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. A1Y motif derived from ten upstream sequences of previously validated ttx-3/ceh-10 dependent genes (Wenick et al. 2004).
**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 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 shows 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 A1Y motif. **Part III'.** -log(P values) from comparison tests of ttx-3/ceh-10 dependent genes (except those used to generate the A1Y motif compared to 1000 random coding genes for TargetOrtho run with the A1Y motif.
**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

**region=upstream and intron**

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

\[ \text{cumulative site score} = \left( \frac{\sum_{i=1}^{n} w_i c_i}{\sum_{i=1}^{n} a_i} \right) / j \]

\( c_i = \text{raw value of ranking criteria } i. \)
\( a_i = \text{maximum } c_i \)
\( E_i = \text{minimum } c_i \)

**region=upstream and intron**

Representative p values (*log_{10}(p)*) for data shown in CDF plots A-I and A'-G' from each comparison test.

HMFU comparison test results for individual ranking criteria in UNC-3 known target genes vs random coding genes (ERF motif)

K  
\( -\log_{10}(p \text{-value}) \)

L  
\( -\log_{10}(p \text{-value}) \)

**region=upstream and intron**

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

**region=upstream and intron**

Values in the grey area are weights, w, applied to ranking criteria A-G used to derive the cumulative site score* in H, I, and J.
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.** \(-\log_{10}(P \text{ value})\) 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.** \(-\log_{10}(P \text{ values})\) 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’).
CHEs known target gene motif match data vs. random coding gene motif match data (ARE motif) from upstream and intronic regions

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

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

\[
\text{cumulative site score} = \left( \frac{c_i - E_i}{a_i - E_i} \right) \cdot Z
\]

\(c_i\) = raw value of ranking criteria \(i\).
\(a_i\) = maximum \(c_i\)
\(E_i\) = minimum \(c_i\)

Figure S4

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

K MNU comparison test results for individual ranking criteria in CHEs known target genes vs. random coding genes (ARE motif)

L MNU comparison test results cumulative site scores in CHEs known target genes vs. random coding genes (ARE motif)
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 H, I, H', I', J'. CDF plot 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 the cumulative site scores derived from all averaged species ranking criteria (A'-G').

K. -log10(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. –log10(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').
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

Averaged species data

- Log(p value) vs cumulative site score

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

**region=upstream and intron

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

C. elegans data

Averaged Species data

- Log(p value) vs cumulative site score

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

**region=upstream and intron

ORJ 10 (p value)

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’).

L. Glenwinkel et al. 9 S1
**C. elegans Ranking Criteria**

<table>
<thead>
<tr>
<th></th>
<th>target gene region** site score</th>
<th>target gene region** score</th>
<th>target gene region** site count</th>
<th>target gene total score</th>
<th>target gene total site count</th>
<th>target gene region** conservation</th>
<th>target gene offset variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Average Species Ranking Criteria**

<table>
<thead>
<tr>
<th></th>
<th>target gene region** site score</th>
<th>target gene region** score</th>
<th>target gene region** site count</th>
<th>target gene total score</th>
<th>target gene total site count</th>
<th>target gene region** conservation</th>
<th>target gene offset variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

* cumulative site score = \(\sum_{i=1}^{n} c_i \times \alpha_i \times \varepsilon_i \times \omega_i \times a_i \)

\(c_i\) - raw value of ranking criteria i.
\(\alpha_i\) - maximum value of ranking criteria i.
\(\varepsilon_i\) - minimum value of ranking criteria i.
\(\omega_i\) - weight for ranking criteria i (shown in gray boxes)
\(a_i\) - number of criteria with weight \(\omega_i\)
Figure S6  AIY motif analysis. ceh-10/tx-3 dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized ceh-10/tx-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 H,I,H',I',J' 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 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').
**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**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>C. elegans</th>
<th>H. sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B'</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C'</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D'</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E'</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F'</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G'</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**region=upstream and intron**

- **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**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>C. elegans</th>
<th>H. sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**cumulative site score = \[ \sum_{i=1}^{n} E_i \]**

- **C. elegans**
- **H. sapiens**

---

**Figure 3**

- Representative p values (-log10(p)) for data shown in CDF plots A-I and A'-G' from each comparison test
- MSHV comparison test results for individual ranking criteria in C. elegans vs. H. sapiens (default method J' used by TargetOrtho for generation of cumulative site scores)
Figure S7  AIY motif analysis with verification bias corrected data. *ceh-10/tx-3* dependent target gene data (blue) compared to random coding gene data (grey). The set of previously characterized *ceh-10/tx-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. -log10(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. –log10(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.
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. Background nucleotide frequencies are generally chosen as species-specific upstream nucleotide frequencies (Figure 3A). Background letter frequencies are used with the background 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: \( |m_{ij}| = 100 \log_2(p/f) \) 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 \( |m_{ij}| = f_i \) where f is the letter frequency of nucleotide at each position i 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


Tables S1-S12

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)