expression, again suggesting UNC-3 dependence (Figure 9; Table S1). While these results confirm UNC-3 dependence, they do not distinguish direct

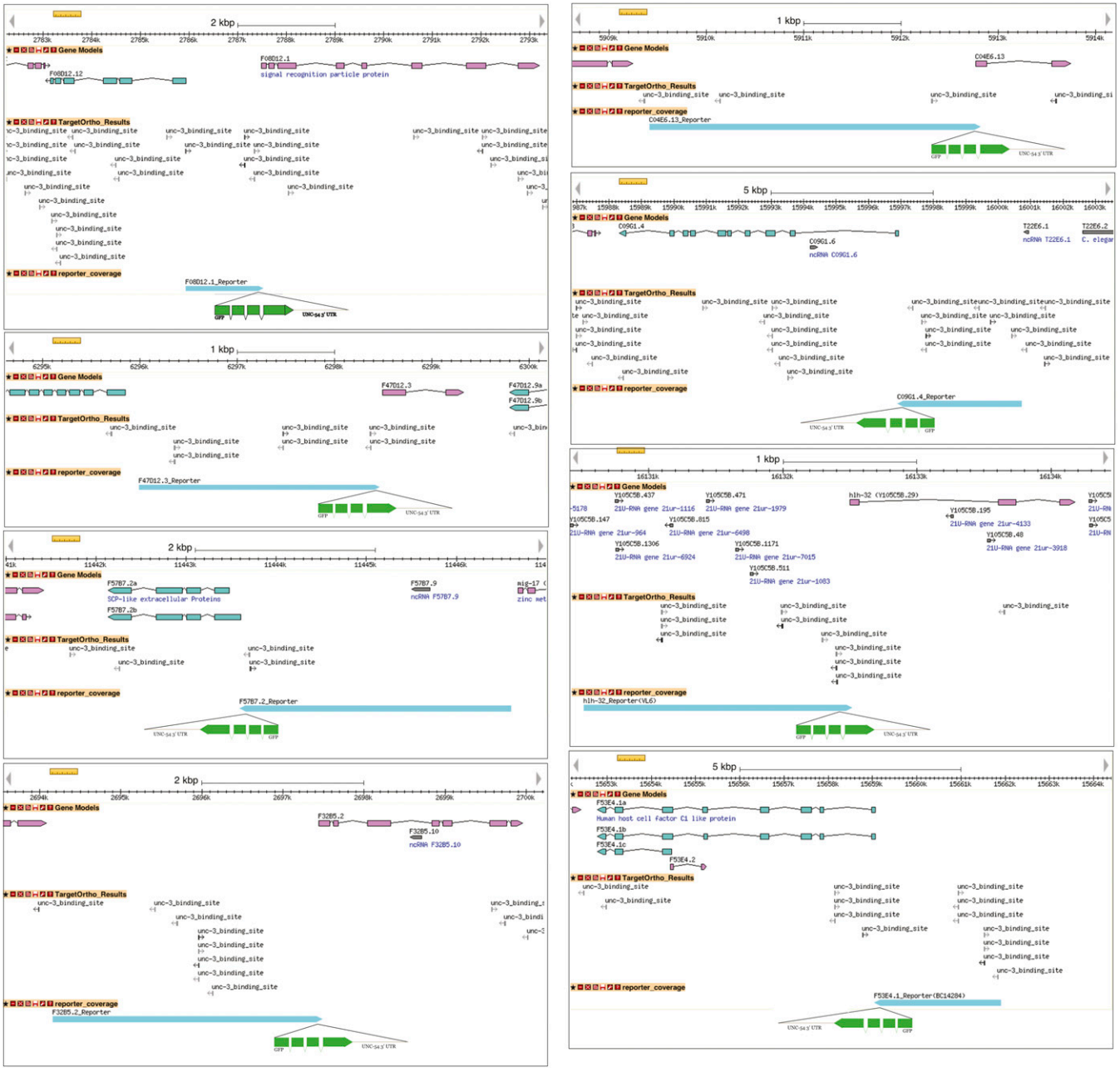
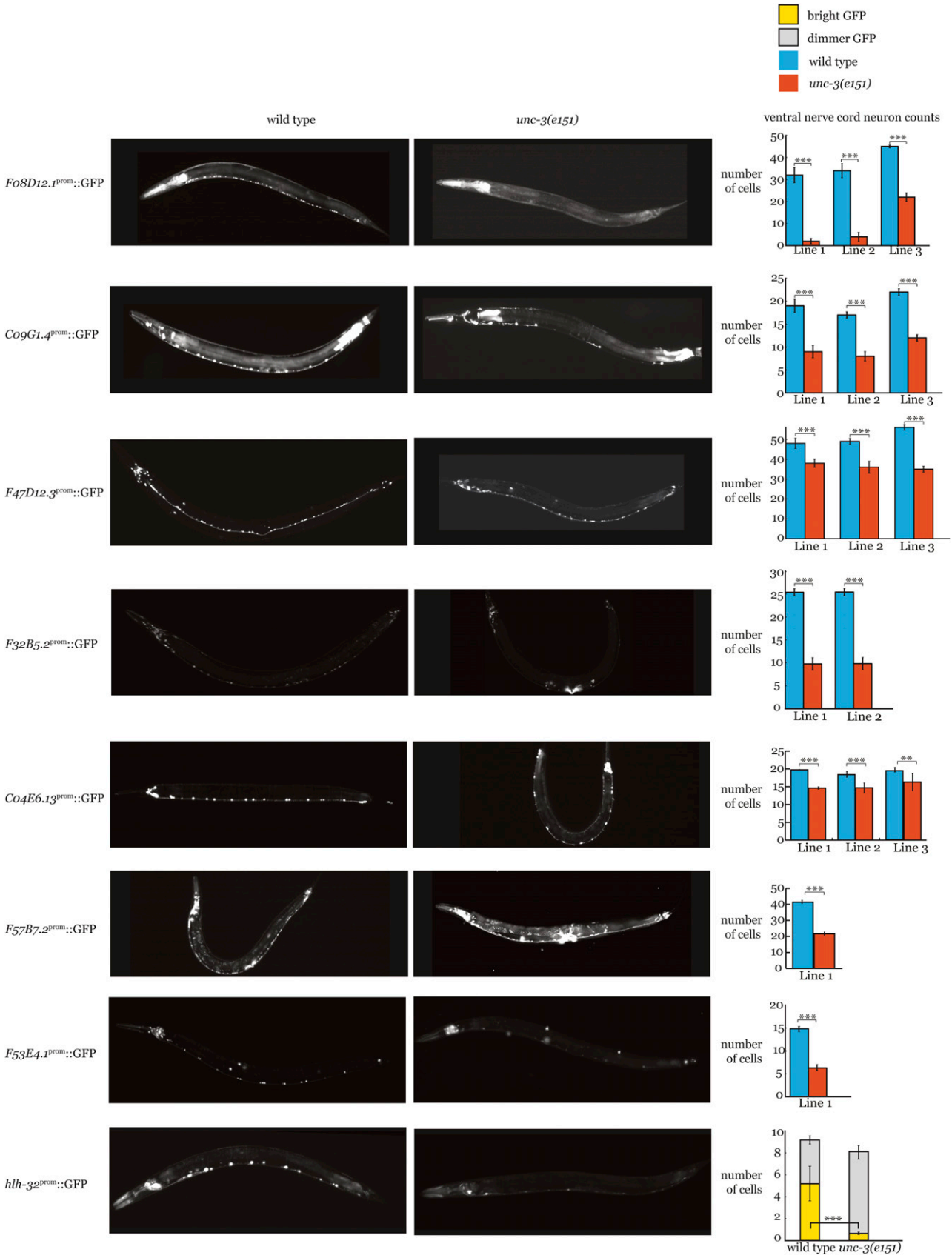


Figure 8 Gbrowse shots of novel *unc-3* target genes. TargetOrtho genome browser track files from an UNC-3 whole genome were uploaded to Wormbase's Gbrowse tool using the custom tracks option (TargetOrtho_results). This track shows each *unc-3* motif match as a shaded arrow. The direction of the arrow indicates the strand while the shading of the arrow corresponds to the strength of the motif match. Darker shading indicates higher log-likelihood motif match scores where the raw log-likelihood motif match score is scaled between 500 and 1000 using the maximum and minimum *C. elegans* (reference genome) scores from the TargetOrtho run. The reporter coverage track shows the coordinates of each GFP fusion reporter used for validation of TargetOrtho in wild-type and UNC-3 mutant animals.

UNC-3 regulation via binding to UNC-3 sites in each promoter from indirect regulation by downstream UNC-3 effectors. Deletion analysis of candidate UNC-3-binding sites in UNC-3-dependent genes is necessary to confirm direct UNC-3 regulation of the candidate target genes.

Utility of TargetOrtho in other species

The utility of TargetOrtho for identification of TF target genes is useful beyond *C. elegans*. To expand the functionality of TargetOrtho, we have implemented the pipeline for the *melanogaster* subgroup species with *D. melanogaster* as the reference genome. Numerous studies have utilized sequence conservation among closely related species to identify biologically functional elements. The relatively close phylogenetic distance between species in the *melanogaster* subgroup makes it amenable to conservation-based prediction of sequence function and suitable for target gene prediction with TargetOrtho.


bright GFP
dimmer GFP
wild type
unc-3(e151)
wild type
unc-3(e151)
ventral nerve cord neuron counts
F08D12.1prom::GFP
C09G1.4prom::GFP
F47D12.3prom::GFP
F32B5.2prom::GFP
C04E6.13prom::GFP
F57B7.2prom::GFP
F53E4.1prom::GFP
hlh-32prom::GFP
number of cells
Line 1
Line 2
Line 3
wild type
unc-3(e151)

The ASE motif used for TargetOrtho analysis in *C. elegans* is conserved in *D. melanogaster* and is bound by the *Drosophila* GLASS TF (Moses *et al.* 1989), the ortholog of CHE-1 in *C. elegans*. Two previously characterized GLASS-binding sites in Lz and Rh1 are highly ranked by TargetOrtho using the *melanogaster* species subgroup to comprise the five species genomes. Other CHE-1 target genes with ASE regulatory motifs are also conserved in *D. melanogaster* and are highly ranked by TargetOrtho (data not shown). The UNC-3 motif is also conserved in *D. melanogaster*, and preliminary analysis suggests that the *unc-17* ortholog, a validated UNC-3 target gene in *C. elegans*, is highly ranked by TargetOrtho in *D. melanogaster*. This trend is apparent in other UNC-3 target orthologs in *Drosophila* as well (data not shown). These preliminary results support a role for TargetOrtho target gene prediction in other species.

Discussion

We have demonstrated the predictive power of TargetOrtho using two approaches: bioinformatic validation of previously characterized TF-dependent genes compared to randomized coding genes and *in vivo* validation of novel TargetOrtho-predicted target genes. The bioinformatic validation supports a multi-species approach to candidate target gene prediction with averaged-species-derived TargetOrtho rankings showing the most discrimination between validated target genes and randomized genes. Similar trends were observed for PWM scans done on subsets of previously validated target genes not used to construct the PWM itself showing a conservative estimate of TargetOrtho's predictive power. The latter approach suggests that whole-genome PWM scans utilizing the multi-species ranking criteria results in novel target gene predictions that are strong with 6/6 scored reporter constructs showing expression in TF-expressing cells in which the expression displays TF dependence.

TargetOrtho provides an effective *in silico* approach for the identification of novel TF target genes. It offers a complementary approach to existing software that focuses mainly on *de novo* motif discovery by instead beginning with an experimentally validated motif and searching for conserved regulatory target genes. In this respect, TargetOrtho allows one to greatly expand the repertoire of TF target genes for a more complete understanding of the extensive regulatory networks controlled by TFs. TargetOrtho employs an alignment-independent method of conservation assignment necessary to accommodate the characteristic sequence degeneracy in TFBSs as well as motif repositioning within promoters due to sequence indels introduced over evolutionary time. The ability to overlay TargetOrtho-ranked results with other experimental data such as expression profiling, ChIP, or gene ontology data allows for additional layers of filtering to narrow down the best candidate target genes for further experimentation. In this respect, TargetOrtho serves as a powerful supplement to existing data.

The compactness of its genome and the often-observed proximity of *cis*-regulatory elements to their target genes make *C. elegans* particularly suited for TargetOrtho-based analysis of TF targets. However, increases in genome size and the sometimes very distal location of *cis*-regulatory control elements complicates target gene assignment in more complex metazoan species so that the utility of TargetOrtho may be limited. Another caveat of TargetOrtho use is that, while it has proved to work well for the three test cases presented here, its predictive power is expected to diminish with low-information-content motifs. A motif that occurs frequently in a given genome is likely to be conserved in orthologous genomes by chance alone, thus increasing the likelihood of false-positive target gene predictions. In cases where PWM information content is high, but the motif length is low (four to seven nucleotides), the same problem is expected.

Alternatively, true TF target genes may be ranked low if appropriate ortholog assignments have not been made. In these cases, TargetOrtho will underestimate the cumulative site score due to lack of nonreference genome species information. Often target genes with nonconserved sites may also be highly ranked due to strong reference genome results (such as motif count or log-likelihood site score). A second reference genome target gene may have identical rankings, but by averaging the ranking criteria across species, there is potential to lower the overall score even though clearly having even poor scoring sites in additional species is better than having no additional sites in additional species. Assuming that conservation increases the likelihood of biological functionality, one may choose to weight the conservation score (1–5) so that, despite underperforming averaged species data, the overall extent of conservation is considered. For our analysis of three well-characterized TFs, known TF target genes outperformed randomized coding genes despite this flaw. Additionally, weighting schemas may be explored for a given TargetOrtho run by adjusting the rank scaling parameters at run time. TargetOrtho results are also available as tab-delimited text so that the user may re-sort the data as appropriate. While adjustable input parameters allow flexibility in the ranking schema, users

Figure 9 Novel *unc-3*-predicted target genes validated *in vivo*. *unc-3* mutants show loss of reporter expression compared to wild-type worms in VNC motor neurons where UNC-3 is known to be a terminal selector of cholinergic motor neuron fate. (Left) Wild-type *C. elegans* worms. (Right) *unc-3(151)* worms. GFP fusions were injected into wild-type worms and then crossed into *unc-3(151)* for scoring. Bar charts show the VNC neuron counts for wild-type and *unc-3* mutant worms in all scored lines. All reporter constructs are complex extrachromosomal arrays except *hlh-32 (VL6)*.

must consider carefully the implications of tweaking individual ranking criteria. Such exploratory adjustments may result in user-biased predictions with the potential for an increase in the false discovery rate. To address this issue, we recommend running TargetOrtho using the query list option with a query list of previously characterized target genes so that ranking of these target genes may be assessed among whole-genome results.

In conclusion, TargetOrtho provides a cost- and time-efficient *in silico* approach for the identification of novel TF target genes, and, together with its CRM search function, is poised to unravel the regulatory logic of diverse biological processes.

Acknowledgments

We thank Q. Chen for expert assistance in generating transgenic strains, H. Bussemaker for valuable suggestions, and members of the Hobert lab for comments on the manuscript. This work was funded by the National Institutes of Health (R01NS039996-05; R01NS050266-03; 5T32DK007328-33). O.H. is an Investigator of the Howard Hughes Medical Institute.

Literature Cited

Aerts, S., 2012 Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr. Top. Dev. Biol. 98: 121–145.

Aerts, S., D. Lambrechts, S. Maity, P. Van Loo, B. Coessens *et al.*, 2006 Gene prioritization through genomic data fusion. Nat. Biotechnol. 24: 537–544.

Bailey, T. L., and C. Elkan, 1994 Fitting a mixture model by expectation maximization to discover motifs in biopolymers, pp. 38–36 in UCSD Technical Report CS94–351. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, August 1994.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37: W202–W208.

Bigelow, H. R., A. S. Wenick, A. Wong, and O. Hobert, 2004 CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. BMC Bioinformatics 5: 27.

Carey, M. F., C. L. Peterson, and S. T. Smale, 2009 Chromatin immunoprecipitation (ChIP). Cold Spring Harb. Protoc. 9: Prot5279

Eden, E., D. Lipson, S. Yogev, and Z. Yakhini, 2007 Discovering motifs in ranked lists of DNA sequences. PLOS Comput. Biol. 3: e39.

Eden, E., R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, 2009 GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48.

Elemento, O., and S. Tavazoie, 2005 Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. 6: R18.

Etchberger, J. F., A. Lorch, M. C. Sleumer, R. Zapf, S. J. Jones *et al.*, 2007 The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. Genes Dev. 21: 1653–1674.

Goecks, J., A. Nekrutenko, and J. Taylor, 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11: R86.

Gordân, R., L. Narlikar, and A. J. Hartemink, 2010 Finding regulatory DNA motifs using alignment-free evolutionary conservation information. Nucleic Acids Res. 38: e90.

Grant, C. E., T. L. Bailey, and W. S. Noble, 2011 FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018.

Hallikas,, O., K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen *et al.* 2006 Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124: 47–59.

Hellman, L. M., and M. G. Fried, 2007 Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat. Protoc. 2: 1849–1861.

Herrmann, C., B. Van de Sande, D. Potier, and S. Aerts, 2012 i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Res. 40: e114.

Hobert, O., 2002 PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. Biotechniques 32: 728–730.

Kratsios, P., A. Stolfi, M. Levine, and O. Hobert, 2012 Coordinated regulation of cholinergic motor neuron traits through a conserved terminal selector gene. Nat. Neurosci. 15: 205–214.

Moses, K., M. C. Ellis, and G. M. Rubin, 1989 The glass gene encodes a zinc-finger protein required by Drosophila photoreceptor cells. Nature 340(6234): 531–536.

Odenwald, W. F., W. Rasband, A. Kuzin, and T. Brody, 2005 EVOPRINTER, a multigenomic comparative tool for rapid identification of functionally important DNA. Proc. Natl. Acad. Sci. USA 102: 14700–14705.

Siddharthan, R., E. D. Siggia, and E. van Nimwegen, 2005 PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLOS Comput. Biol. 1: e67.

Sinha, S., M. Blanchette, and M. Tompa, 2005 PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics 5: 170.

Smedley, D., S. Haider, B. Ballester, R. Halland, D. London *et al.*, 2009 BioMart: biological queries made easy. BMC Genomics 10: 22.

Song, L., and G. E. Crawford, 2010 DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. 2010(2): pdb.prot5384.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100: 9440–9445.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin *et al.*, 2009 EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19: 327–335.

Wang, T., 2007 Using PhyloCon to identify conserved regulatory motifs. Curr. Protoc. Bioinformatics, Chapter 2, Unit 2.12.

Ward, L. D., and H. J. Bussemaker, 2008 Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics 24: i165–i171.

Wenick, A. S., and O. Hobert, 2004 Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans. Dev. Cell 6: 757–770.

Wright, W. E., M. Binder, and W. Funk, 1991 Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Mol. Cell. Biol. 11(8): 4104–4110.

Communicating editor: P. Sengupta

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160721/-/DC1

TargetOrtho: A Phylogenetic Footprinting Tool to Identify Transcription Factor Targets

Lori Glenwinkel, Di Wu, Gregory Minevich, and Oliver Hobert

DOI: 10.1534/genetics.113.160721

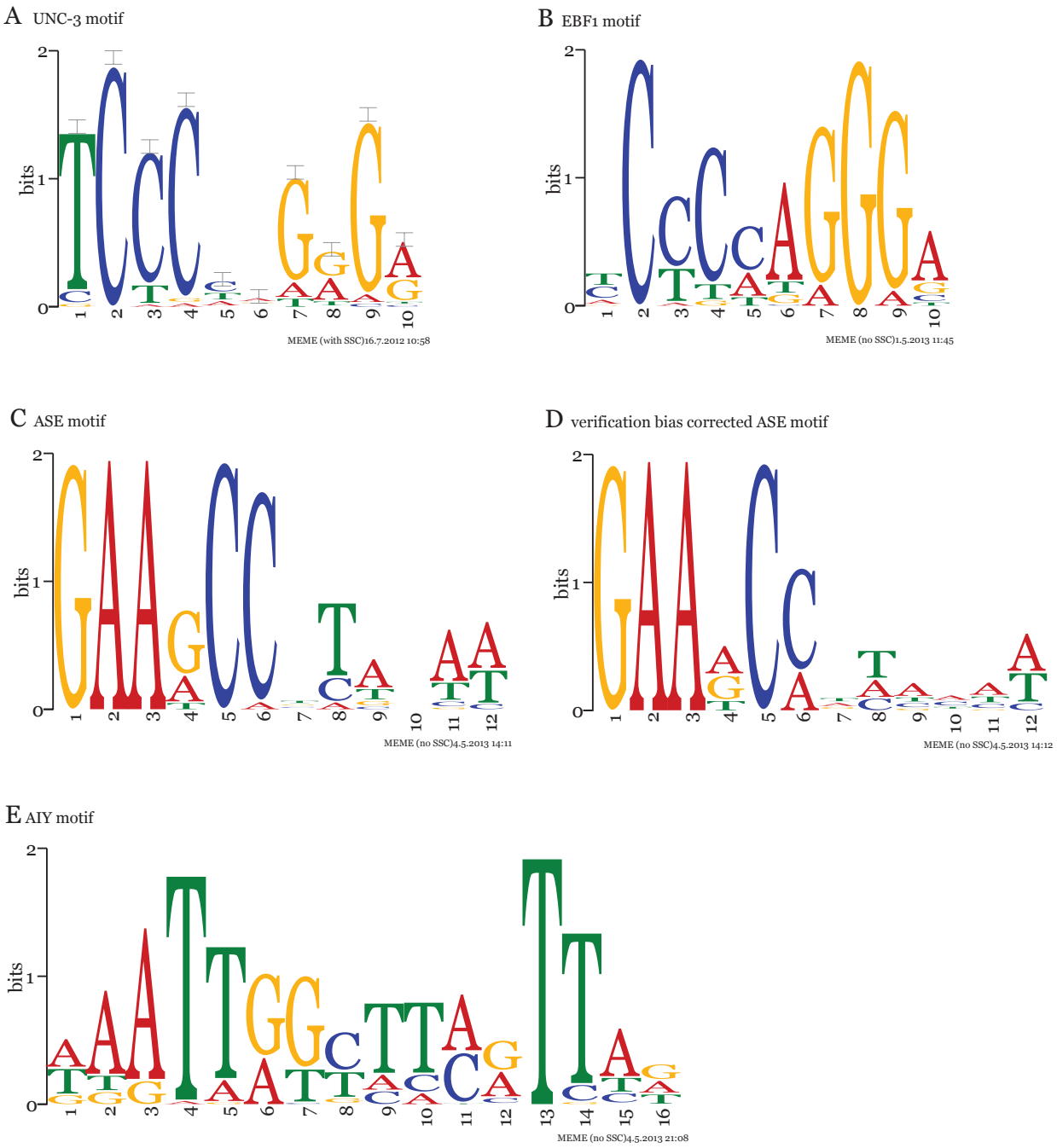


Figure S1

Figure S1 Motif logos. All logos were generated with Meme (Bailey *et al.* 2009) from upstream sequences of previously validated transcription factor dependent genes (Wenick *et al.* 2004, Kim *et al.* 2005, Etchberger *et al.* 2007, Kratsios *et al.* 2012) with background nucleotide frequencies were set to A: 0.325 C: 0.175 G: 0.175 T: 0.325 as determined from *C. elegans* upstream sequences. **A.** UNC-3 motif derived from previously characterized *unc-3* dependent target gene sequences (Kim *et al.* 2005, Kratsios *et al.* 2012). **B.** EBF1 motif derived from mouse DNA sequences from ChIP binding experiments (Treiber *et al.* 2010). **C.** ASE motif derived from upstream sequences of previously validated CHE-1 dependent genes(Etchberger *et al*. 2007). **D.** ASE (verification bias corrected) motif derived from upstream sequences of all CHE-1 dependent genes except those used to generate the ASE motif. **E.** AIY motif derived from ten upstream sequences of previously validated *ttx-3*/*ceh-10* dependent genes (Wenick *et al.* 2004).

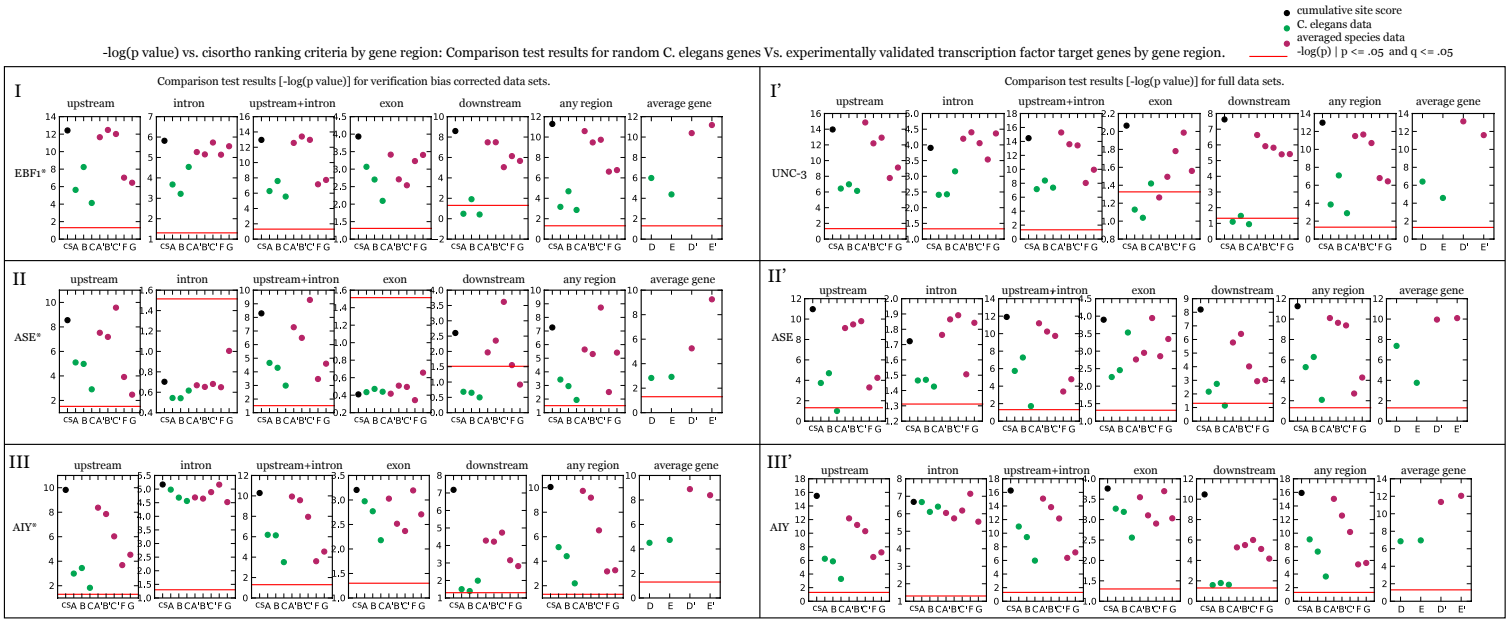


Figure S2

Figure S2 -log(P values) from comparison tests. TargetOrtho ranking criteria and cumulative site scores were compared between transcription factor dependent genes compared to 1000 random coding genes by gene region. Combining upstream and Intronic best motif match data per gene results in the most significant difference between comparison groups. **Part I.** TargetOrtho run with the EBF1 motif. Each TargetOrtho ranking criteria is compared among previously characterized *unc-3* dependent genes and 1000 random coding genes. Each ranking criteria P value is represented as the -log(P value) for each gene region (upstream, downstream, intron, exon, all regions) where each comparison group is composed of the best motif match value (for ranking criteria A,B,C,A',B',C',D,E,D',E',F,G described in the figure key) in a given region for each gene in the comparison group. Black dots represent the cumulative site score (cs) where the cumulative site score is derived from each averaged species ranking criteria in the given region (A'-C',F,G) and the averaged species total gene ranking criteria (D', E'). For example, part I upstream plot: The cumulative site score is derived from the best averaged species motif match score per gene for each ranking criteria A'-C', F, G and D',E' where A'-G are all determined from upstream ranking criteria and D',E' are averaged values across all gene regions (averaged species gene log-likelihood score and averaged species total gene site count). Green dots (A-C) represent the significance of the difference between comparison groups for upstream C. elegans ranking criteria while A'-C' represent the corresponding ranking criteria derived from averaged species data and F and G represent the conservation and offset variance criteria. Points above the red line are significant such that p<.05 and q < .05. Data shown in the intron, upstream + intron, exon, and downstream plots are as described for the upstream plot. The first 'all' plot represents -log(P values) derived from taking the best motif match value from any gene region for comparison tests. The final all plot show the significance of the total gene data comparisons using either C. elegans total gene score averaged log-likelihood score (D) and the total gene site count (E) across all gene regions or the averaged species data corresponding to D and E (D', E'). **Part I'**. -log(p values) from comparison tests of *unc-3* dependent genes compared to 1000 random coding genes for TargetOrtho run with the UNC-3 motif. **Part II.** -log(P values) from comparison tests of CHE-1 dependent genes compared to 1000 random coding genes for TargetOrtho run with the ASE motif. **Part II'.** -log(P values) from comparison tests of CHE-1 dependent genes (except those used to construct ASE-2 motif) compared to 1000 random coding genes for TargetOrtho run with the ASE-2 motif. **Part III.** -log(P values) from comparison tests of ttx-3/ceh-10 dependent genes compared to 1000 random coding genes for TargetOrtho run with the AIY motif. **Part III'** -log(P values) from comparison tests of ttx-3/ceh-10 dependent genes (except those used to generate the AIY motif compared to 1000 random coding genes for TargetOrtho run with the AIY motif.

UNC-3 known target gene motif match data Vs. random coding gene motif match data (EBF1 motif) from upstream and intronic regions

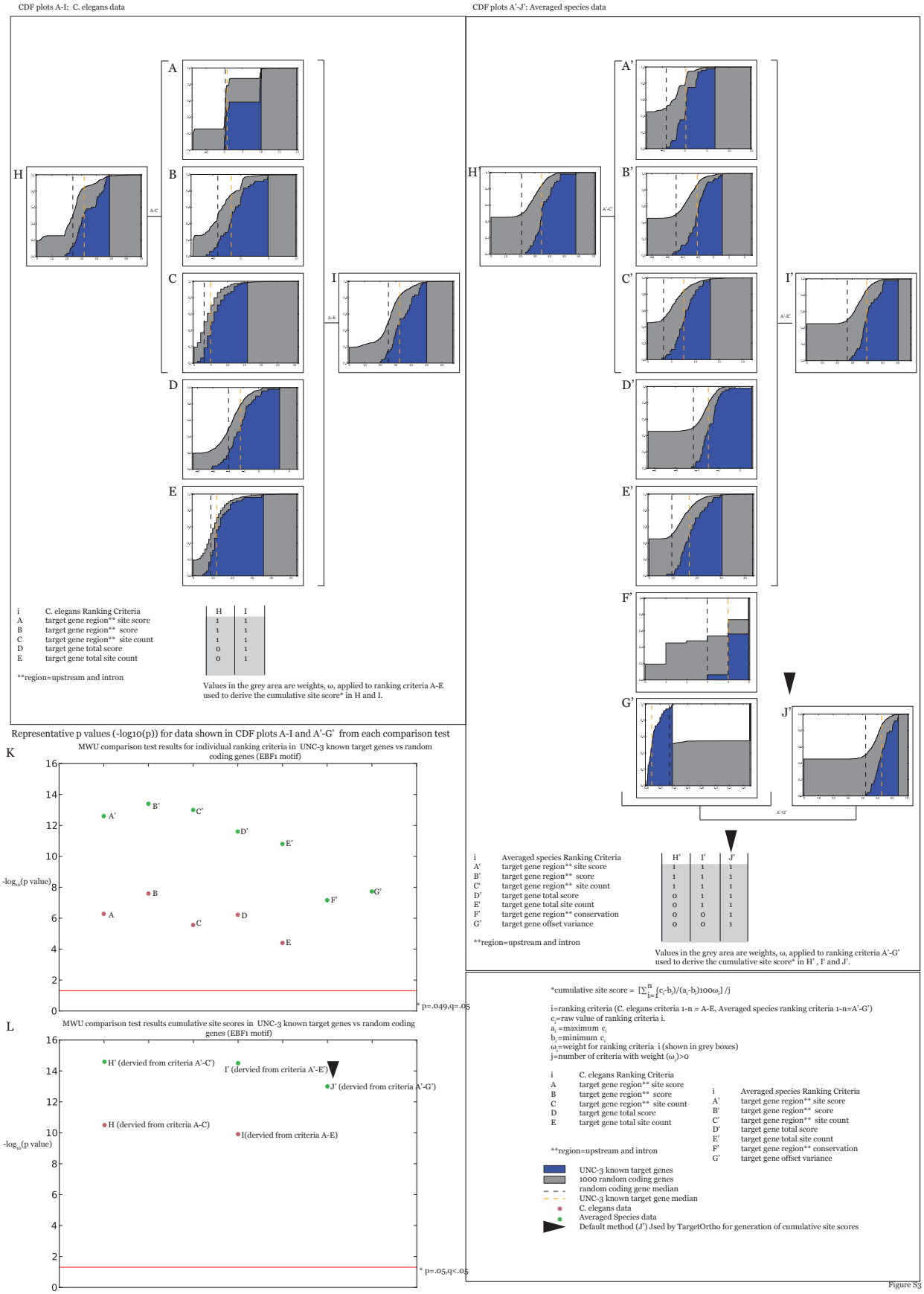
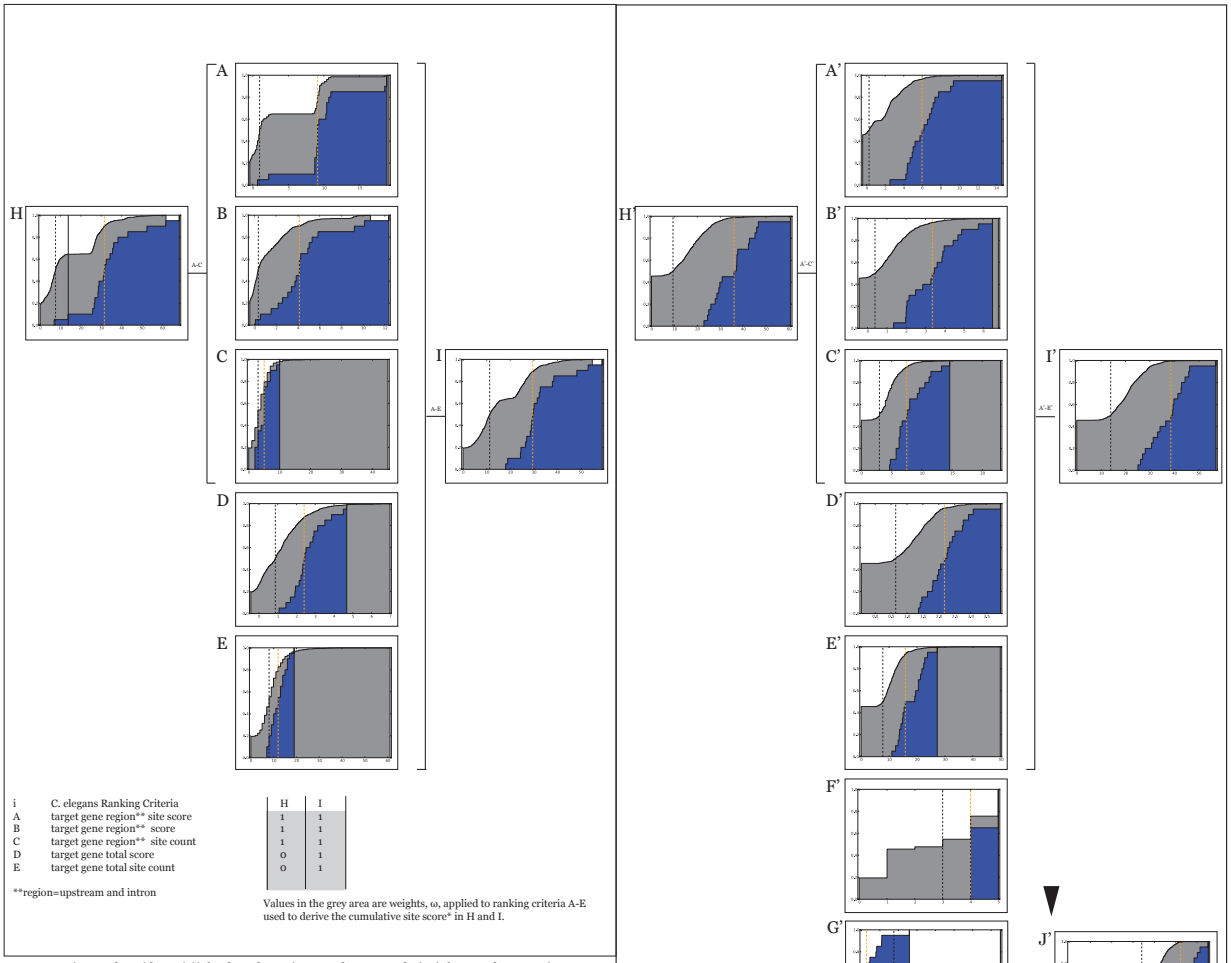


Figure S3

Figure S3 EBF1 motif analysis for verification bias correction of UNC-3 analysis-Unc-3 dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized *unc-3* dependent genes and 1000 random coding genes were submitted to TargetOrtho using the EBF1 motif as input (Figure S1B). Data distributions for each TargetOrtho ranking criteria were compared between known target genes and random coding genes.
CDF plots of individual ranking criteria (plots A-E and plots A'-G'): CDF plots are shown for individual ranking criteria A-E and A'-G'. TargetOrtho ranking criteria derived from averaged species data (A'-G') better distinguish previously validated TF target genes from random genes compared to using *C. elegans* (reference genome) data alone (A-E). **CDF plots A-E** show ranking criteria derived from *C. elegans* genome data only while **CDF plots A'-E**' show the corresponding ranking criteria derived from averaged species data. **CDF plot F' and G'** show averaged species data having no reference genome counterpart including the conservation and offset variance data distributions.
CDF plots of cumulative site scores (plots H, I and plots H', I', J'): Data distributions for cumulative site scores derived from unique combinations of TargetOrtho ranking criteria are shown in CDF plots H,I,H',I',J'. **CDF plot H** shows the cumulative site score distributions derived from *C. elegans* upstream and intronic data only calculated from A-C. The left panel, plots A'-C' shows the cumulative site score CDF plots calculated from the corresponding averaged species upstream and intronic data. **CDF plot I** shows cumulative site scores derived from criteria shown in CDF plots A-E where **CDF plots D and E** represent total gene ranking criteria in *C. elegans* only (**D.** *C. elegans* averaged upstream and intronic site scores and **E.** *C. elegans* averaged site score across all gene regions). **CDF plot I'** (left panel) shows the data distribution of cumulative site scores derived from A'-E' where **CDF plots D' and E'** represent the corresponding total gene ranking criteria averaged across species. **CDF plot J'** shows cumulative site scores derived from all averaged species ranking criteria (A'-G').
K. $-\log_{10}$(P value) for each ranking criteria comparison test where transcription factor dependent genes were compared to 1000 random coding genes. Compare *C. elegans* data A-E to average species data A'-E' plus F' and G'.
L. –$\log_{10}$(P values) for each comparison test where cumulative sites scores in transcription factor dependent genes are compared to scores in random coding genes. Compare *C. elegans* derived cumulative site score (H and I) to averaged species derived cumulative sites scores (H', I', and J').

CDF plots A-I: C. elegans data

CDF plots A'-J': Averaged species data



i	C. elegans Ranking Criteria
A	target gene region** site score
B	target gene region** score
C	target gene region** site count
D	target gene total score
E	target gene total site count

**region=upstream and intron

H	I
1	1
1	1
1	1
0	1
0	1

Values in the grey area are weights, ω, applied to ranking criteria A-E used to derive the cumulative site score* in H and I.

i	Averaged species Ranking Criteria
A'	target gene region** site score
B'	target gene region** score
C'	target gene region** site count
D'	target gene total score
E'	target gene total site count
F'	target gene region** conservation
G'	target gene offset variance

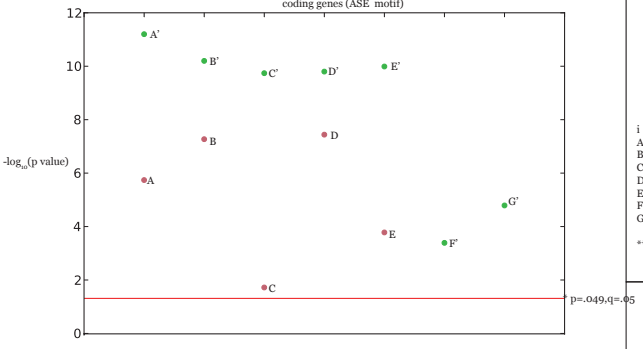
**region=upstream and intron

H'	I'	J'
1	1	1
1	1	1
1	1	1
0	1	1
0	1	1
0	0	1
0	0	1

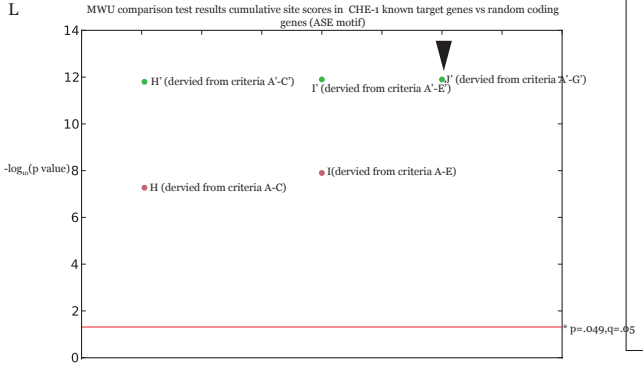
Values in the grey area are weights, ω, applied to ranking criteria A'-G' used to derive the cumulative site score* in H' , I' and J'.

Representative p values (-log10(p)) for data shown in CDF plots A-I and A'-G' from each comparison test

K



L



*cumulative site score = $[\sum_{i=1}^{n}(c_i-b_i)/(a_i-b_i)100\omega_i]\ /j$

i=ranking criteria (C. elegans criteria 1-n = A-E, Averaged species ranking criteria 1-n=A'-G')
c_i=raw value of ranking criteria i.
a_i =maximum c_i
b_i=minimum c_i
ω_i=weight for ranking criteria i (shown in grey boxes)
j=number of criteria with weight (ω_i)>0

i	C. elegans Ranking Criteria
A	target gene region** site score
B	target gene region** score
C	target gene region** site count
D	target gene total score
E	target gene total site count

**region=upstream and intron

i	Averaged species Ranking Criteria
A'	target gene region** site score
B'	target gene region** score
C'	target gene region** site count
D'	target gene total score
E'	target gene total site count
F'	target gene region** conservation
G'	target gene offset variance

CHE-1known target genes
1000 random coding genes
random coding gene median
CHE-1 known target gene median
C. elegans data
Averaged Species data
Default method (J') Jsed by TargetOrtho for generation of cumulative site scores

Figure S4

Figure S4 ASE motif analysis. *che-1* dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized *che-1* dependent genes and 1000 random coding genes were submitted to TargetOrtho using the ASE motif as input (Figure S1C). Data distributions for each TargetOrtho ranking criteria were compared between known target genes and random coding genes.

CDF plots of individual ranking criteria (plots A-E and plots A'-G'): CDF plots are shown for individual ranking criteria A-E and A'-G'. TargetOrtho ranking criteria derived from averaged species data (A'-G') better distinguish previously validated TF target genes from random genes compared to using *C. elegans* (reference genome) data alone (A-E). **CDF plots A-E** show ranking criteria derived from *C. elegans* genome data only while **CDF plots A'-E**' show the corresponding ranking criteria derived from averaged species data. **CDF plot F' and G'** show averaged species data having no reference genome counterpart including the conservation and offset variance data distributions.

CDF plots of cumulative site scores (plots H, I and plots H', I', J'): Data distributions for cumulative site scores derived from unique combinations of TargetOrtho ranking criteria are shown in CDF plots H,I,H',I',J'. **CDF plot H** shows the cumulative site score distributions derived from *C. elegans* upstream and intronic data only calculated from A-C. The left panel, plots A'-C' shows the cumulative site score CDF plots calculated from the corresponding averaged species upstream and intronic data. **CDF plot I** shows cumulative site scores derived from criteria shown in CDF plots A-E where **CDF plots D and E** represent total gene ranking criteria in *C. elegans* only (**D.** *C. elegans* averaged upstream and intronic site scores and **E.** *C. elegans* averaged site score across all gene regions). **CDF plot I'** (left panel) shows the data distribution of cumulative site scores derived from A'-E' where **CDF plots D' and E'** represent the corresponding total gene ranking criteria averaged across species. **CDF plot J'** shows cumulative site scores derived from all averaged species ranking criteria (A'-G').

K. $-\log_{10}$(P value) for each ranking criteria comparison test where transcription factor dependent genes were compared to 1000 random coding genes. Compare *C. elegans* data A-E to average species data A'-E' plus F' and G'.

L. –$\log_{10}$(P values) for each comparison test where cumulative sites scores in transcription factor dependent genes are compared to scores in random coding genes. Compare *C. elegans* derived cumulative site score (H and I) to averaged species derived cumulative sites scores (H', I', and J').

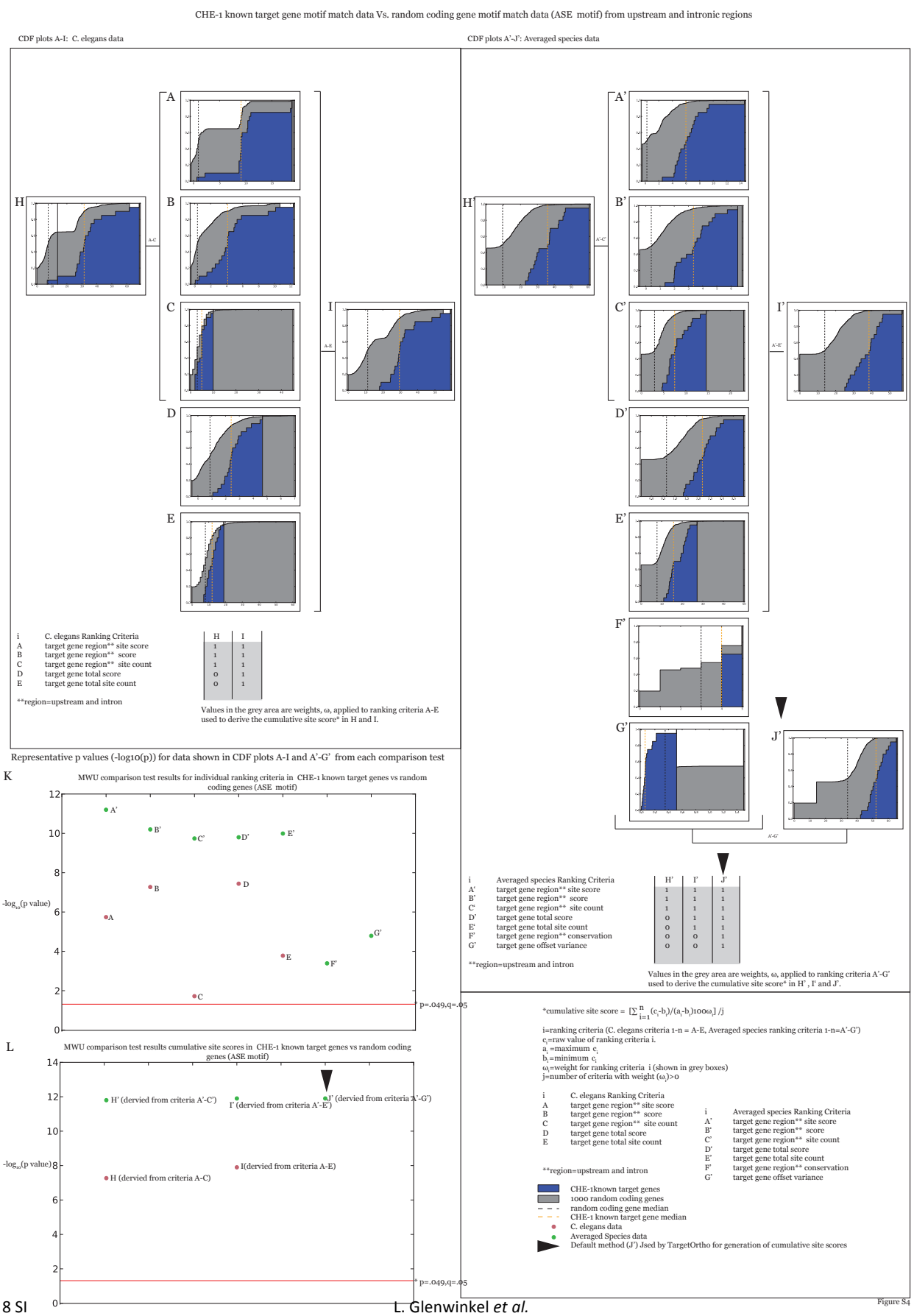

CHE-1 known target gene motif match data Vs. random coding gene motif match data (ASE motif) from upstream and intronic regions
CDF plots A-I: C. elegans data
CDF plots A'-J': Averaged species data
A
B
C
D
E
H
I
A'
B'
C'
D'
E'
F'
G'
H'
I'
J'
A-C
A-E
A'-C'
A'-E'
A'-G'
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
**region=upstream and intron
H I
1 1
1 1
1 1
0 1
0 1
Values in the grey area are weights, ω, applied to ranking criteria A-E used to derive the cumulative site score* in H and I.
Representative p values (-log10(p)) for data shown in CDF plots A-I and A'-G' from each comparison test
K
MWU comparison test results for individual ranking criteria in CHE-1 known target genes vs random coding genes (ASE motif)
-log10(p value)
p=.049,q=.05
L
MWU comparison test results cumulative site scores in CHE-1 known target genes vs random coding genes (ASE motif)
H' (dervied from criteria A'-C')
I' (dervied from criteria A'-E')
J' (dervied from criteria A'-G')
I(dervied from criteria A-E)
H (dervied from criteria A-C)
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
**region=upstream and intron
H' I' J'
1 1 1
1 1 1
1 1 1
0 1 1
0 1 1
0 0 1
0 0 1
Values in the grey area are weights, ω, applied to ranking criteria A'-G' used to derive the cumulative site score* in H' , I' and J'.
*cumulative site score = [∑ n i=1 (ci-bi)/(ai-bi)100ωi] /j
i=ranking criteria (C. elegans criteria 1-n = A-E, Averaged species ranking criteria 1-n=A'-G')
ci=raw value of ranking criteria i.
ai =maximum ci
bi =minimum ci
ωi=weight for ranking criteria i (shown in grey boxes)
j=number of criteria with weight (ωi)>0
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
**region=upstream and intron
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
CHE-1known target genes
1000 random coding genes
random coding gene median
CHE-1 known target gene median
C. elegans data
Averaged Species data
Default method (J') Jsed by TargetOrtho for generation of cumulative site scores

Figure S4

Figure S5 ASE motif analysis with verification bias correction. *che-1* dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized *che-1* dependent genes (except those used to construct the bias corrected ASE motif) and 1000 random coding genes were submitted to TargetOrtho using the bias corrected ASE motif as input (Figure S1D). Data distributions for each TargetOrtho ranking criteria were compared between known target genes and random coding genes.

CDF plots of individual ranking criteria (plots A-E and plots A'-G'): CDF plots are shown for individual ranking criteria A-E and A'-G'. TargetOrtho ranking criteria derived from averaged species data (A'-G') better distinguish previously validated TF target genes from random genes compared to using *C. elegans* (reference genome) data alone (A-E). **CDF plots A-E** show ranking criteria derived from *C. elegans* genome data only while **CDF plots A'-E**' show the corresponding ranking criteria derived from averaged species data. **CDF plot F' and G'** show averaged species data having no reference genome counterpart including the conservation and offset variance data distributions.

CDF plots of cumulative site scores (plots H, I and plots H', I', J'): Data distributions for cumulative site scores derived from unique combinations of TargetOrtho ranking criteria are shown in CDF plots H,I,H',I',J'. **CDF plot H** shows the cumulative site score distributions derived from *C. elegans* upstream and intronic data only calculated from A-C. The left panel, plots A'-C' shows the cumulative site score CDF plots calculated from the corresponding averaged species upstream and intronic data. **CDF plot I** shows cumulative site scores derived from criteria shown in CDF plots A-E where **CDF plots D and E** represent total gene ranking criteria in *C. elegans* only (**D.** *C. elegans* averaged upstream and intronic site scores and **E.** *C. elegans* averaged site score across all gene regions). **CDF plot I'** (left panel) shows the data distribution of cumulative site scores derived from A'-E' where **CDF plots D' and E'** represent the corresponding total gene ranking criteria averaged across species. **CDF plot J'** shows cumulative site scores derived from all averaged species ranking criteria (A'-G').

K. $-\log_{10}$(P value) for each ranking criteria comparison test where transcription factor dependent genes were compared to 1000 random coding genes. Compare *C. elegans* data A-E to average species data A'-E' plus F' and G'.

L. $-\log_{10}$(P values) for each comparison test where cumulative sites scores in transcription factor dependent genes are compared to scores in random coding genes. Compare *C. elegans* derived cumulative site score (H and I) to averaged species derived cumulative sites scores (H' , I', and J').

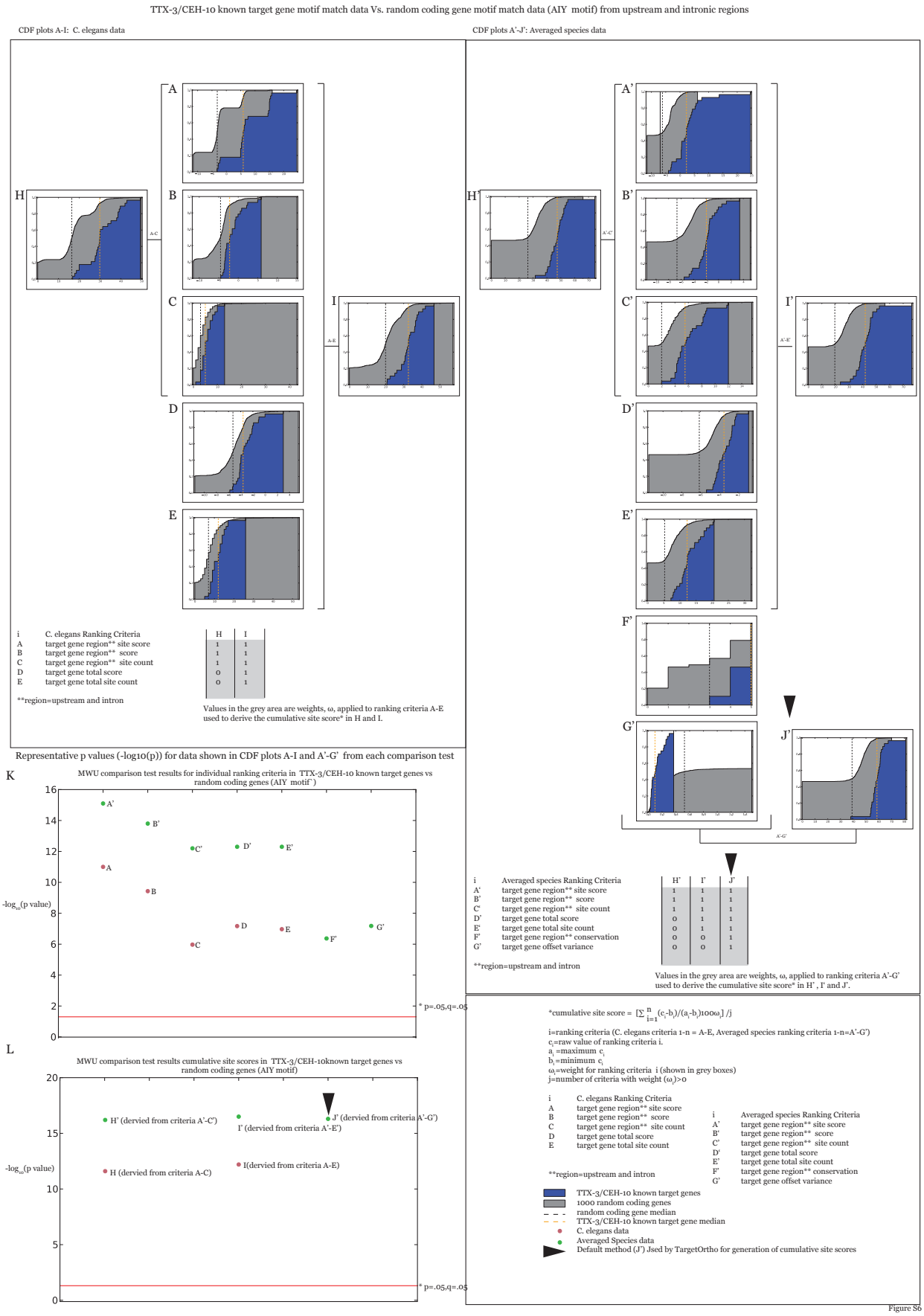

TTX-3/CEH-10 known target gene motif match data Vs. random coding gene motif match data (AIY motif) from upstream and intronic regions
CDF plots A-I: C. elegans data
CDF plots A'-J': Averaged species data
H
A
B
C
D
E
I
A-C
A-E
H'
A'
B'
C'
D'
E'
F'
G'
I'
J'
A'-C'
A'-E'
A'-G'
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
**region=upstream and intron
H I
1 1
1 1
1 1
0 1
0 1
Values in the grey area are weights, ω, applied to ranking criteria A-E used to derive the cumulative site score* in H and I.
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
**region=upstream and intron
H' I' J'
1 1 1
1 1 1
1 1 1
0 1 1
0 1 1
0 0 1
0 0 1
Values in the grey area are weights, ω, applied to ranking criteria A'-G' used to derive the cumulative site score* in H' , I' and J'.
Representative p values (-log10(p)) for data shown in CDF plots A-I and A'-G' from each comparison test
K
MWU comparison test results for individual ranking criteria in TTX-3/CEH-10 known target genes vs random coding genes (AIY motif)
-log10(p value)
* p=.05,q=.05
L
MWU comparison test results cumulative site scores in TTX-3/CEH-10known target genes vs random coding genes (AIY motif)
H' (dervied from criteria A'-C')
I' (dervied from criteria A'-E')
J' (dervied from criteria A'-G')
H (dervied from criteria A-C)
I(dervied from criteria A-E)
*cumulative site score = [∑ n i=1 (ci-bi)/(ai-bi)100ωi] /j
i=ranking criteria (C. elegans criteria 1-n = A-E, Averaged species ranking criteria 1-n=A'-G')
ci=raw value of ranking criteria i.
ai =maximum ci
bi =minimum ci
ωi=weight for ranking criteria i (shown in grey boxes)
j=number of criteria with weight (ωi)>0
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
**region=upstream and intron
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
TTX-3/CEH-10 known target genes
1000 random coding genes
random coding gene median
TTX-3/CEH-10 known target gene median
C. elegans data
Averaged Species data
Default method (J') Jsed by TargetOrtho for generation of cumulative site scores

Figure S6

Figure S6 AIY motif analysis. *ceh-10/ttx-3* dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized *ceh-10/ttx-3* dependent genes and 1000 random coding genes were submitted to TargetOrtho using the ASE motif as input (Figure S1E). Data distributions for each TargetOrtho ranking criteria were compared between known target genes and random coding genes.

CDF plots of individual ranking criteria (plots A-E and plots A'-G'): CDF plots are shown for individual ranking criteria A-E and A'-G'. TargetOrtho ranking criteria derived from averaged species data (A'-G') better distinguish previously validated TF target genes from random genes compared to using *C. elegans* (reference genome) data alone (A-E). **CDF plots A-E** show ranking criteria derived from *C. elegans* genome data only while **CDF plots A'-E**' show the corresponding ranking criteria derived from averaged species data. **CDF plot F' and G'** show averaged species data having no reference genome counterpart including the conservation and offset variance data distributions.

CDF plots of cumulative site scores (plots H, I and plots H', I', J'): Data distributions for cumulative site scores derived from unique combinations of TargetOrtho ranking criteria are shown in CDF plots H,I,H',I',J'. **CDF plot H** shows the cumulative site score distributions derived from *C. elegans* upstream and intronic data only calculated from A-C. The left panel, plots A'-C' shows the cumulative site score CDF plots calculated from the corresponding averaged species upstream and intronic data. **CDF plot I** shows cumulative site scores derived from criteria shown in CDF plots A-E where **CDF plots D and E** represent total gene ranking criteria in *C. elegans* only (**D.** *C. elegans* averaged upstream and intronic site scores and **E.** *C. elegans* averaged site score across all gene regions). **CDF plot I'** (left panel) shows the data distribution of cumulative site scores derived from A'-E' where **CDF plots D' and E'** represent the corresponding total gene ranking criteria averaged across species. **CDF plot J'** shows cumulative site scores derived from all averaged species ranking criteria (A'-G').

K. $-\log_{10}$(P value) for each ranking criteria comparison test where transcription factor dependent genes were compared to 1000 random coding genes. Compare *C. elegans* data A-E to average species data A'-E' plus F' and G'.

L. –$\log_{10}$(P values) for each comparison test where cumulative sites scores in transcription factor dependent genes are compared to scores in random coding genes. Compare *C. elegans* derived cumulative site score (H and I) to averaged species derived cumulative sites scores (H' , I', and J').

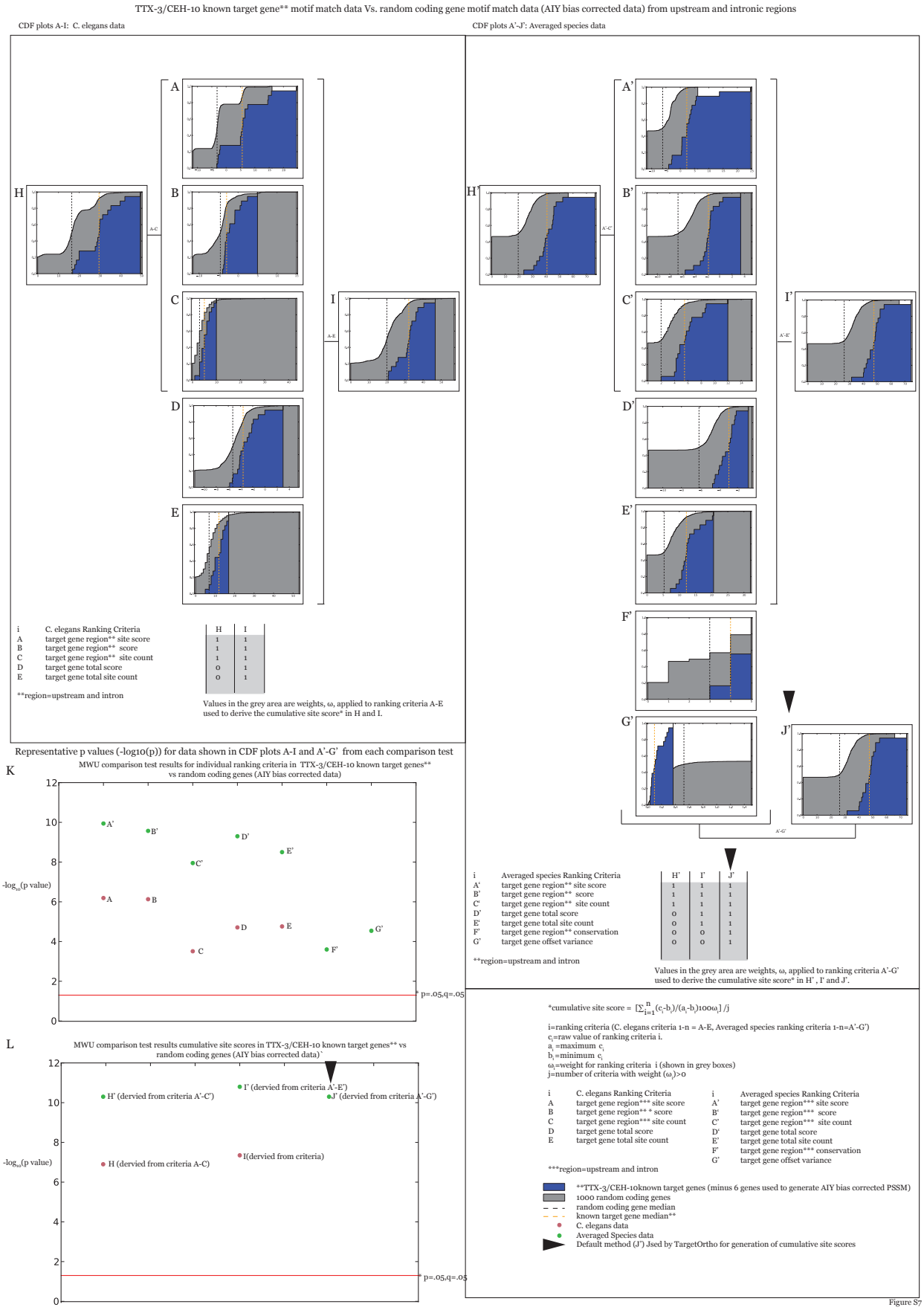

TTX-3/CEH-10 known target gene** motif match data Vs. random coding gene motif match data (AIY bias corrected data) from upstream and intronic regions
CDF plots A-I: C. elegans data
CDF plots A'-J': Averaged species data
A
B
C
D
E
H
I
A'
B'
C'
D'
E'
F'
G'
H'
I'
J'
A-C
A-E
A'-C'
A'-E'
A'-G'
i C. elegans Ranking Criteria
A target gene region** site score
B target gene region** score
C target gene region** site count
D target gene total score
E target gene total site count
**region=upstream and intron
H I
1 1
1 1
1 1
0 1
0 1
Values in the grey area are weights, ω, applied to ranking criteria A-E used to derive the cumulative site score* in H and I.
i Averaged species Ranking Criteria
A' target gene region** site score
B' target gene region** score
C' target gene region** site count
D' target gene total score
E' target gene total site count
F' target gene region** conservation
G' target gene offset variance
**region=upstream and intron
H' I' J'
1 1 1
1 1 1
1 1 1
0 1 1
0 1 1
0 0 1
0 0 1
Values in the grey area are weights, ω, applied to ranking criteria A'-G' used to derive the cumulative site score* in H' , I' and J'.
*cumulative site score = [Σ_{i=1}^{n} (c_i-b_i)/(a_i-b_i)100ω_i] /j
i=ranking criteria (C. elegans criteria 1-n = A-E, Averaged species ranking criteria 1-n=A'-G')
c_i=raw value of ranking criteria i.
a_i=maximum c_i
b_i=minimum c_i
ω_i=weight for ranking criteria i (shown in grey boxes)
j=number of criteria with weight (ω_i)>0
i C. elegans Ranking Criteria
A target gene region*** site score
B target gene region** * score
C target gene region*** site count
D target gene total score
E target gene total site count
i Averaged species Ranking Criteria
A' target gene region*** site score
B' target gene region*** score
C' target gene region*** site count
D' target gene total score
E' target gene total site count
F' target gene region*** conservation
G' target gene offset variance
***region=upstream and intron
**TTX-3/CEH-10known target genes (minus 6 genes used to generate AIY bias corrected PSSM)
1000 random coding genes
random coding gene median
known target gene median**
C. elegans data
Averaged Species data
Default method (J') Jsed by TargetOrtho for generation of cumulative site scores
Representative p values (-log10(p)) for data shown in CDF plots A-I and A'-G' from each comparison test
K
MWU comparison test results for individual ranking criteria in TTX-3/CEH-10 known target genes** vs random coding genes (AIY bias corrected data)
-log10(p value)
* p=.05,q=.05
L
MWU comparison test results cumulative site scores in TTX-3/CEH-10 known target genes** vs random coding genes (AIY bias corrected data)`
H' (dervied from criteria A'-C')
I' (dervied from criteria A'-E')
J' (dervied from criteria A'-G')
H (dervied from criteria A-C)
I(dervied from criteria)
-log10(p value)
* p=.05,q=.05

Figure S7

Figure S7 AIY motif analysis with verification bias corrected data. *ceh-10/ttx-3* dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized *ceh-10/ttx-3* dependent genes (except those used to construct the AIY motif) and 1000 random coding genes were submitted to TargetOrtho using the ASE motif as input (Figure S1E). Data distributions for each TargetOrtho ranking criteria were compared between known target genes and random coding genes.

CDF plots of individual ranking criteria (plots A-E and plots A'-G'): CDF plots are shown for individual ranking criteria A-E and A'-G'. TargetOrtho ranking criteria derived from averaged species data (A'-G') better distinguish previously validated TF target genes from random genes compared to using *C. elegans* (reference genome) data alone (A-E). **CDF plots A-E** show ranking criteria derived from *C. elegans* genome data only while **CDF plots A'-E**' show the corresponding ranking criteria derived from averaged species data. **CDF plot F' and G'** show averaged species data having no reference genome counterpart including the conservation and offset variance data distributions.

CDF plots of cumulative site scores (plots H, I and plots H', I', J'): Data distributions for cumulative site scores derived from unique combinations of TargetOrtho ranking criteria are shown in CDF plots H,I,H',I',J'. **CDF plot H** shows the cumulative site score distributions derived from *C. elegans* upstream and intronic data only calculated from A-C. The left panel, plots A'-C' shows the cumulative site score CDF plots calculated from the corresponding averaged species upstream and intronic data. **CDF plot I** shows cumulative site scores derived from criteria shown in CDF plots A-E where **CDF plots D and E** represent total gene ranking criteria in *C. elegans* only (**D.** *C. elegans* averaged upstream and intronic site scores and **E.** *C. elegans* averaged site score across all gene regions). **CDF plot I'** (left panel) shows the data distribution of cumulative site scores derived from A'-E' where **CDF plots D' and E'** represent the corresponding total gene ranking criteria averaged across species. **CDF plot J'** shows cumulative site scores derived from all averaged species ranking criteria (A'-G').

K. -$\log_{10}$(P value) for each ranking criteria comparison test where transcription factor dependent genes were compared to 1000 random coding genes. Compare *C. elegans* data A-E to average species data A'-E' plus F' and G'.

L. –$\log_{10}$(P values) for each comparison test where cumulative sites scores in transcription factor dependent genes are compared to scores in random coding genes. Compare *C. elegans* derived cumulative site score (H and I) to averaged species derived cumulative sites scores (H', I', and J').

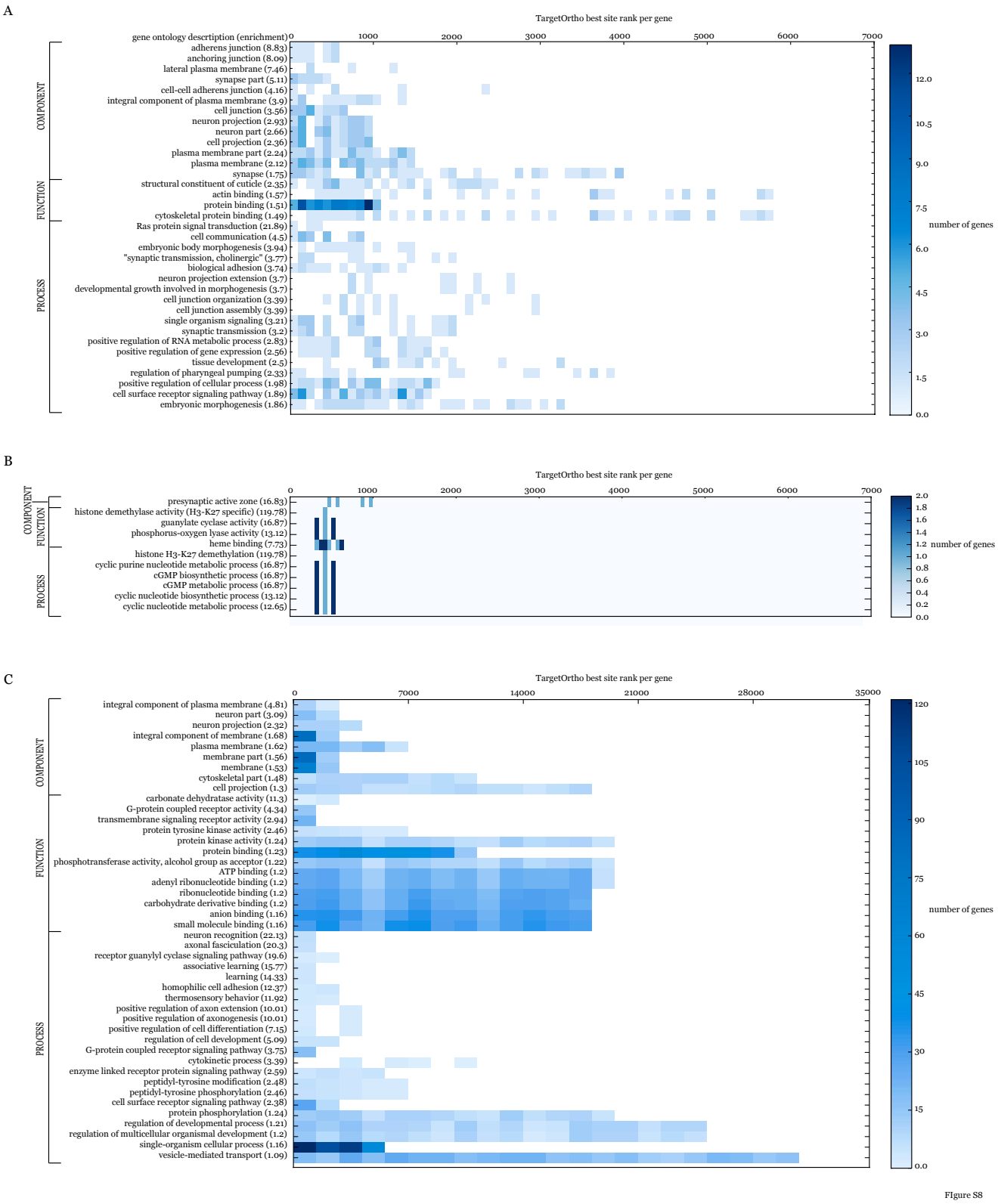


Figure S8 Heatmaps of gene ontology results from Gorilla analysis. A. Gene ontology enrichments of UNC-3 candidate target genes in the top ranked genes from UNC-3 motif whole genome run. The x-axis show the TargetOrtho best site rank per gene where the rank represents the best motif match cumulative score for each candidate target gene in the genome. Site ranks for

each gene ontology shown on the left (y-axis) are binned. The shading of each bin represents the number of genes within a unique rank bin in a particular gene ontology category. **B.** Gene ontology enrichments of candidate CHE-1 target genes in the top ranked genes from the ASE motif whole genome run. **C.** Gene ontology enrichments of candidate *ttx-3*/*ceh-10* target genes in the top ranked genes from the AIY motif whole genome run. The resulting ontologies among highly ranked predicted TF target genes show enrichments in neurogenesis pathway genes for all three terminal selector genes providing ample candidates for further *in vivo* experimentation.

File S1

Program overview and features

Query list filtering. Further filtering may be applied through user selected query lists (Figure. 2B, Table S3) that restrict the results and/or report specifically on a subset of genes such as putative target genes determined through expression profiling experiments, ChIP-ChIP/ChIP-seq data, or gene ontology associations. The option is especially useful for preliminary TargetOrtho runs as the user may restrict initial analysis to a subset of query genes (option -w) in order to fine tune initial TargetOrtho input parameters. Positive or negative control target genes may be uploaded as a 'training set' using the query list only option so that the user may determine trends in true regulatory target genes. Observations made in this way may be used to weight the final ranking criteria in future TargetOrtho runs (see binding site ranking criteria and the adjustable cumulative site score for weighting details). Upon experimental validation of novel target genes, novel target gene binding sites may be used to improve the initial input PWM and may be added to the initial query list input file for re-evaluation of the ranking criteria weighting schemes.

Genomes. Currently, two reference genomes are available: *C. elegans* and *D. melanogaster*. The reference genome is the genome from which candidate transcription factor target genes are reported. All motif matches in the reference genome are matched to sites in other species' genome to determine the level of motif match conservation among orthologous gene regions. The *C. elegans* reference genome option searches five nematode genomes in the Caenorhabditis genus including *C. elegans, C. briggsae, C. remanei, C. brenneri*, *and C. japonica* while the *D. melanogaster* option searches the melanogaster species subgroup including *D*. *melanogaster, D. sechellia, D. simulans, D. yakuba*, and *D. erecta*. The decision to use these genomes stems from their relatively short evolutionary distance given the availability of complete whole genome sequence. By choosing genomes with limited divergence between them, we expect enough *cis*-regulatory functional conservation to provide strong candidates for *in vivo* validation. Because sequence conservation in regulatory regions may persist despite loss, sub- or neo-functionalization among recently diverged genomes, conservation alone may not be sufficient to predict function. This may be especially true in cases where binding sites and their corresponding binding proteins have co-evolved to allow a certain level of binding site sequence degeneracy. TargetOrtho overcomes this constraint by implementing multiple validating criteria in addition to conservation (see **Binding site ranking criteria below**).

Motif search and scoring procedure. Genome-wide motif searches and motif match scoring utilize the FIMO tool (Grant *et al.* 2011) from the MEME suit (Bailey *et al.* 2009) . Briefly, beginning with a set of experimentally derived binding sites, a consensus

PWM is constructed by the user in meme plain text format (MEME documentation). This input PWM file must include at least one log-odds matrix and/or letter probability formatted matrix**Error! Bookmark not defined.** together with the background ucleotide frequencies (Figure 3A). Background letter frequencies are generally chosen as species-specific upstream nucleotide frequencies and affect the motif match log-likelihood score (See MEME documentation to learn more about building PWMs and choosing appropriate background frequencies). The log-odds matrix used as input for TargetOrtho is an n x 4 matrix where n is the nucleotide length of the binding site alignment. The log-odds format PWM is of the form: $|m_{ij}| = 100*\log_2(p_{ij}/f_j)$ where the matrix is a log-odds matrix calculated by taking 100 times the log (base 2) of the ratio p/f at each position ij in the motif. p is the probability of the nucleotide letter j at position i in the motif, and f is the background frequency of the nucleotide letter j. Columns of the matrix correspond to the letters of the nucleotide alphabet and rows correspond to the positions of the motif with position one coming first (see Meme documentation for a complete description) (meme documentation). The letter-probability matrix is of the form $|m_{ij}| = f_{ij}$ where f is the letter frequency of nucleotide at each position ij in the motif. TargetOrtho accepts direct input from MEME (text format) or the user may submit a MEME formatted log-odds motif as a plain text file and assign a unique name above the motif header in the form "MOTIF name". Up to five separate PWMs may be submitted in the same text file. Each of five species genomes and each input motif (up to five) is searched in parallel resulting in DNA hit coordinates, and motif match scores for each site as the log-likelihood ratio of the motif match compared to the background letter frequency. The motif match results from FIMO may be limited by setting a P value threshold (option -p). The -p option may be of interest for preliminary TargetOrtho runs. Combined with a query list (option -q) of experimentally determined or suspected candidate target genes, the user may restrict initial analysis to a subset of query genes (option -w) in order to fine tune initial TargetOrtho input parameters (Table S3).

Exon association. Each site from each genome is associated with the nearest upstream exon and nearest downstream exon to generate the associated exons tables (Figure 1, Figure 5A,5B). The user may define the number of intervening genes allowed between a site and its associated exon (option -Z). The filter exons option (option -e) allows for the association of sites with only intergenic and intronic genomic regions. Removing all exons from the association step will result in missed sites that reside in single exon genes such as non-coding RNAs. It may be desirable to identify these sites and associate them with the nearest coding gene in which case, the filter exons option should not be used. The offset distance from the first exon or last exon of a gene is then determined for each site where a negative offset represents an upstream distance and a positive offset represents the downstream nucleotide distance. This step is followed by distance filtering using the user defined maximum upstream (option -x) and maximum downstream distance (option -i) as well as the nucleotide distance allowed (option -Z) from the first

exon or last exon if more than 1 intervening genes are positioned between the site and its associated gene. Each step in the exon-association procedure is executed in parallel for each genome for each input motif.

Orthology matching. Each site in the reference genome is then matched to the site having an orthologous gene association in each non-reference genome where the matched site has the smallest variance in offset between species (Figure 5B) within the user defined limit (option -P). The offset variance of a matched group of orthologous sites is defined as the absolute value of the variance of the group of offsets (Figure 5B). This parameter allows for constraint on the positional conservation allowed between species and is scalable via a user defined limit and ranking weight. If the require-region-overlap option is used (option -k) then each matched site must be in the same region as the reference genome site where regions include upstream, downstream, intronic, and exonic loci. If more than one ortholog is associated with a site in a given genome (as may occur with one to many ortholog mapping relationships between genomes), then each site in the reference genome is matched to each orthologous site in each non-reference genome. This may result in one site having multiple unique combinations of orthologous matches of which each is separately ranked in the final results.

Conservation assignment: Each site in the reference genome is assigned a conservation score between 1 and 5 representative of the number of species in which at least one site is associated with an orthologous gene. The conservation assignment is constrained by the require-region-overlap parameter (option -p). For example, if require-region-overlap is set to True, then a reference genome site found upstream of gene X is considered conserved only if the corresponding site in another genome is in the same orthologous region, i.e. upstream of an ortholog of gene X. A conservation score of 1 indicates that the site is only associated with a gene in the reference genome and therefore not conserved, while a conservation score of 5 is assigned when all five genomes have at least one site upstream of a gene and its corresponding orthologs. All site-gene associations, together with general conservation, log-likelihood scores, and offsets are combined into the All-conserved-hits-ranked table (Figure 1) for each motif input taken by TargetOrtho. Each orthology matching step is executed in parallel for each genome and each input motif.

Parameters for TargetOrtho runs. Each P value threshold for the FIMO (Grant *et al.* 2011) genome wide-motif scans was determined by setting the threshold to the highest motif match sequence P value among experimentally validated TF target genes for each PWM. TargetOrtho was set to filter out sites beyond 20,000 nucleotides upstream (-i 20,000) and 20,000 nucleotides downstream (-x 20000), 20 genes were allowed between a site and an associated gene (-Z 20) if the site was within

6,000 bases (-z 6000) of the first or last annotated exon. By allowing 20 annotated genes within 6000 bases, motif matches in promoters with multiple intervening single exon or non-coding RNA genes are still associated with important protein coding genes. Exonic sites were not filtered out (-e False), the query list option (-q) was used to report only on (-w True) the specified TF-dependent genes plus 1,000 random coding genes. See File S3 for all TargetOrtho input parameters for each analysis performed.

Motif construction and data sets. Each motif used for analysis was generated using experimentally validated transcription factor dependent sequences from (Wenick *et al.* 2004, Kim *et al.* 2005, Etchberger *et al.* 2007, Kratsios *et al.* 2012) using the MEME tool (Bailey *et al.* 2009) (See File S1 and File S3 for parameters used for TargetOrtho runs). All analyses were done using the set of previously validated TF-dependent target genes for *unc-3* (Figure S1A motif logos, File S2- gene list 1), ASE (Figure S1C motif logos, File S2 gene list-2), and AIY (Figure S1E motif logos, File S2-gene list 4) motifs respectively compared to 1000 random *C. elegans* protein coding genes (File S2-gene list 6).

PWM verification bias correction and analysis. Because each PWM is constructed from a set of validated DNA sequences whose content determines the resulting log-likelihood score of a given motif match, and because the final cumulative site score for each motif match is constructed using this PWM derived log-likelihood score, all analyses were done in parallel with a motif constructed from promoter sequences of genes not included in the set of validated TF-dependent genes used for comparison to random coding genes. This approach provides a conservative estimate of the significance of the scoring schema. This approach was achieved using the following motifs and gene list combinations for comparative analysis: the EBF-1 motif (Figure S1B motif logos), the mouse UNC-3 homolog binding site, was constructed from mouse DNA sequences derived from ChIP binding data (Treiber *et al.* 2010) and the set of all 50 previously characterized UNC-3 dependent genes (S1–gene list 1) were compared to the set of 1,000 random coding genes (File S2-gene list 6) for analysis; the ASE verification bias corrected motif (Figure S1D motif logos) constructed from a subset of CHE-1 dependent promoter sequences with all CHE-1 dependent gene promoter sequences except those used to constructed the PWM (File S2-gene list 3) compared to 1,000 random coding genes (File S2-gene list 7); the AIY motif (Figure S1E motif logos), generated from ten TTX-3/CEH-10 dependent gene promoter sequences with all TTX-3/CEH-10 dependent genes except those ten used to generate the PWM (File S2-gene list 5) compared to 1,000 random protein coding genes (File S2-gene list 6).

REFERENCES

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., *et al.* 2009 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37: W202-8

Etchberger, J. F., Lorch, A., Sleumer, M. C., Zapf, R., Jones, S. J. *et al.* 2007 The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes. Dev.* 21:1653-74

Grant, C. E., Bailey, T. L., Noble, W. S. 2011 FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27:1017–8

Kim, K. Colosimo, M. E., Yeung, H. *et al.* 2005 The UNC-3 Olf/EBF protein represses alternate neuronal program to specify chemosensory neuron identity. Dev. Biol. 286(1):136-48

Kratsios, P., Stolfi, A., Levine, M., Hobert, O. 2012 Coordinated regulation of cholinergic motor neuron traits through a conserved terminal selector gene. *Nat. Neurosci.* 15:205-14

Treiber, T., Mandel E. M., Pott, S., Györy, I., Firner, S. *et al.* 2010 Early B cell factor 1 regulates B cell gene networks by activation, repression, and transcription- independent poising of chromatin. *Immunity* 32:714-25

Wenick, A.S. & Hobert, O. 2004 Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans. *Dev. Cell* 6:757-70

Tables S1-S12

Available for download as Excel files at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160721/-/DC1

Table S1 Comparison test results for ventral nerve cord neuron counts of GFP fusion reporters in 1. wild type (N2) or 2. *unc-3*(e151) animals.

Table S2 TargetOrtho output files

Table S3 TargetOrtho input parameters

Table S4 Gene lists

Table S5 TargetOrtho parameters for motif analysis

Table S6 Comparison test results for UNC-3 motif analysis

Table S7 Comparison test results for EBF1 motif analysis

Table S8 Comparison test results for ASE motif analysis

Table S9 Comparison test results for ASE verification bias corrected analysis

Table S10 Comparison test results AIY motif analysis

Table S11 Comparison test results for AIY motif analysis (verification bias corrected)

Table S12 Gene Ontology Enrichment Results from Gorilla (Gene Ontology enrichment analysis and visualization tool)