2015-04-23

Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities

Kamesh Narasimhan
University of Toronto

Et al.

Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities

Kamesh Narasimhan, Samuel A Lambert, Ally W H Yang, Jeremy Riddell, Sanie Mnaimneh, Hong Zheng, Mihai Albu, Hamed S Najafabadi, John S Reece-Hoyes, Juan I Fuxman Bass, Albertha J M Walhout, Matthew T Weirauch, Timothy R Hughes

DOI: http://dx.doi.org/10.7554/eLife.06967

Cite as: eLife 2015;10.7554/eLife.06967

Received: 13 February 2015
Accepted: 22 April 2015
Published: 23 April 2015

Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities

Kamesh Narasimhan[1*], Samuel A. Lambert[2*], Ally W.H. Yang[1*], Jeremy Riddell[3], Sanie Mnaimneh[1], Hong Zheng[1], Mihai Albu[1], Hamed S. Najafabadi[1], John S. Reece-Hoyes[4], Juan I. Fuxman Bass[4], Albertha J.M. Walhout[4], Matthew T. Weirauch[5¶] & Timothy R. Hughes[1,2,6¶]

[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada

[2]Department of Molecular Genetics, University of Toronto, Canada

[3]Department of Molecular and Cellular Physiology, Systems Biology and Physiology Program, University of Cincinnati, Cincinnati, OH, USA

[4]Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA, USA

[5]Center for Autoimmune Genomics and Etiology (CAGE) and Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA

[6]Canadian Institutes For Advanced Research, Toronto, Canada

*Authors contributed equally

¶To whom correspondence should be addressed

t.hughes@utoronto.ca or matthew.weirauch@cchmc.org

Ph: 416-946-8260

FAX: 416-978-8287

ABSTRACT

Caenorhabditis elegans **is a powerful model for studying gene regulation, as it has a compact genome and a wealth of genomic tools. However, identification of regulatory elements has been limited, as DNA-binding motifs are known for only 71 of the estimated 763 sequence-specific transcription factors (TFs). To address this problem, we performed protein binding microarray experiments on representatives of canonical TF families in *C. elegans*, obtaining motifs for 129 TFs. Additionally, we predict motifs for many TFs that have DNA-binding domains similar to those already characterized, increasing coverage of binding specificities to 292 *C. elegans* TFs (~40%). These data highlight the diversification of binding motifs for the nuclear hormone receptor and C2H2 zinc finger families, and reveal unexpected diversity of motifs for T-box and DM families. Motif enrichment in promoters of functionally related genes is consistent with known biology, and also identifies putative regulatory roles for unstudied TFs.**

INTRODUCTION

Transcription factors (TF) are sequence-specific DNA binding proteins that control gene expression, often regulating specific biological processes such as pluripotency and differentiation (Takahashi and Yamanaka 2006), tissue patterning (Lemons and McGinnis 2006), the cell cycle (Evan et al. 1994), metabolic pathways (Blanchet et al. 2011), and responses to environmental stimuli (Benizri et al. 2008). The nematode *C. elegans* is a powerful model for studying gene regulation as it is a complex and motile animal, yet has a compact genome (~100 Mbp) (C.elegans consortium 1998) featuring relatively short intergenic regions (mean 1,389 bp; median 662 bp). Indeed, the observation that proximal promoter sequence is often sufficient to produce complex tissue-specific gene expression patterns (Dupuy et al. 2004; Zhao et al. 2007; Grove et al. 2009; Sleumer et al. 2009; Niu et al. 2011) indicates that long-range gene regulation through enhancers is not as abundant in *C. elegans* as it is in flies or mammals (Gaudet and McGhee 2010; Reinke et al. 2013).

C. elegans has 934 annotated TFs (Reece-Hoyes et al. 2005), and 744 proteins that possess a well-characterized sequence-specific DNA-binding domain (Weirauch and Hughes 2011; Weirauch et al. 2014). *C. elegans* contains major expansions of several specific TF families, with Nuclear Hormone Receptor (NHR), Cys_2His_2 (C2H2) zinc finger, homeodomain, bHLH, bZIP, and T-box together comprising 74% of the TF repertoire (Reece-Hoyes et al. 2005; Haerty et al. 2008). The lineage-specific expansion of C2H2 zinc finger TFs is similar to that observed in many animals, including diversification of DNA-contacting "specificity residues", suggesting diversification in DNA binding specificity (Stubbs et al. 2011). The *C. elegans* genome encodes an unusually large number of NHRs (274 members), more than five times the number in human (48 members) (Enmark and Gustafsson 2001; Reece-Hoyes et al. 2005). It is speculated that the

63 NHRs may serve as environmental sensors (Enmark and Gustafsson 2001; Arda et al. 2010),
64 providing a possible explanation for their variety and numbers. Five of the six major NHR sub-
65 families found across metazoa are also found in *C. elegans* (NR3 is lacking), but the vast
66 majority of *C. elegans* NHRs define novel sub-families that are not present in other metazoans
67 (Van Gilst et al. 2002) and which are derived from an ancestral gene most closely resembling
68 HNF4 (aka NR2A) (Robinson-Rechavi et al. 2005). Extensive variation in the DNA-contacting
69 recognition helix or "P-box" suggests that *C. elegans* NHRs, like C2H2 and bHLH families,
70 have diversified DNA sequence specificities, and that many will recognize novel motifs (Van
71 Gilst et al. 2002). The T-box gene family presents another example of a nematode-specific
72 expansion, with 22 members in *C. elegans*, of which 18 lack one-to-one orthologs in other
73 metazoan lineages (Minguillon and Logan 2003). Only four have known binding motifs, and
74 unlike most other TFs, T-box binding motifs are virtually identical across the metazoa (Sebe-
75 Pedros et al. 2013; Weirauch et al. 2014); the diversification of TFs is often associated not only
76 with changes in DNA sequence specificity, but also alteration in protein-protein interactions and
77 expression of the TF gene itself (Grove et al. 2009; Reece-Hoyes et al. 2013).

78 Despite extensive study of gene regulation, including several large-scale efforts (Deplancke et al.
79 2006; Grove et al. 2009; Lesch et al. 2009; Gerstein et al. 2010; Niu et al. 2011; Sarov et al.
80 2012; Reece-Hoyes et al. 2013; Araya et al. 2014), the landscape of *C. elegans* TF sequence
81 specificities remains largely unknown. To our knowledge, motifs are currently known for only
82 71 *C. elegans* TFs, including those determined in single-gene studies, previous PBM analyses,
83 and modENCODE TF ChIP-seq data (Matys et al. 2006; Araya et al. 2014; Mathelier et al. 2014;
84 Weirauch et al. 2014). It has been surprisingly difficult to obtain motifs from ChIP-seq data
85 (Niu et al. 2011; Araya et al. 2014), possibly due to indirect binding, or a dominant role of

86 chromatin structure in either determining *in vivo* binding sites (Song et al. 2011) or in the
87 purification of chromatin fragments (Teytelman et al. 2013). Yeast one-hybrid (Y1H) assays
88 (Reece-Hoyes et al. 2011) cannot be used easily to derive TF motifs, because the DNA
89 sequences tested are too large (~2 kb on average). However, there is a strong statistical
90 correspondence between motifs determined by Protein Binding Microarrays (PBMs) and Y1H
91 data (Reece-Hoyes et al. 2013). Computational approaches coupling promoter sequence
92 conservation and/or gene expression data to identify TF motifs *de novo* have collectively
93 produced many more motifs than there are TFs (Beer and Tavazoie 2004; Sleumer et al. 2009;
94 Zhao et al. 2012), and also do not inherently reveal the cognate TFs that correspond to each
95 putative motif. Multimeric binding represents one possible complication in the analysis of *in*
96 *vivo* TF binding data (Ao et al. 2004). Indeed, TF co-associations were identified based on
97 ChIP-seq peak binding overlaps in *C. elegans* modENCODE studies (Araya et al. 2014), but the
98 underlying sequence recognition mechanisms were not apparent.

99 Here, we use PBMs to systematically identify *C. elegans* TF DNA-binding motifs. We selected
100 a diverse set of TFs to assay, ultimately obtaining 129 motifs from different TF families and
101 subclasses. The data show that the expansion of most major TF families is associated with
102 diversification of DNA-binding motifs. Motif enrichment in promoters reveals that our motif
103 collection readily associates individual TFs with putative regulated processes and pathways.

104

105 **RESULTS**

106 **Overview of the PBM data**

107 The key goal of this project was to expand our knowledge of DNA sequence specificities of *C.*
108 *elegans* TFs. To do this we analyzed a diverse set of TF DNA-binding domains (DBDs) (see
109 below) with PBM assays (Berger et al. 2006; Weirauch et al. 2014). Briefly, the PBM method
110 works by "hybridizing" a GST-tagged DNA-binding protein (in our assays, the DNA-binding
111 domain of a TF plus 50 flanking amino acids) to an array of ~41,000 defined 35-mer double-
112 stranded DNA probes. The probes are designed such that all 10-mer sequences are present once
113 and all non-palindromic 8-mers are thus present 32 times in difference sequence contexts
114 (palindromic 8-mers occur 16 times). A fluorescently labelled anti-GST antibody illuminates the
115 extent to which each probe is bound by the assayed TF. Using the signal intensity for each
116 probe, the specificity of the TF is derived. For each individual 8-mer, we derive both E-scores
117 (which represent the relative rank of microarray spot intensities, and range from -0.5 to + 0.5
118 (Berger et al. 2006)) and Z-scores (which scale approximately with binding affinity (Badis et al.
119 2009)). PBMs also allow derivation of Position Weight Matrices (PWMs) up to 14 bases long
120 (Berger et al. 2006; Mintseris and Eisen 2006; Badis et al. 2009; Weirauch et al. 2013)
121 (hereafter, we take "motif" to mean PWM). To determine PWMs, we used the data from PBM
122 assays performed on two different array designs to score the performance of PWMs obtained
123 from different algorithms, as previously described (Weirauch et al. 2013; Weirauch et al. 2014).

124 In this study, we selected TFs to analyze on the basis of their DBD sequence, aiming to examine
125 at least one TF from each group of paralogous TFs, and biasing against TFs that have known
126 PBM motifs, or close orthologs or paralogs with known motifs (see **Methods** for full description

127 of selection scheme). The selections were guided by previous PBM analyses that determined
128 sequence identity thresholds for each DBD class that correspond to motif identity (Weirauch et
129 al. 2014). To identify TFs, we used the CisBP definition of DBDs (Weirauch et al. 2014), which
130 employs a list of well-characterized eukaryotic DBDs and a distinct significance threshold for
131 each DBD class. CisBP identified 744 *C. elegans* proteins, encompassing 52 domain types
132 (listed in **Figure 1–source data 1**). 689 (93%) of these 744 are present in the wTF catalog of
133 934 annotated TFs (Reece-Hoyes et al. 2005); thus these sets are largely overlapping. We
134 manually examined the differences between the two TF lists (see **Supplemental File 1**) and
135 found that most of them can be accounted for by (i) changes to the *C. elegans* protein catalog
136 over time; (ii) differences in domain classes included; (iii) differences in domain score threshold,
137 (iv) fewer manual annotations in CisBP, and (v) ambiguity in classifying C2H2 zinc fingers as
138 TFs. Overall, wTF2.0 contains only 19 proteins that are not in CisBP and that are very likely
139 *bona fide* sequence-specific TFs. wTF2.0 also contains 83 C2H2 proteins that fall below the
140 CisBP score threshold, 52 of which have only a single C2H2 domain. DNA recognition
141 typically requires multiple C2H2 domains; however, some fungal TFs do bind DNA with a
142 single C2H2, employing additional structural elements (Wolfe et al. 2000). Thus, these proteins
143 have an ambiguous status. In general, CisBP excludes proteins with lower domain scores and
144 those with little or no evidence for sequence-specific DNA binding, and we therefore refer to the
145 744 in CisBP plus the 19 additional *bona fide* TFs as the 763 "high confidence" *C. elegans* TFs.

146 We attempted to clone DBDs from 552 unique high confidence TFs, ultimately obtaining clones
147 for 449, all of which we assayed by PBMs. After employing stringent success criteria (see
148 **Methods**) we obtained sequence specificity data (8-mer scores and motifs) for 129 DBDs. PBM
149 "failures" may be due to any of several causes, including protein misfolding, requirement for

150 cofactors or protein modifications (e.g. phosphorylation), or *bona fide* lack of sequence-specific
151 DNA binding activity. The overall success rate (29%) is comparable to that we have observed
152 from analysis of thousands of DBDs from diverse species (35%) (Weirauch et al. 2014).

153 A summary of our results is presented in **Figure 1**, broken down by motif numbers and percent
154 coverage for individual DBD classes. Our motif collection encompasses 26 different DBD
155 classes, and greatly increases the number and proportion of *C. elegans* TFs for which motifs
156 have been identified experimentally, from 71 (10%) to 195 (26%) (five of the 129 had
157 previously-known motifs). The new data encompass all of the large TF families, including
158 C2H2 zinc fingers, NHRs, bZIPs, homeodomains, DM domains, and GATA proteins.

159 **Validation of motifs, motif novelty, and motifs predicted using homology**

160 We next asked whether our new data are consistent with previous knowledge. Of the 129 TFs,
161 only five have previously known motifs, all of which we recapitulated (**Figure 1–figure**
162 **supplement 1**). The sequence preferences for most of the 129 TFs were different from those of
163 any previously assayed TF, however. The boxplots in **Figure 2A, B, C,** and **Figure 2–figure**
164 **supplements 1-4** show that, on average, the new TFs we analyzed bound a set of 8-mers that
165 was largely non-overlapping with that of the most similar protein that had been analyzed
166 previously by PBM (red circles indicate the 8-mer overlap between individual TFs analyzed by
167 PBM in our study, and the most similar TF analyzed by PBM in any study). Nonetheless, some
168 pairs of TFs have DBDs that are highly similar, and bind highly overlapping 8-mers. These
169 observations are quantitatively consistent with the prior study we used for guidance in selecting
170 TFs (black box plots) (Weirauch et al. 2014), and thus, we expect that the scheme for predicting
171 sequence specificity via amino acid identity that was proposed in the prior study can also be used

172 in *C. elegans*. In this scheme, TFs without DNA-binding data are simply assigned the motifs and
173 8-mer data for other TFs with DBD amino acid similarity above a threshold, if those data exist.
174 These TFs can be from *C. elegans* or from other species. If we include these predicted motifs,
175 then the number of *C. elegans* TFs with an associated motif increases to 292 (39%), including
176 TFs with motifs predicted from other *C. elegans* TFs (24) and those with motifs predicted from
177 other species (79).

178 **Expert curation of motifs**

179 The entire *C. elegans* motif collection, including our new data, previously published motifs, and
180 those predicted by homology from other TFs in *C. elegans* and other species, encompasses 1,769
181 unique motifs representing only 292 TFs. About half (157, or 54%) of the 292 TFs with motifs
182 are represented by only a single motif, as there was no data prior to our study for these TFs or
183 their close homologs. Some TFs (e.g. homeodomains, PAX, and forkheads), however, are highly
184 conserved and thus have many orthologs above the prediction threshold. In addition, TFs that
185 are known developmental regulators tend to be well studied, and often possess multiple
186 associated motifs. To gain an overview of the full motif collection, and to compare among the
187 multiple motifs for each protein, we used the PWMclus tool (Jiang and Singh 2014), with default
188 settings, to obtain groups of highly-related motifs from all TFs within each DBD class. This tool
189 uses an information-content weighted Pearson correlation between aligned PWM columns as a
190 similarity measure for hierarchical clustering, then selects branches within which the average
191 internal correlation exceeds $R \geq 0.8$. This procedure collapsed the 1,769 motifs into a set of 424
192 clusters. This number is still larger than the number of TFs with either known or predicted
193 motifs (292), since there are many cases in which motifs for a single TF are distributed across

194 multiple clusters, although in 67% of cases in which there are multiple known and predicted
195 motifs for a given protein, the majority of them do form a single cluster.

196 There appear to be several explanations for this phenomenon, as exemplified by the bZIP family
197 shown in **Figure 2D**. First, different studies and different experimental (or computational)
198 techniques often yield motifs for the same protein that are clearly related by visual examination,
199 but score as different from each other using PWMclus. For example, there are four different
200 motifs for *skn-1* (from PBM, Chip-seq, and Transfac) that all contain the same half-site, ATGA,
201 but have different flanking sequence preferences. Similarly, for Forkhead TFs FKH-1 and UNC-
202 130, different methods produce variants with differences in the sequences flanking the core
203 TGTTT Forkhead binding site. A related explanation is that a single motif may not adequately
204 capture all aspects of TF sequence preferences, such as the ability of many TFs to bind as both a
205 monomer and a homodimer (or multimer) with preferred spacing and orientation, variability in
206 the preferred spacing, changes to the preferred monomeric sites that are associated with
207 dimerization, and effects of base stacking that result in preferred polynucleotides at some
208 positions (Jolma et al. 2013). In addition, different experimental methods may capture some
209 aspects of DNA binding complexity better than others.

210 It is inconvenient to have a large number of motifs for a single protein for several reasons. First,
211 it is difficult to peruse the full motif collection. In addition, comprehensive motif scanning is
212 slower with a large number of motifs, and the motif scans produce partially redundant results that
213 require deconvolution and reduce statistical power. We therefore sought to identify a single
214 motif or set of motifs for each protein that are minimally redundant, and are best supported by
215 existing data. We used a semi-automated scheme that considers all data available (similar to that
216 described in (de Boer and Hughes 2011); see **Methods**). Briefly, we prioritized motifs that are

217 (a) measured experimentally, rather than predicted; (b) more similar to other motifs for the same
218 TF, or highly similar TFs, especially if they are derived from *in vitro* data, which would be free
219 of confounding effects present *in vivo*; (c) assigned to the cluster that contains the majority of
220 motifs for that TF; (d) most consistent with the type of sequences that a given DBD class
221 typically binds; (e) best supported by ChIP-seq or Y1H data, if available (see below).

222 This procedure resulted in a set of 284 motifs representing the 292 *C. elegans* TFs with
223 experimentally determined or predicted motifs (**Supplemental File 2**). The outcome for the
224 bZIP family is shown on the right of **Figure 2D**, which illustrates that the motif curation
225 procedure produces motifs that are consistent with known bZIP class binding sites. The curated
226 set also contains 16 cases in which the same protein is represented by multiple motifs
227 (exemplified by the GATA family TF ELT-1, which binds as both a monomer and a homodimer,
228 **Figure 2–figure supplement 5**), and 11 cases in which more than one protein is represented by
229 the same motif (e.g. GATA family TFs MED-1 and MED-2, **Figure 2–figure supplement 5**; in
230 all of these cases, the TFs are highly similar proteins). We also note that PWMclus subdivides
231 the 284 curated motifs into only 127 different clusters (data not shown), because the motif(s)
232 contained in many of the clusters met few or none of the selection criteria above.

233 **Overview of PBM 8-mer data**

234 The majority of the expert curated motifs (237, or 84%) are derived from the PBM data
235 described in this study or from previous studies (compiled in (Weirauch et al. 2014)), which are
236 the only data available for the majority of the 292 TFs with motifs. We reasoned that the PBM
237 data should facilitate direct comparison among TF sequence preferences, as they were generated
238 using identical methodology. In addition, PBMs facilitate comparisons because they produce

239 scores for individual DNA 8-mers. Thus, to complement the PWM analysis above, we examined
240 as a composite the 8-mer E-score data for all of the TFs analyzed in this study using PBMs.
241 **Figure 3** illustrates that the 8-mers recognized by each individual protein are in general distinct,
242 and further highlights the distinctiveness of the sequences preferred by different TFs that share
243 the same type of DBD. For example, *C. elegans* homeodomain and Sox TFs display different
244 sequence preferences that largely reflect the known subclasses (**Figure 3** and data not shown; all
245 data and motifs are available in the Cis-BP database (see Data Access section below). We also
246 observed subtle differences in Forkhead DNA sequence preferences: despite the motifs having
247 similar appearance, the proteins prefer slightly different sets of 8-mers, as previously observed
248 using only PBM data (Badis et al. 2009; Nakagawa et al. 2013), indicating that these variations
249 are not due to differences in methodology. Other large *C. elegans* TF families display
250 undocumented and unexpected diversity in their DNA sequence preferences, which we next
251 examined in greater detail.

252 **Complex relationships between protein sequences and motifs recognized by the NHR**
253 **family**

254 Previously, the literature contained motifs for only eight of the 271 *C. elegans* NHRs, while
255 motifs for an additional 13 could be predicted from orthologs and paralogs (Hochbaum et al.
256 2011; Weirauch et al. 2014). It has also been reported that additional *C. elegans* NHRs bind
257 sequences similar to those bound by their counterparts in other vertebrates (Van Gilst et al.
258 2002), but the data available does not lend itself to motif models that can be used for scanning.
259 We obtained new PBM data for 20 *C. elegans* NHRs (**Figure 4**), among which only one had a
260 previously known motif (DAF-12, which yielded a motif identical to one found by ChIP-chip
261 (Hochbaum et al. 2011)). None of the remaining 19 could have been predicted by simple

262 homology; due to their widespread divergence, and absence of motifs for most NHRs, few motifs
263 can be predicted by homology among the *C. elegans* NHR class at our threshold for motif
264 prediction (70% identity for NHRs). However, these 19 new NHR motifs do lead to predicted
265 motifs for eight additional *C. elegans* NHRs.

266 The most striking feature of the NHR motifs is their diversity, but an equally surprising
267 observation is that very different NHRs can bind very similar sets of sequences. Data from the
268 27 NHRs that have been analyzed by PBMs in our study or others are shown in **Figure 4**. We
269 obtained 13 different groups of motifs, using the PWMclus methodology described above
270 (indicated by shading of dendrogram labels in **Figure 4**). We expected that all 27 of these NHRs
271 might have yielded a distinct motif, as no two are more than 70% identical to each other. In
272 several cases, however, NHRs with very different overall DBD sequences (below the threshold
273 for predicting motif identity) in fact display similar sequence preferences, while more similar
274 NHR TFs often bind different motifs, as the shading on the labels in **Figure 4** does not strictly
275 reflect the dendrogram. We also note that the data for individual 8-mers appears more complex
276 than the motif groups capture (see heatmaps in **Figure 4**). For example, the individual 8-mer
277 scores for TFs represented by the two largest groups of motifs - sets binding sequences related to
278 G(A/T)CACA and (A/T)GATCA, respectively - indicates that they may in fact possess distinct
279 DNA sequence preferences (**Figure 4**, top and bottom). These subtle and complex differences
280 are presumably obscured by the motif derivation process, which tends to produce degenerate (i.e.
281 low information content) motifs for most of these TFs. In addition, or possibly as a
282 consequence, the default correlation threshold used by the PWMclus algorithm groups these TFs
283 together.

284 To examine the determinants of NHR sequence preferences more closely, we considered NHR
285 recognition helix (RH) sequences (**Figure 4**, middle). Of the 95 unique RH sequences found in
286 *C. elegans* NHRs, 15 are found in our data, including multiple representatives of most of the
287 populous RHs (our data contain ten of the 75 with RA-AA; 3 of the 19 with NG-KT; 2 of the 10
288 with NG-KG; and one of the seven with AA-AA). It is believed that identity in the recognition
289 helix corresponds to identity in sequence preference (Van Gilst et al. 2002); surprisingly,
290 however, we found that TFs with identical RH sequences can bind very different DNA
291 sequences. For example, NHR-177 shares the RA-AA recognition helix with nine other NHRs
292 examined in our study, yet binds a completely different set of sequences (resembling CGAGA,
293 unlike the CACA-containing motifs of the others). Conversely, NHRs with different RH
294 sequences can have very similar DNA sequence preferences. NHR-66 and NHR-70, for
295 example, differ at two of the four variable residues in the recognition helix (AA-SA vs. RA-AA),
296 and share only ~49% amino acid identity (and NHR-66 contains a three-residue insertion). Yet
297 they bind highly overlapping sets of 8-mers, and produce motifs featuring CTACA. Thus, there
298 is an imperfect correspondence between identity in the recognition helix and identity in DNA
299 binding sequence preferences, suggesting that additional residues within (or flanking) the DBD
300 contribute to the specificity of *C. elegans* NHR proteins. These observations also show that,
301 when NHRs with very different overall DBD sequences bind similar motifs, it is typically not
302 due to the two proteins sharing the same RH.

303 Only one NHR, SEX-1, produced a motif strongly resembling the canonical steroid hormone
304 response element (SHRE) (GGTCA); SEX-1 shares three of four variable residues in the
305 recognition helix with canonical SHRE binding TFs such as the Estrogen Receptor (SEX-1: EG-

306 KG; ER: EG-KA). Moreover, none of the NHRs examined produced a motif matching that of
307 HNF4, the presumed ancestor of most *C. elegans* NHRs.

308 **Motifs for *C. elegans* C2H2 TFs are supported by the recognition code**

309 We obtained new PBM data for 42 C2H2 zinc finger (ZF) TFs (**Figure 5**), only one of which
310 was previously known (**Figure 1–figure supplement 1**). Previously there were only six
311 experimentally-determined *C. elegans* C2H2 motifs in the literature, and 11 that could be
312 predicted by homology, all of which are well conserved in distant metazoans (members of KLF,
313 SP1, EGR, SNAIL, OSR, SQZ, and FEZF families); seven of these are among our data and have
314 PBM motifs consistent with those predicted (data not shown). Only two additional TFs (ZTF-25
315 and ZTF-30) can be assigned motifs by homology using our new data. Together, the new data
316 and predictions bring the total number of *C. elegans* C2H2 TFs with motifs to 53 (~50% of the
317 107 C2H2s in our list of 763 TFs).

318 The C2H2 motifs are diverse (**Figure 5**), but unlike the NHR family, the molecular determinants
319 of C2H2 DNA sequence specificities are more readily understood. The motifs we obtained are
320 broadly consistent with previously determined relationships between DNA contacting residues
321 and preferred bases (the so-called "recognition code") (Wolfe et al. 2000), although the motifs
322 predicted by the recognition code are not sufficiently accurate to be used in motif scans (median
323 R^2 = 0.21 vs. predictions made by an updated recognition code that surpasses all previous
324 recognition codes when compared against gold standards (Najafabadi et al. 2015). While most
325 of the motifs are similar to those predicted by the recognition code (**Figure 5–figure**
326 **supplement 1**), lower similarity is observed for TFs with unusual inter-C2H2 linker lengths and
327 atypical zinc-coordinating residues (**Figure 5–figure supplement 1**). In some cases, differences

328 in the motifs obtained from related C2H2 TFs can be rationalized: **Figure 5** (right) shows the
329 example of paralogs EGRH-1 and EGRH-3, in which the motifs obtained by PBM closely reflect
330 those predicted by the recognition code, which differ at several positions. **Figure 5** also shows
331 the example of Snail homologs CES-1 and K02D7.2, in which a short linker between fingers 2
332 and 3 may explain the truncated motif in K02D7.2, and may also explain the differences
333 previously observed between these two proteins in Y1H assays (Reece-Hoyes et al. 2009).

334 **Unexpected diversity in T-box DNA binding specificities**

335 We obtained motifs for four nematode-specific T-box TFs (i.e. lacking one-to-one orthologs in
336 other phyla): TBX-33, TBX-38, TBX-39, and TBX-43. In addition, TBX-40 was previously
337 analyzed by PBM, and our motif for the related protein TBX-39 (93% identical) is very similar.
338 T-box TFs can bind to dimeric sites, with the characteristic spacing and orientation varying
339 among different T-box proteins (Jolma et al. 2013). The monomeric sequence preference
340 (resembling "GGTGTG") is thought to be constant, however, as it is observed across different T-
341 box classes and in distant phyla (Sebe-Pedros et al. 2013; Weirauch et al. 2014). Strikingly, our
342 new PBM data indicate that monomeric T-box sites can also vary considerably (**Figure 6A**).
343 While the motifs for TBX-38 and TBX-43 are highly similar to the canonical "GGTGTG" motif,
344 TBX-33, TBX-39 and TBX-40 exhibit novel recognition motifs.

345 The primary determinants of sequence specificity of T-box TFs are believed to reside in amino-
346 acid residues located in α-helix 3 and the 3_{10}-helixC, which contact the major and minor groove,
347 respectively (Muller and Herrmann 1997; Coll et al. 2002; Stirnimann et al. 2002), and indeed,
348 the DNA contacting residues in TBX-33, -39, and -40 are different from those in T-box TFs that
349 bind the canonical motif (**Figure 6–figure supplement 1**). In addition, TBX-39 and -40 exhibit

350 sequence deletion in the "variable region", and TBX-33 has an 18 amino acid insertion in the
351 region leading up to the β-strand e', which could also potentially alter sequence preferences via
352 structural rearrangements.

353 **Variation in motifs for DM domains highlights nematode-specific expansions**

354 DM TFs are well studied because of their established roles in sex determination, and previous
355 analyses established that different DM TFs often bind distinct motifs that typically contain a
356 TGTAT core, including Drosophila doublesex, for which the family is named (Gamble and
357 Zarkower 2012). *C. elegans* and other nematodes encode several lineage-specific DM TFs in
358 addition to orthologs shared across metazoans, with eight of the eleven *C. elegans* DM domains
359 having less than 85% identity (our threshold for DM motif prediction) to any DM domain in
360 insects and vertebrates (Weirauch et al. 2014). Accordingly, most of the *C. elegans* DM
361 domains have highest preference for sequences that are different from TGTAT, although in all
362 but two cases the motifs do contain a TGT (**Figure 6B**). DM domains encode intertwined
363 CCHC and HCCC zinc binding sites, and are hypothesized to bind primarily in the minor groove
364 (Zhu et al. 2000; Narendra et al. 2002). A DNA-protein structure has not yet been described for
365 any DM protein, however; mapping the determinants of their variable DNA sequence
366 preferences will therefore require further study.

367 **Motif enrichment in Y1H and ChIP-seq data**

368 We next examined whether motifs from our collection correspond to modENCODE TF ChIP-seq
369 data (Araya et al. 2014), and to TF prey - promoter bait interactions from Y1H experiments
370 ((Reece-Hoyes et al. 2013) and J.F-B. and A.J.M.W., unpublished data). Among the 40 TFs
371 analyzed by ChIP-seq and present in our motif collection, peaks for 20 TFs displayed central

372 enrichment of motif scores (q-value < 0.05) using the CentriMo algorithm on the top 250 peaks
373 (Bailey and Machanick 2012)) (**Figure 7**). Similarly, among 145 TFs both analyzed by Y1H and
374 present in our motif collection, motif affinity scores for 103 were significantly enriched (Mann-
375 Whitney U test; q-value < 0.05) among promoter sequences scoring as positive by Y1H, relative
376 to those scoring as negative by Y1H (**Figure 7–figure supplement 1**). The correspondence
377 among these data sets is presumably imperfect due to indirect DNA binding *in vivo*, and/or the
378 impact of chromatin and cofactors on binding site selection (Liu et al. 2006), both of which occur
379 in *C. elegans* and yeast. We note that, among the 25 TFs that are present in Y1H data, ChIP-seq
380 data, and our motif collection, 11 are only significantly enriched in Y1H (using the cutoff
381 above), five are only significantly enriched in ChIP-seq, and only five are significantly enriched
382 in both. Thus, most motifs (21/25; 84%) can be supported by independent assays, although the
383 *in vivo* assays appear to capture different aspects of TF binding. Overall, the clear relationship
384 between our motifs and independent data sets strongly supports direct *in vivo* relevance of the
385 motifs.

386 We also examined whether we could detect multimeric or composite motifs (CMs) in existing
387 ChIP-seq data sets by searching for enrichment of patterns in which there is fixed spacing and
388 orientation between two or more motifs within the peaks, one of which corresponds to the TF
389 that was ChIPed. We identified 185 significantly enriched CMs (see **Methods**) involving 11/40
390 ChIPed TFs, and 14 different TF families (including partner motifs) (**Figure 8, Figure 8–source**
391 **data 1**). As an example, the most highly significant result involves NHR-28, in which the six
392 base core sequence "ACTACA" (which could correspond to NHR-28 or NHR-70) is found
393 repeated in both dimeric and trimeric patterns (**Figure 8A, top**). We also identified a CM
394 involving LSY-2 (a bZIP protein) and NHR-232 with a spacing of one base between the core

395 motifs (**Figure 8A, middle**). A subset of these instances included an additional ZIP-6 (bZIP)
396 motif at a one base distance 3' of the NHR-232 motif, yielding a multi-family trimeric CM
397 (**Figure 8A, bottom**).

398 PWMclus grouped the 185 CMs into 37 clusters (**Figure 8–figure supplements 1-5**). Most of
399 the CMs were identified repeatedly for the same TF ChIPped in different developmental stages;
400 these instances were considered separately in the analysis above, and highlight the robustness of
401 the observations. Some of the clusters also correspond to CMs containing motifs for related TFs,
402 demonstrating robustness to the exact motif employed. Our methodology allowed the individual
403 motifs to overlap, and half of the 37 CM clusters represent such overlaps. However, the majority
404 of overlaps occur in the flanking low-information-content sections of motifs, such that most the
405 37 CM clusters resemble a concatenation of two motif "cores" with or without a small gap (1-4
406 bases). Surprisingly, 17 of the 37 clusters (~46%) were obtained from the embryonic ChIP-seq
407 data for the poorly-characterized, essential bZIP protein F23F12.9 (ZIP-8), which is most similar
408 to human ATF TFs and binds both the ATF site and the CREB site (**Figure 8–figure**
409 **supplements 1-3**). In total, these results suggest that multimeric interactions within and between
410 TF families may be a prevalent phenomenon in *C. elegans*.

411 **Motif enrichment in tissue and developmental-stage specific expression data**

412 To identify potential roles for TFs in the regulation of specific groups of functionally related
413 genes, we asked whether the set of promoters containing a strong motif match to each TF (FIMO
414 P-value < 10^{-4} in the region -500 to +100 relative to TSS) overlapped significantly with any
415 tissue expression (Spencer et al. 2010), GO categories (Ashburner et al. 2000), or KEGG
416 pathways (Kanehisa et al. 2014) (Fisher's exact test, one-sided probability, FDR < 0.05). We

417 obtained dozens of significant relationships (**Figure 9**), including known roles for GATA TFs in
418 the regulation of intestinal gene expression (and related GO categories) (Pauli et al. 2006;
419 McGhee 2007), HLH-1 in the regulation of muscle gene expression (Fukushige et al. 2006),
420 DAF-19 (an RFX TF) in the regulation of ciliary genes (Swoboda et al. 2000), and PHA-4 in
421 development of the pharynx (Gaudet and Mango 2002) (boxed in **Figure 9**). We also note that
422 the association of the motif for ZTF-19 (PAT-9), a C2H2 zinc finger protein, with genes
423 expressed in L2 body wall muscle tissue is consistent with observed expression patterns for this
424 gene in body wall muscle, as well as defective muscle development in a mutant (Liu et al. 2012).
425 The ZTF-19 binding motif may therefore enable identification of specific downstream targets.
426 Most of the associations in **Figure 9**, however, appear to represent potentially undocumented
427 regulatory interactions, suggesting that the motif collection can be used to gain new biological
428 insight.

429 **DISCUSSION**

430 The collection of motifs described here will further advance *C. elegans* as a major model system
431 for the study of gene regulation. TF DNA-binding motifs enable dissection of promoters,
432 prediction of new targets of TFs, and identification of putative new regulatory mechanisms.
433 Statistical associations between motif matches in promoters and expression patterns or functional
434 categories of genes also provide a ready starting point for directed experimentation; for example,
435 analysis of gene expression in mutants. Apparent position and orientation constraints between
436 motif matches also suggest functional relationships. Our observation that the largely unstudied
437 bZIP TF F23F12.9 (ZIP-8) was involved in almost half of all CMs identified in this study
438 suggests that it may function as a cofactor for targeting to open chromatin: pioneer TF activity
439 and partnering with other TFs has previously been proposed for other Creb/ATF proteins in
440 mouse embryonic stem cells (Sherwood et al. 2014).

441 A key observation in this study is that all of the large groups of TFs in *C. elegans* are malleable
442 in their DNA binding sequence preferences. The NHR is a striking case, even more so when we
443 consider that our motifs encompass only monomeric binding sites. A previous analysis classified
444 the *C. elegans* NHRs into four subtypes, on the basis of their recognition helix sequences, and
445 predicted that all of those in Class I (those most similar to canonical NHRs such as the Estrogen
446 Receptor) would likely bind canonical SHRE GGTCA subsites (Van Gilst et al. 2002). Instead,
447 we find that those in Class I (the entire lower half in **Figure 4**) bind a wide range of sequences,
448 and that the recognition helix cannot be the only determinant of sequence specificity. Our
449 observations are consistent with the previous demonstration that mutation of one or a few
450 residues in the NHR recognition helix can result in dramatic changes in sequence preferences,
451 but that mutations elsewhere in the DBD play a role in sequence selectivity (e.g. (McKeown et

452 al. 2014)). Recent analyses of other DBD classes (e.g. C2H2 and Forkhead) also highlight the
453 importance of residues beside the canonical specificity residues (Nakagawa et al. 2013; Siggers
454 et al. 2014). Together, these analyses strongly confirm that alteration of binding motifs is
455 widespread among TF classes throughout evolution.

456 Our study experimentally determined motifs for 129 TFs, all but five of which were previously
457 unstudied, bringing the total number of *C. elegans* TFs with motifs to 292 (including predicted
458 motifs and data already in the literature). We estimate that the remaining 453 *C. elegans* TFs
459 encode as many as 409 different DNA binding motifs, most of which correspond to NHRs,
460 C2H2 ZFs, bHLHs, and Homeodomains (**Supplemental File 3**). Additional effort will thus be
461 required to obtain a complete motif collection. For instance, even with our new motifs, and
462 including motifs predicted by homology, coverage for the *C. elegans* NHR family is only 17%.

463 For some classes of DBDs, most of the PBM assays yielded negative data. The NHRs in
464 particular yielded only 20% success (27/135). In addition, of the 108 that failed by PBM, we
465 have tested 100 by Y1H, of which only 17 succeeded (three or more detected interactions; data
466 not shown). We observed no obvious property of their DNA contacting residues that strongly
467 predicts success or failure and hypothesize that requirement for ligand binding, dimerization,
468 cofactors, or protein modifications may represent other potential explanations for failures in
469 heterologous assays. Like human NHRs, the *C. elegans* NHRs have a ligand-binding domain
470 that is distinct from the DNA-binding domain, and is thought to primarily regulate interactions
471 with coactivators and corepressors (Sonoda et al. 2008). Thus, ligand-dependent DNA binding
472 seems unlikely, especially for an *in vitro* assay. Yeast two-hybrid screens have identified several
473 interactions between different NHRs (Simonis et al. 2009; Reece-Hoyes et al. 2013), suggesting
474 that heterodimerization may be prevalent. If DNA binding is often dependent on

475 heterodimerization, then ChIP-seq should often succeed where PBMs and Y1H fail, and
476 heterodimeric motifs should be identified. To our knowledge, however, there is only one
477 published ChIP-seq data set for a *C. elegans* NHR that has yielded a motif *de novo* (NHR-25)
478 (Araya et al. 2014; Boyle et al. 2014). We did not test NHR-25 by PBM, although our predicted
479 monomeric motif (from Drosophila Ftz-f1) resembles the motif identified by ChIP-seq. As noted
480 above, our motif for DAF-12 is also consistent with the motif obtained by ChIP-chip (Hochbaum
481 et al. 2011). In support of heteromeric binding, however, CMs involving NHRs and other TF
482 families were prevalent in modENCODE ChIP-seq data (**Figure 8–source data 1**). Further
483 analysis will be required to explore the role of multimerization in *C. elegans* TF DNA binding
484 and gene regulation.

485 Finally, we note that there are numerous similarities between the TF collections of human,
486 *Drosophila*, and *C. elegans*. First, the total number TFs containing a canonical DBD varies only
487 by a factor of ~2 (~1734, 701, and 744 for human, *Drosophila*, and *C. elegans*, respectively
488 (Weirauch et al. 2014)). The total number of groups of closely related paralogs (taken using the
489 thresholds established in (Weirauch et al. 2014)), which should approximate the number of
490 distinct motifs that can eventually be expected, also varies by only a factor of ~2 (the TFs fall
491 into 1339, 656, and 632 groups of proteins expected to have very similar sequence specificity,
492 respectively). Our study also brings the proportion of *C. elegans* TFs with a known or predicted
493 motif much closer to the proportion in human and *Drosophila*, (56.1%, 54.1%, and 39.2% for
494 human, *Drosophila*, and *C. elegans*, respectively). In addition, all three species possess a large
495 number of diverse lineage-specific TFs, which are known or expected to bind different motifs:
496 in human, ~700 C2H2 ZF TFs; in Drosophila, 230 C2H2 ZF TFs, and in *C. elegans*, 266 NHRs,
497 99 C2H2 ZF TFs and – as we show here – roughly a dozen T-box and DM TFs. Thus, despite

498 widespread conservation in TF number and gene expression (Stuart et al. 2003) among
499 metazoans, extensive rewiring of the metazoan *trans* regulatory network is apparently common.

500

501 **MATERIALS AND METHODS**

502 **Selection of TFs for analysis.** We compiled a list of 874 known and putative *C. elegans* TFs.
503 We took 740 from build 0.90 of the CisBP database (Weirauch et al. 2014), plus additional
504 candidate TFs that lack canonical DBDs (Reece-Hoyes et al. 2005). We considered selecting
505 each TF for characterization using the following criteria:

506 1. Include the TF if its characterization would provide multiple motif predictions for
507 other TFs, based on established prediction thresholds for the given DBD class
508 (Weirauch et al. 2014);
509 2. Include the TF if it is a member of a DBD class with relatively little available motif
510 information;
511 3. Include the TF if it has a known important biological role;
512 4. Exclude the TF if it has already been characterized by PBM in another study;
513 5. Exclude the TF if it is a member of a DBD class with a low PBM success rate;
514 6. Exclude the TF if the resulting construct would be excessively long (for example,
515 exclude C2H2 ZF TFs with many DBDs)

516 **Cloning of *C. elegans* TF DBDs.** We identified putative DBDs for all TFs by scanning their
517 protein sequences using the HMMER tool (Eddy 2009), and a collection of 81 Pfam (Finn et al.
518 2010) models taken from (Weirauch and Hughes 2011), as described previously (Weirauch et al.
519 2014). For some TFs, we could not identify DBDs using this procedure. In such cases, DBDs
520 were manually detected by lowering HMMER scanning thresholds, using DBDs annotated in the
521 SMART database (Letunic et al. 2012), or performing literature searches. Using the above
522 criteria for selection of TFs, we initially chose 398 TFs from the Walhout clone collection for

523 characterization. We designed primers (**Supplementary File 4**) to clone open reading frames
524 (ORFs) comprising the DBDs plus additional flanking sequences (50 endogenous amino acid
525 flanking residues, or until the end of the protein). We inserted the resulting sequences using *Asc*I
526 and *Sbf*I restriction sites into a modified T7-driven expression vector (pTH6838) that expresses
527 N-terminal GST fusion proteins (**Supplementary File 5**). In a first round of cloning we
528 attempted cloning using both individual plasmids and pooled mRNA (by RT-PCR) or cDNA.
529 After PBM analysis with the resulting clones, we then considered remaining uncharacterized
530 TFs, and selected an additional 154 TFs using the same criteria as above. The DBDs and
531 flanking bases of these TFs were created using gene synthesis (BioBasic), and inserted into
532 vectors as described above. Primers and insert sequences are provided on our project web site.
533 All clones were sequence verified.

534 **PBMs and data processing.** PBM laboratory methods were identical to those described
535 previously (Lam et al. 2011; Weirauch et al. 2013). Each plasmid was analyzed in duplicate on
536 two different arrays with differing probe sequences. Microarray data were processed by
537 removing spots flagged as 'bad' or 'suspect', and employing spatial de-trending (using a 7x7
538 window centered on each spot) (Weirauch et al. 2013). Calculation of 8-mer Z- and E-scores
539 was performed as previously described (Berger et al. 2006). Z-scores are derived by taking the
540 average spot intensity for each probe containing the 8-mer, then subtracting the median value for
541 each 8-mer, and dividing by the standard deviation, thus yielding a distribution with a median of
542 zero and a standard deviation of one. E-scores are a modified version of the AUROC statistic,
543 which consider the relative ranking of probes containing a given 8-mer, and range from -0.5 to
544 +0.5, with E>0.45 taken as highly statistically significant (Berger et al. 2008). We deemed
545 experiments successful if at least one 8-mer had an E-score > 0.45 on both arrays, the

546 complimentary arrays produced highly correlated E- and Z-scores, and the complimentary arrays
547 yielded similar PWMs based on the PWM_align algorithm (Weirauch et al. 2013).

548 **Generation of PWMs from PBMs.** Motif derivation followed steps as outlined previously
549 (Weirauch et al. 2014). Briefly, to obtain a single representative motif for each protein, we
550 generated motifs for each array using four different algorithms: BEEML-PBM (Zhao and Stormo
551 2011), FeatureREDUCE (manuscript in prep, source code available at
552 http://rileylab.bio.umb.edu/content/software), PWM_align (Weirauch et al. 2013), and
553 PWM_align_Z (Ray et al. 2013). We scored each motif on the complimentary array using the
554 energy scoring system utilized by the BEEML-PBM algorithm (Zhao and Stormo 2011). We
555 then compared these PWM-based probe score predictions with the actual probe intensities using
556 (1) the Pearson correlation coefficient (PCC) and (2) the AUROC of "bright probes" (defined by
557 transforming all probe intensities to Z-scores, and selecting probes with Z-scores >= 4),
558 following (Weirauch et al. 2013). Finally, we chose a single PWM for each DBD construct
559 using these two criteria, as previously described (Weirauch et al. 2014).

560 **Expert curation.** For every TF with motif information we selected a representative motif, or
561 motifs if that TF appears to have multiple binding modes (e.g multimers), using the following
562 scheme. If the TF has an experimentally derived motif it is selected as the primary motif. If
563 there are multiple such motifs we selected one that was derived *in vitro*, if any. If the TF had
564 multiple *in vitro* motifs, then we ranked PBM>B1H>SELEX, to maximize comparability among
565 motifs, and excluded motifs that are inconsistent with known motifs for the same or highly
566 related proteins. If the TF had only predicted motifs, we selected a motif from a highly similar
567 TF that is: preferably derived from an *in vitro* method (PBM>B1H>SELEX); assigned to the

568 cluster that contains the majority of motifs for that TF in our PWMclus analysis; consistent with
569 known DBD preferences; and best supported by ChIP-seq or Y1H data, if available.

570 **Motif enrichment with Y1H and ChIP-seq data.** We calculated motif enrichment in ChIP-seq
571 peaks using CentriMo (Bailey and Machanick 2012), which uses TF motifs to look for central
572 enrichment of motifs in ChIP-seq peaks, as an indication of direct binding by that TF. We
573 obtained ChIP-seq peaks from the *C. elegans* modENCODE consortium (Araya et al. 2014). We
574 used the top 250 peaks ranked by Irreproducible Discovery Rate (Landt et al. 2012) as the input
575 datasets. We scored the curated set of motifs for TFs with peak datasets across all the peaks.
576 We report false discovery rate (FDR)-adjusted p-values for a motif's central enrichment in TF
577 peak datasets.

578 For yeast one-hybrid (Y1H) data, we assigned motif scores to promoter bait sequences using the
579 BEEML scoring system (Zhao et al. 2009). We included TFs in the analysis only if they bound
580 five or more promoters in Y1H (those with 3 or 4 promoters bound were excluded to minimize
581 sampling error in Mann-Whitney tests). We scored only the promoter-proximal 500 bp of Y1H
582 bait sequences, as activating TF binding sites are mainly effective within a few hundred bases of
583 TSS in *S. cerevisiae* (Dobi and Winston 2007). We calculated motif enrichment or depletion for
584 motifs using a two-tailed Mann–Whitney *U* test and reported with FDR-corrected p-values, with
585 Y1H interactors as positives and the remaining non-interacting baits as the background.

586 We performed composite motif (CM) analysis by scanning 77 *C. elegans* ChIP-seq top 250 peak
587 sequences for all pairwise combinations between the 40 ChIPed TFs (using the curated list of
588 PWMs) and 129 PBM-derived PWMs from this study. Relative PWM spacing was restricted to -
589 5 (overlapping) to +10 (gapped) bp separation, with four possible stereospecific arrangements of

590 TFs: TF-1 forward TF-2 forward (1F2F), TF-1 forward TF-2 reverse (1F2R), 2R1F, and 2F1F,
591 yielding 64 stereospecific combinations. We identified sequence matches using the standard log-
592 likelihood scoring framework (Stormo 2000), with a threshold of 0.50*max_score for each
593 PWM, where max_score is the highest possible score for the given PWM. We created 10 sets of
594 background sequences by scrambling the input sequences (maintaining dinucleotide
595 frequencies). We calculated sample z-scores and p-values by comparing the number of sequence
596 matches observed in the "real" sequence to the number observed in the random sequences, and
597 applied a Bonferroni correction to each p-value. To identify significant composite motifs, we
598 filtered to retain only results with sequence match counts ≥ 10% of the number of input peak
599 sequences and Bonferroni-adjusted p-values ≤ 0.05 (alpha=0.05). We also considered an
600 alternative null model, in which we shuffled the non-ChIPed motif, and counted matches in the
601 original DNA sequences (this procedure was repeated 10x). Overall, we found very good
602 agreement using this approach and our original null model. Out of the 635,712 possible patterns
603 we tested, both methods call 635,483 insignificant, both call 49 positive, and they disagree on
604 180 (**Figure 8C**). **Figure 8D** plots the number of significant hits identified relative to
605 dinucleotide scrambled sequences using shuffled (blue) and non-shuffled (red) non-ChIPed
606 motifs. This plot indicates, however, that the shuffled motif null model over-estimates the
607 significance of CMs as the overlap of their constituent motifs increases, presumably due to
608 dispersal of high information content "core" positions, which are typically adjacent in the real
609 motifs. We therefore use and report results based only on null model 1. Sequence logos were
610 constructed using the actual matches obtained in the ChIP-seq peak sequences, and the WebLogo
611 3.4 tool (Crooks et al. 2004). For each TF family *F*, we calculated an odds ratio (*OR*) comparing
612 the ratio of families in CMs to the ratio of families in the motif list. We define *OR* as *(a/b)/(c/d)*,

613 where a is the number of TFs of family F involved in a CM; b is the total number of unique TF
614 pairs involved in a CM, minus a; c is the number of TFs of family F in the motif list; and d is the
615 total number of TFs in the motif list, minus c. We calculated the standard error (SE) as
616 $\sqrt{(1/a+1/b+1/c+1/d)}$, and the 95% confidence interval as $e^{ln(OR)\pm 1.96SE}$.

617 **Motif enrichment in co-regulated tissue/developmental stage-specific genes, KEGG and GO**
618 **categories.** We obtained selectively enriched gene sets for each tissue from
619 (http://www.vanderbilt.edu/wormdoc/wormmap/) GO annotations from
620 (http://www.geneontology.org/) and KEGG pathway modules
621 (http://www.genome.jp/kegg/module.html). We ran FIMO (Grant et al. 2011) with default
622 parameters.

623 **DATA ACCESS.** PBM microarray data are available at GEO (www.ncbi.nlm.nih.gov/geo/)
624 under accession number GSE65719. Motifs and 8-mer data (E- and Z-scores) are available at
625 www.cisbp.ccbr.utoronto.ca. Supplementary data files, including plasmid and primer
626 information, motifs, source data for figure heatmaps, and lists of TFs are found on our project
627 web site http://hugheslab.ccbr.utoronto.ca/supplementary-data/CeMotifs/.

628 **ACKNOWLEDGEMENTS**. This work was supported by CIHR Operating grants MOP-
629 111007 and MOP-77721 to T.R.H. S.A.L. is supported by an Ontario Graduate Scholarship.
630 H.S.N. is supported by a CIHR Banting Fellowship. We are grateful to Andy Fraser, Susan
631 Mango, Bob Waterston, Carlos Araya, and Don Moerman for helpful discussions and materials.

632 **AUTHOR CONTRIBUTIONS**. KN, AY, SM, HZ, and SL performed cloning and PBM
633 analyses. KN and SL performed motif curation and carried out the bulk of the computational
634 analyses. JRR and MTW performed the multimer analysis of ChIP-seq data. MTW and MA

635 performed motif generation and derivation, and integrated the motif collection into the CisBP
636 database. HSN performed the C2H2 recognition code analysis. JSRH and AJMW contributed
637 clones and, with JIFB, provided Y1H data. TRH conceived of the study and coordinated the
638 analyses. KN, SL, and TRH wrote the manuscript with significant input from JSRH, JAFB,
639 AJMW, and MTW.

640 **FIGURE AND FIGURE SUPPLEMENT LEGENDS**

641 **Figure 1. Motif status by DBD class.** Stacked bar plot depicting the number of unique *C.*
642 *elegans* TFs for which a motif has been derived using PBM (this study), previous literature
643 (including PBMs), or by homology-based prediction rules (see main text). The y-axis is
644 displayed on a $\log_2$ scale for values greater than zero. See **Figure 1-source data 1** for DBD
645 abbreviations. Correspondence between motifs identified in current study and previously
646 reported motifs are shown in **Figure 1-figure supplement 1**.

647

648 **Figure 1-figure supplement 1. Correspondence between TF motifs identified from our PBM**
649 **study and previously reported motifs from several types of experimental data.**

650

651 **Figure 2. Motif prediction, motif clustering, and identification of representative motifs. (A-**
652 **C),** Boxplots depict the relationship between the %ID of aligned AAs and % of shared 8-mer
653 DNA sequences with E-scores exceeding 0.45, for the three DBD classes, as indicated. %ID bins
654 range from 0 to 100, of size 10, in increments of five. Red dots indicate individual TFs in this
655 study, vs. the next closest TF with PBM data. Vertical lines indicate AA %ID threshold above
656 which motifs can be predicted using homology, taken from (Weirauch et al. 2014). Boxplots for
657 all other DBDs in current study are shown in **Figure 2 - figure supplements 1-4**. **(D)** Clustering
658 analysis of motifs of bZIP domains using PWMclus (Jiang and Singh 2014). Coloured gridlines
659 indicate clusters. Cluster centroids are shown along the diagonal; expert curated motifs are
660 shown within the box at right. 'E' indicates experimentally determined motifs; 'P' indicates
661 predicted motifs. Source of motif is also indicated. Results of motif curation for GATA family
662 TFs is displayed in **Figure 2-figure supplement 5**.

663

664 **Figure 2-figure supplements 1-4. *C. elegans* TFs adhere to established thresholds for motif**
665 **inference.** Boxplots depict the relationship between the %ID of aligned AAs and % of shared 8-
666 mer DNA sequences with E-scores exceeding 0.45, for the DBD classes of TFs with PBMs from
667 this study. %ID bins range from 0 to 100, of size 10, in increments of five. Red dots indicate

individual proteins in this study, vs. the next closest protein with PBM data. Vertical blue lines indicate AA %ID threshold above which motifs can be predicted using homology.

Figure 2-figure supplement 5. GATA TF motif clustering and identification of representative motifs. Clustering analysis of *C.elegans* GATA TF's motifs using PWMclus (Jiang and Singh 2014). Coloured gridlines indicate clusters. Cluster centroids are shown along the diagonal, while manually curated motifs are shown within the box at right. Bolded row names represent motifs obtained from this study.

Figure 3. Overview of 8-mer sequences preferences for the 129 *C. elegans* TFs analyzed by PBM in this study. 2-D Hierarchical agglomerative clustering analysis of E-scores performed on all 5,728 8-mers bound by at least one TF (average E > 0.45 between ME and HK replicate PBMs). Coloured boxes represent DBD classes for each TF.

Figure 4. 8-mer binding profiles of NHR family reveal distinct sequence preferences. *Left*, ClustalW phylogram of NHR DBD amino acid sequences with corresponding motifs. TF labels are shaded according to motif similarity groups identified by PWMclus. *Center*, Heatmap showing E-scores. NHRs are ordered according to the phylogram at left. The 1,406 8-mers with E-score > 0.45 for at least one family member on at least one PBM array were ordered using hierarchical agglomerative clustering. Each TF has one row for each of two replicate PBM experiments (ME or HK array designs). *Right*; recognition helix (RH) sequences for the corresponding proteins, with identical RH sequence types highlighted by colored asterisks. Variant RH residues are underlined at bottom. *Right*, matrix indicates cluster membership according to PWMclus. *Top and bottom.* Pullouts show re-clustered data including only the union of the top ten most highly scoring 8-mers (taking the average E-score from the ME and HK arrays) for each of the selected proteins.

695 **Figure 5. C2H2 motifs relate to DBD similarity and to the recognition code.** *Left*, ClustalW
696 phylogram of C2H2 ZF amino acid sequences with corresponding motifs. *Right*, examples in
697 which motifs predicted by the ZF recognition code are compared to changes in DNA sequences
698 preferred by paralogous C2H2 ZF TFs. Cartoon shows individual C2H2 ZFs and their specificity
699 residues. Dashed lines correspond to 4-base subsites predicted from the recognition code.

700

701 **Figure 5-figure supplement 1. Comparison of C2H2 zinc finger recognition model with**
702 **motifs derived PBM.** Motif correlations between PBM derived motifs and ZF-model based
703 predictions for TFs with both typical and atypical **(A)** linker lengths between ZF modules that
704 are longer than 6 amino acids or shorter than 4 amino acids **(B)** Zinc coordinating cysteine or
705 histidine structural motifs and **(C)** differing length of the ZF array. Examples of recognition code
706 predictions (sequence logos) for both typical and atypical TFs are compared with PBM motifs
707 for each case. The *p*-values shown are estimated from Student's *t*-test. The number of TFs in
708 each boxplot is shown above in parentheses.

709

710 **Figure 6. Nematode-specific sequence preferences in T-box and DM TFs.** PBM data
711 heatmaps of preferred 8-mers for T-box **(A)**, and DM **(B)** TFs. TFs are clustered using ClustalW;
712 8-mers were selected (at least one instance of E>0.45) and clustered using hierarchical
713 agglomerative clustering, as in Figures 4 and 5. Ten representative 8-mers (those with highest E-
714 scores) are shown below for each of the clusters indicated in cyan. *C. elegans* TFs with data from
715 this study are bolded.

716

717 **Figure 6-figure supplement 1. T-box sequence alignments and the crystal structure of**
718 **mTBX3 illustrate C. elegans specific variations. (A)** Multiple sequence alignment of T-box
719 DBDs from *C. elegans*, the protist *C. owczarzaki* CoBra and mouse Eomes and TBX3. Key
720 DNA binding residues identified from crystal structure of mTBX3 are highlighted in red.
721 Sequence insertions (for TBX-33), changes in the variable region (for TBX-39/40) and
722 significant sequence changes in the key 3_{10}C recognition helix (for TBX- 39/40) are highlighted

723 in blue frames. **(B)** Crystal structure model of mTBX3 is used as a prototype to illustrate *C.*
724 *elegans* specific sequence variations. The primary recognition helix, 3_{10}C, is highlighted in
725 yellow.

726

727 **Figure 7. The *C. elegans* curated motif collection explains ChIP-seq and Y1H TF binding**
728 **data.** Heatmap of CentriMo -$\log_{10}$(*q-values*) for central enrichment of TF motifs in the top 250
729 peaks for each ChIP experiment. Motif enrichment in Y1H data is presented in **Figure 7-figure**
730 **supplement 1.** Heatmaps are symmetric with duplicate rows to ensure the diagonal represents
731 TF motif enrichment in it's matching dataset(s). Red and blue colouring depicts statistically
732 significant enrichments and depletions (q <= 0.05).

733

734 **Figure 7-figure supplement 1. The *C. elegans* curated motif collection explains Y1H TF**
735 **binding data.** Heatmap depicting enrichment or depletion (Mann-Whitney U test) of TF motifs
736 in the interactions of TF's with a collection of promoter bait sequences in Y1H experiments
737 compared to all non-interacting bait sequences.

738

739 **Figure 8. Composite motifs enriched in *C. elegans* ChIP-seq peaks. (A)** Stereospecificity
740 plots showing enriched CM configurations for pairs of TF motifs. The identical "1F2F" and
741 "2F1F" results in A (top row) demonstrate homodimer and homotrimer CMs, while those
742 involving LSY-2, NHR-232, and R07H5.10 demonstrate heterodimer and heterotrimer CMs
743 (middle and bottom rows, respectively). Black arrows represent orientation of the motif within
744 CMs, while gray dashed arrows designate shadow motifs within trimeric CMs. Error bars are
745 ±S.D., *corrected $p < 0.05$. **(B)** Forest plot of odds ratios for TF family enrichment in CMs vs.
746 input TF list. **(C)** Venn diagram showing overlap of significant CMs identified by null model 1
747 (dinucleotide shuffled sequence) and null model 2 (motif shuffling). **(D)** Number of significant
748 CMs identified relative to dinucleotide scrambled sequences using shuffled and non-shuffled
749 non-ChIPed motifs, as a function of motif pair distance.

Figure 8-figure supplements 1-5. Summary of clustered CMs enriched in *C. elegans* ChIP-seq peaks. CM cluster centroids are shown for the enriched motifs. For each cluster the ChIPed TF(s) and potential partner TF(s) are listed along with information about motif overlap (OLAP), spacing (GAP) and enrichment. Coloured arrows over motif indicate high information-content portions of either factor.

Figure 9. Enrichment of motifs upstream of gene sets. Each row of the heatmap represents a motif from our curated collection that is enriched ($q < 0.05$) in at least one gene set category. Known regulatory interactions between TFs and gene sets are highlighted (black outlines). 'E' indicates experimentally determined motifs; 'P' indicates predicted motifs. Source of motif is also indicated.

SOURCE DATA AND SUPPLEMENTARY FILE LEGENDS

Figure 1-source data 1. Table of *C. elegans* TF repertoire motif coverage and list of TF DBDs present in *C. elegans*. The number of unique *C. elegans* TFs by DNA-binding domain family for which a motif has been derived using PBM (this study), previous literature (including PBMs), or by homology-based prediction rules and the list of *C. elegans* TFs by DNA-binding domain family type.

Figure 3-source data 1. Table showing 8-mers bound by at least one TF with an E-score >=0.45 for all the 129 *C. elegans* TFs analyzed by PBMs in this study.

Figure 4- source data 1. Table showing 8-mer E-score profiles of NHRs analyzed by PBMs. 8-mers bound by at least one NHR with an E-score >=0.45 for all the *C. elegans* NHRs that have been analyzed by PBMs (center panel) and a table of pullouts (top and bottom panel) showing average (ME and HK) E-scores of the union of the top ten highly scoring 8-mers bound by at least one NHR within the selected motif cluster.

Figure 6-source data 6. Table showing 8-mer E-score profiles of T-box and DM TFs from *C. elegans* and other metazoans that have been analyzed by PBMs.

Figure 9-source data 1. Table of motif enrichments –$\log_{10}$(p-values) in the promoters of gene set categories identified from KEGG pathway modules, Gene Ontology processes and tissue/developmental stage specific expression lists.

Supplementary File 1. Comparison of CisBP TF collection with wTF2.0. Includes comments of overlaps and differences between two lists and whether each entry is likely a *bona fide* TF.

Supplementary File 2. *C. elegans* curated motif collection. This spreadsheet contains the curated motif IDs for each *C. elegans* TF along with their source and experimental support.

Supplementary File 3. Number of experiments required for complete coverage of human, fly and worm TF collections. This spreadsheet contains numbers of experiments needed for each DBD class to have complete coverage of the motif collection based on previously described DBD prediction thresholds.

Supplementary File 4. List of primers and gene systhesis constructs used to obtain TF clones in this study. This spreadsheet contains primers used to clone TFs as well as gene synthesis constructs that were cloned in to the PBM plasmid backbone (**Supplementary File 5**).

Supplementary File 5. PBM plasmid (pTH6838) backbone map. Information on the expression vector used in PBM experiments.

802

803 **REFERENCES**

804 Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE. 2004. Environmentally induced foregut remodeling by
805 PHA-4/FoxA and DAF-12/NHR. *Science* **305**(5691): 1743-1746.
806 Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D et
807 al. 2014. Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature*
808 **512**(7515): 400-405.
809 Arda HE, Taubert S, MacNeil LT, Conine CC, Tsuda B, Van Gilst M, Sequerra R, Doucette-Stamm L,
810 Yamamoto KR, Walhout AJ. 2010. Functional modularity of nuclear hormone receptors in a
811 Caenorhabditis elegans metabolic gene regulatory network. *Mol Syst Biol* **6**: 367.
812 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT
813 et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
814 *Nat Genet* **25**(1): 25-29.
815 Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A,
816 Chen X et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science*
817 **324**(5935): 1720-1723.
818 Bailey TL, Machanick P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40**(17):
819 e128.
820 Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* **117**(2): 185-198.
821 Benizri E, Ginouves A, Berra E. 2008. The magic of the hypoxia-signaling cascade. *Cell Mol Life Sci* **65**(7-
822 8): 1133-1149.
823 Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S,
824 Botvinnik OB, Chan ET et al. 2008. Variation in homeodomain DNA binding revealed by high-
825 resolution analysis of sequence preferences. *Cell* **133**(7): 1266-1276.
826 Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. 2006. Compact, universal DNA
827 microarrays to comprehensively determine transcription-factor binding site specificities. *Nat*
828 *Biotechnol* **24**(11): 1429-1435.
829 Blanchet E, Annicotte JS, Lagarrigue S, Aguilar V, Clape C, Chavey C, Fritz V, Casas F, Apparailly F, Auwerx
830 J et al. 2011. E2F transcription factor-1 regulates oxidative metabolism. *Nature cell biology*
831 **13**(9): 1146-1152.
832 Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L et al.
833 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature*
834 **512**(7515): 453-456.
835 C.elegans consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating
836 biology. *Science* **282**(5396): 2012-2018.
837 Coll M, Seidman JG, Muller CW. 2002. Structure of the DNA-bound T-box domain of human TBX3, a
838 transcription factor responsible for ulnar-mammary syndrome. *Structure* **10**(3): 343-356.
839 Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*
840 **14**(6): 1188-1190.
841 de Boer CG, Hughes TR. 2011. YeTFaSCo: a database of evaluated yeast transcription factor sequence
842 specificities. *Nucleic Acids Res*.
843 Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm
844 L, Reece-Hoyes JS, Hope IA et al. 2006. A gene-centered C. elegans protein-DNA interaction
845 network. *Cell* **125**(6): 1193-1205.
846 Dobi KC, Winston F. 2007. Analysis of transcriptional activation at a distance in Saccharomyces
847 cerevisiae. *Mol Cell Biol* **27**(15): 5575-5586.

848 Dupuy D, Li QR, Deplancke B, Boxem M, Hao T, Lamesch P, Sequerra R, Bosak S, Doucette-Stamm L,
849 Hope IA et al. 2004. A first version of the Caenorhabditis elegans Promoterome. *Genome*
850 *research* **14**(10B): 2169-2175.
851 Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome*
852 *informatics International Conference on Genome Informatics* **23**(1): 205-211.
853 Enmark E, Gustafsson JA. 2001. Comparing nuclear receptors in worms, flies and humans. *Trends in*
854 *pharmacological sciences* **22**(12): 611-615.
855 Evan G, Harrington E, Fanidi A, Land H, Amati B, Bennett M. 1994. Integrated control of cell proliferation
856 and cell death by the c-myc oncogene. *Philosophical transactions of the Royal Society of London*
857 **345**(1313): 269-275.
858 Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K
859 et al. 2010. The Pfam protein families database. *Nucleic Acids Res* **38**(Database issue): D211-222.
860 Fukushige T, Brodigan TM, Schriefer LA, Waterston RH, Krause M. 2006. Defining the transcriptional
861 redundancy of early bodywall muscle development in C. elegans: evidence for a unified theory
862 of animal muscle development. *Genes Dev* **20**(24): 3395-3406.
863 Gamble T, Zarkower D. 2012. Sex determination. *Current biology : CB* **22**(8): R257-262.
864 Gaudet J, Mango SE. 2002. Regulation of organogenesis by the Caenorhabditis elegans FoxA protein
865 PHA-4. *Science* **295**(5556): 821-825.
866 Gaudet J, McGhee JD. 2010. Recent advances in understanding the molecular mechanisms regulating C.
867 elegans transcription. *Dev Dyn* **239**(5): 1388-1404.
868 Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami
869 K et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE
870 project. *Science* **330**(6012): 1775-1787.
871 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*
872 **27**(7): 1017-1018.
873 Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, Walhout AJ. 2009. A
874 multiparameter network reveals extensive divergence between C. elegans bHLH transcription
875 factors. *Cell* **138**(2): 314-327.
876 Haerty W, Artieri C, Khezri N, Singh RS, Gupta BP. 2008. Comparative analysis of function and interaction
877 of transcription factors in nematodes: extensive conservation of orthology coupled to rapid
878 sequence evolution. *BMC genomics* **9**: 399.
879 Hochbaum D, Zhang Y, Stuckenholz C, Labhart P, Alexiadis V, Martin R, Knolker HJ, Fisher AL. 2011. DAF-
880 12 regulates a connected network of genes to ensure robust developmental decisions. *PLoS*
881 *genetics* **7**(7): e1002179.
882 Jiang P, Singh M. 2014. CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic*
883 *acids research* **42**(5): 2833-2847.
884 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et
885 al. 2013. DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**(1-2): 327-339.
886 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge
887 and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**(Database issue): D199-205.
888 Lam KN, van Bakel H, Cote AG, van der Ven A, Hughes TR. 2011. Sequence specificity is obtained from
889 the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res* **39**(11): 4680-4690.
890 Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB,
891 Cayting P et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE
892 consortia. *Genome Res* **22**(9): 1813-1831.
893 Lemons D, McGinnis W. 2006. Genomic evolution of Hox gene clusters. *Science* **313**(5795): 1918-1922.
894 Lesch BJ, Gehrke AR, Bulyk ML, Bargmann CI. 2009. Transcriptional regulation and stabilization of left-
895 right neuronal identity in C. elegans. *Genes Dev* **23**(3): 345-358.

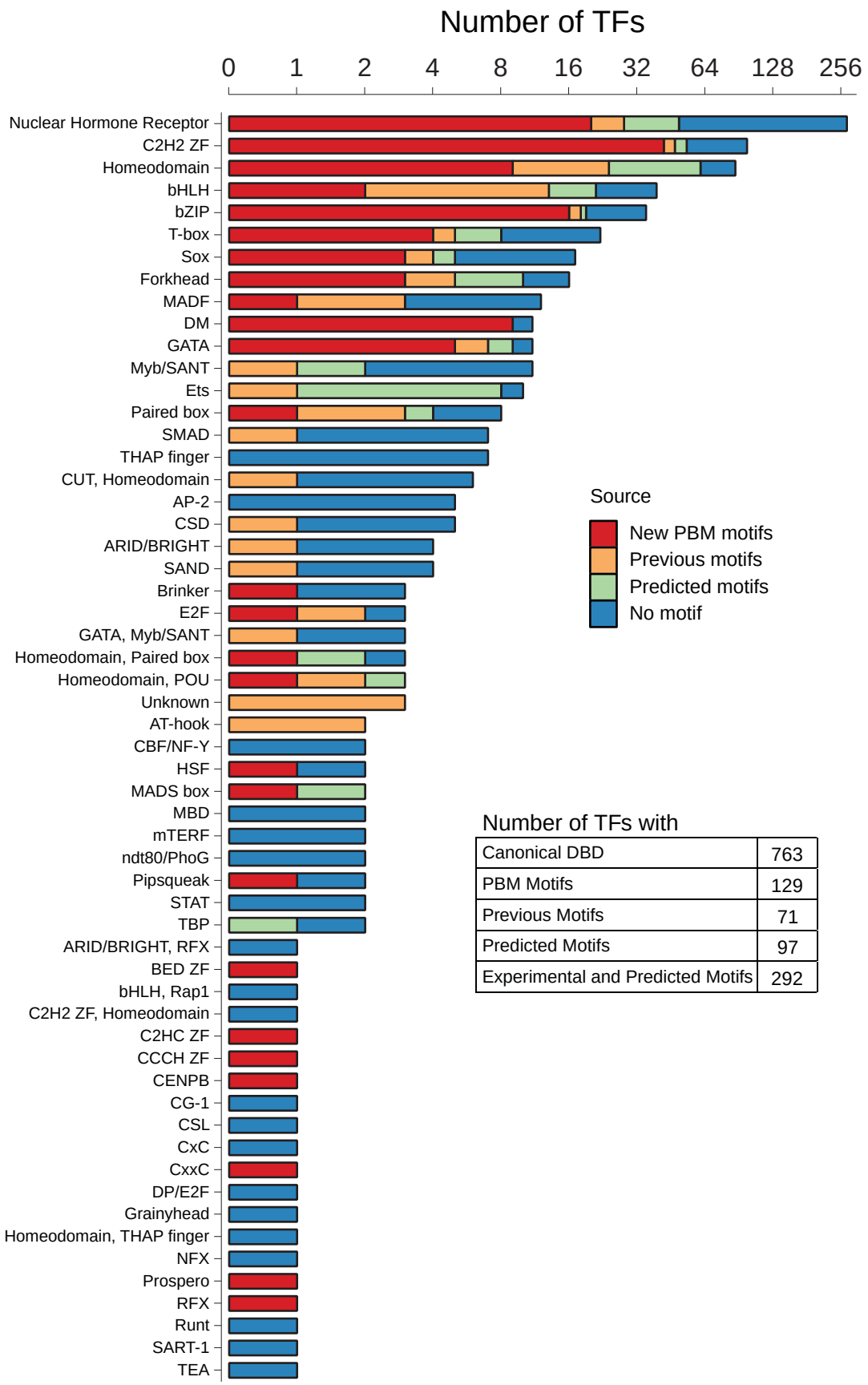
896 Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource.
897 *Nucleic acids research* **40**(Database issue): D302-305.
898 Liu Q, Jones TI, Bachmann RA, Meghpara M, Rogowski L, Williams BD, Jones PL. 2012. C. elegans PAT-9 is
899 a nuclear zinc finger protein critical for the assembly of muscle attachments. *Cell & bioscience*
900 **2**(1): 18.
901 Liu X, Lee CK, Granek JA, Clarke ND, Lieb JD. 2006. Whole-genome comparison of Leu3 binding in vitro
902 and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome*
903 *Res* **16**(12): 1517-1528.
904 Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A,
905 Ienasescu H et al. 2014. JASPAR 2014: an extensively expanded and updated open-access
906 database of transcription factor binding profiles. *Nucleic Acids Res* **42**(Database issue): D142-
907 147.
908 Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M,
909 Hornischer K et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene
910 regulation in eukaryotes. *Nucleic Acids Res* **34**(Database issue): D108-110.
911 McGhee JD. 2007. The C. elegans intestine. *WormBook : the online review of C elegans biology*: 1-36.
912 McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. 2014. Evolution of
913 DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell*
914 **159**(1): 58-68.
915 Minguillon C, Logan M. 2003. The comparative genomics of T-box genes. *Briefings in functional genomics*
916 *& proteomics* **2**(3): 224-233.
917 Mintseris J, Eisen MB. 2006. Design of a combinatorial DNA microarray for protein-DNA interaction
918 studies. *BMC Bioinformatics* **7**: 429.
919 Muller CW, Herrmann BG. 1997. Crystallographic structure of the T domain-DNA complex of the
920 Brachyury transcription factor. *Nature* **389**(6653): 884-888.
921 Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani
922 E, Kim PM et al. 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon.
923 *Nat Biotechnol*.
924 Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. 2013. DNA-binding specificity changes in
925 the evolution of forkhead transcription factors. *Proc Natl Acad Sci U S A* **110**(30): 12349-12354.
926 Narendra U, Zhu L, Li B, Wilken J, Weiss MA. 2002. Sex-specific gene regulation. The Doublesex DM motif
927 is a bipartite DNA-binding domain. *J Biol Chem* **277**(45): 43463-43473.
928 Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, Janette J, Chen C, Alves P, Preston E et al. 2011.
929 Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans.
930 *Genome research* **21**(2): 245-254.
931 Pauli F, Liu Y, Kim YA, Chen PJ, Kim SK. 2006. Chromosomal clustering and GATA transcriptional
932 regulation of intestine-expressed genes in C. elegans. *Development* **133**(2): 287-295.
933 Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al.
934 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**(7457):
935 172-177.
936 Reece-Hoyes JS, Deplancke B, Barrasa MI, Hatzold J, Smit RB, Arda HE, Pope PA, Gaudet J, Conradt B,
937 Walhout AJ. 2009. The C. elegans Snail homolog CES-1 can activate gene expression in vivo and
938 share targets with bHLH transcription factors. *Nucleic Acids Res* **37**(11): 3689-3698.
939 Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout AJ. 2005. A compendium of
940 Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription
941 regulatory networks. *Genome Biol* **6**(13): R110.

942 Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S, Pesyna C, Dekker J, Myers CL,
943 Walhout AJ. 2011. Enhanced yeast one-hybrid assays for high-throughput gene-centered
944 regulatory network mapping. *Nat Methods* **8**(12): 1059-1064.
945 Reece-Hoyes JS, Pons C, Diallo A, Mori A, Shrestha S, Kadreppa S, Nelson J, Diprima S, Dricot A, Lajoie BR
946 et al. 2013. Extensive rewiring and complex evolutionary dynamics in a C. elegans
947 multiparameter transcription factor network. *Mol Cell* **51**(1): 116-127.
948 Reinke V, Krause M, Okkema P. 2013. Transcriptional regulation of gene expression in C. elegans.
949 *WormBook : the online review of C elegans biology*: 1-34.
950 Robinson-Rechavi M, Maina CV, Gissendanner CR, Laudet V, Sluder A. 2005. Explosive lineage-specific
951 expansion of the orphan nuclear receptor HNF4 in nematodes. *Journal of molecular evolution*
952 **60**(5): 577-586.
953 Sarov M, Murray JI, Schanze K, Pozniakovski A, Niu W, Angermann K, Hasse S, Rupprecht M, Vinis E,
954 Tinney M et al. 2012. A genome-scale resource for in vivo tag-based protein function exploration
955 in C. elegans. *Cell* **150**(4): 855-866.
956 Sebe-Pedros A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, Adamski M, Adamska M,
957 Hughes TR, Gomez-Skarmeta JL et al. 2013. Early evolution of the T-box transcription factor
958 family. *Proceedings of the National Academy of Sciences of the United States of America*
959 **110**(40): 16050-16055.
960 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford
961 DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling
962 DNase profile magnitude and shape. *Nat Biotechnol* **32**(2): 171-178.
963 Siggers T, Reddy J, Barron B, Bulyk ML. 2014. Diversification of transcription factor paralogs via
964 noncanonical modularity in C2H2 zinc finger DNA binding. *Mol Cell* **55**(4): 640-648.
965 Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, Hirozane-Kishikawa T, Hao T, Sahalie JM,
966 Venkatesan K, Gebreab F et al. 2009. Empirically controlled mapping of the Caenorhabditis
967 elegans protein-protein interactome network. *Nat Methods* **6**(1): 47-54.
968 Sleumer MC, Bilenky M, He A, Robertson G, Thiessen N, Jones SJ. 2009. Caenorhabditis elegans cisRED: a
969 catalogue of conserved genomic elements. *Nucleic acids research* **37**(4): 1323-1334.
970 Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D et al.
971 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape
972 cell-type identity. *Genome research* **21**(10): 1757-1767.
973 Sonoda J, Pei L, Evans RM. 2008. Nuclear receptors: decoding metabolic disease. *FEBS Lett* **582**(1): 2-9.
974 Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT,
975 Widmer C, Jo J et al. 2010. A spatial and temporal map of C. elegans gene expression. *Genome*
976 *Res* **21**(2): 325-341.
977 Stirnimann CU, Ptchelkine D, Grimm C, Muller CW. 2002. Structural basis of TBX5-DNA recognition: the
978 T-box domain in its DNA-bound and -unbound form. *Journal of molecular biology* **400**(1): 71-81.
979 Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**(1): 16-23.
980 Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of
981 conserved genetic modules. *Science* **302**(5643): 249-255.
982 Stubbs L, Sun Y, Caetano-Anolles D. 2011. Function and Evolution of C2H2 Zinc Finger Arrays. *Subcell*
983 *Biochem* **52**: 75-94.
984 Swoboda P, Adler HT, Thomas JH. 2000. The RFX-type transcription factor DAF-19 regulates sensory
985 neuron cilium formation in C. elegans. *Mol Cell* **5**(3): 411-421.
986 Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult
987 fibroblast cultures by defined factors. *Cell* **126**(4): 663-676.

988 Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. 2013. Highly expressed loci are vulnerable to
989 misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A* **110**(46):
990 18602-18607.
991 Van Gilst M, Gissendanner CR, Sluder AE. 2002. Diversity and function of orphan nuclear receptors in
992 nematodes. *Crit Rev Eukaryot Gene Expr* **12**(1): 65-88.
993 Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A,
994 Talukder S et al. 2013. Evaluation of methods for modeling transcription factor sequence
995 specificity. *Nat Biotechnol* **31**(2): 126-134.
996 Weirauch MT, Hughes TR. 2011. A catalogue of eukaryotic transcription factor types, their evolutionary
997 origin, and species distribution. *Subcell Biochem* **52**: 25-73.
998 Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA,
999 Mann I, Cook K et al. 2014. Determination and inference of eukaryotic transcription factor
1000 sequence specificity. *Cell* **158**(6): 1431-1443.
1001 Wolfe SA, Nekludova L, Pabo CO. 2000. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev*
1002 *Biophys Biomol Struct* **29**: 183-212.
1003 Zhao G, Ihuegbu N, Lee M, Schriefer L, Wang T, Stormo GD. 2012. Conserved Motifs and Prediction of
1004 Regulatory Modules in Caenorhabditis elegans. *G3 (Bethesda)* **2**(4): 469-481.
1005 Zhao G, Schriefer LA, Stormo GD. 2007. Identification of muscle-specific regulatory modules in
1006 Caenorhabditis elegans. *Genome research* **17**(3): 348-357.
1007 Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. *PLoS*
1008 *computational biology* **5**(12): e1000590.
1009 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors require only
1010 simple models of specificity. *Nat Biotechnol* **29**(6): 480-483.
1011 Zhu L, Wilken J, Phillips NB, Narendra U, Chan G, Stratton SM, Kent SB, Weiss MA. 2000. Sexual
1012 dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Genes*
1013 *Dev* **14**(14): 1750-1764.

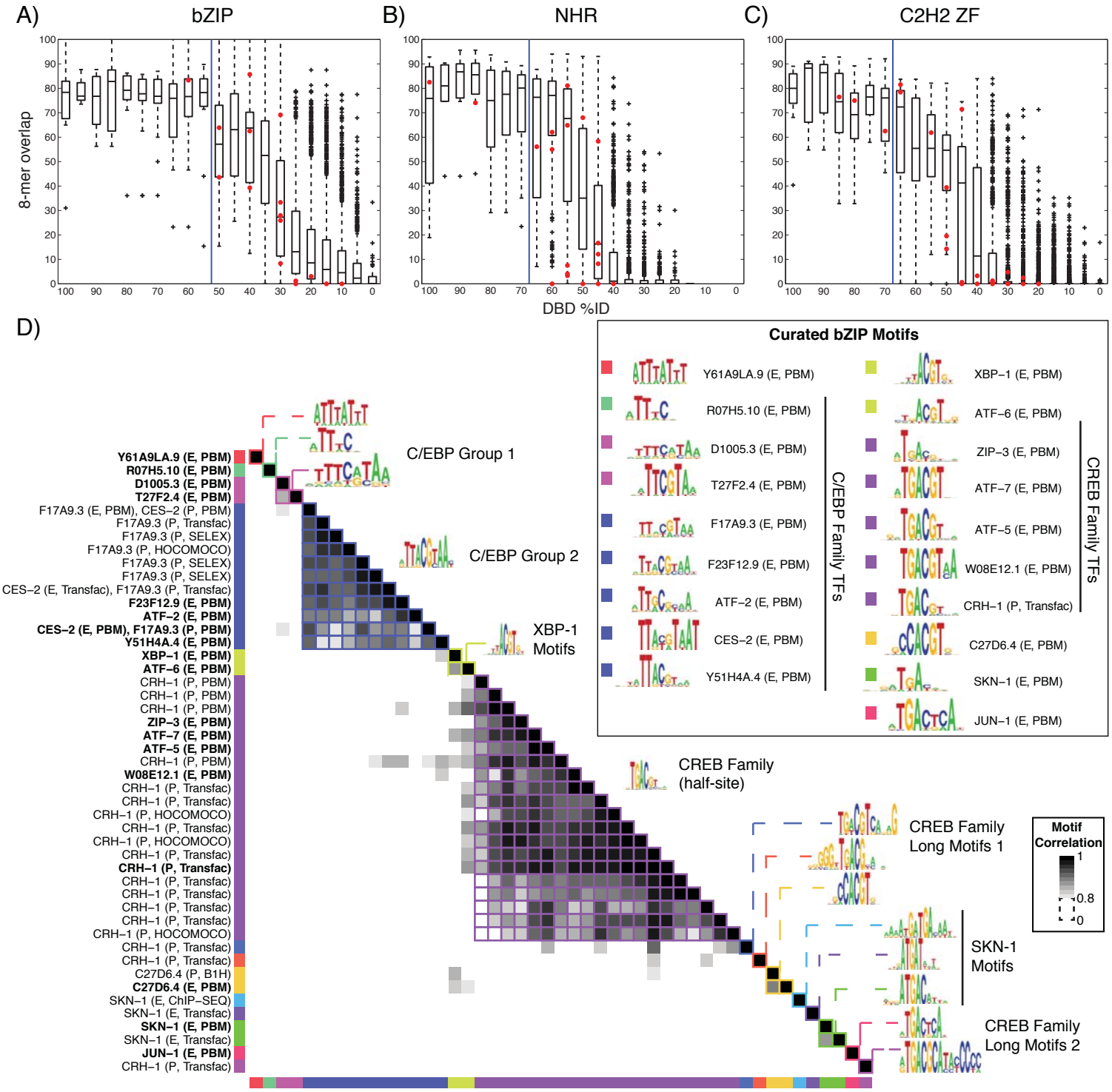
1014

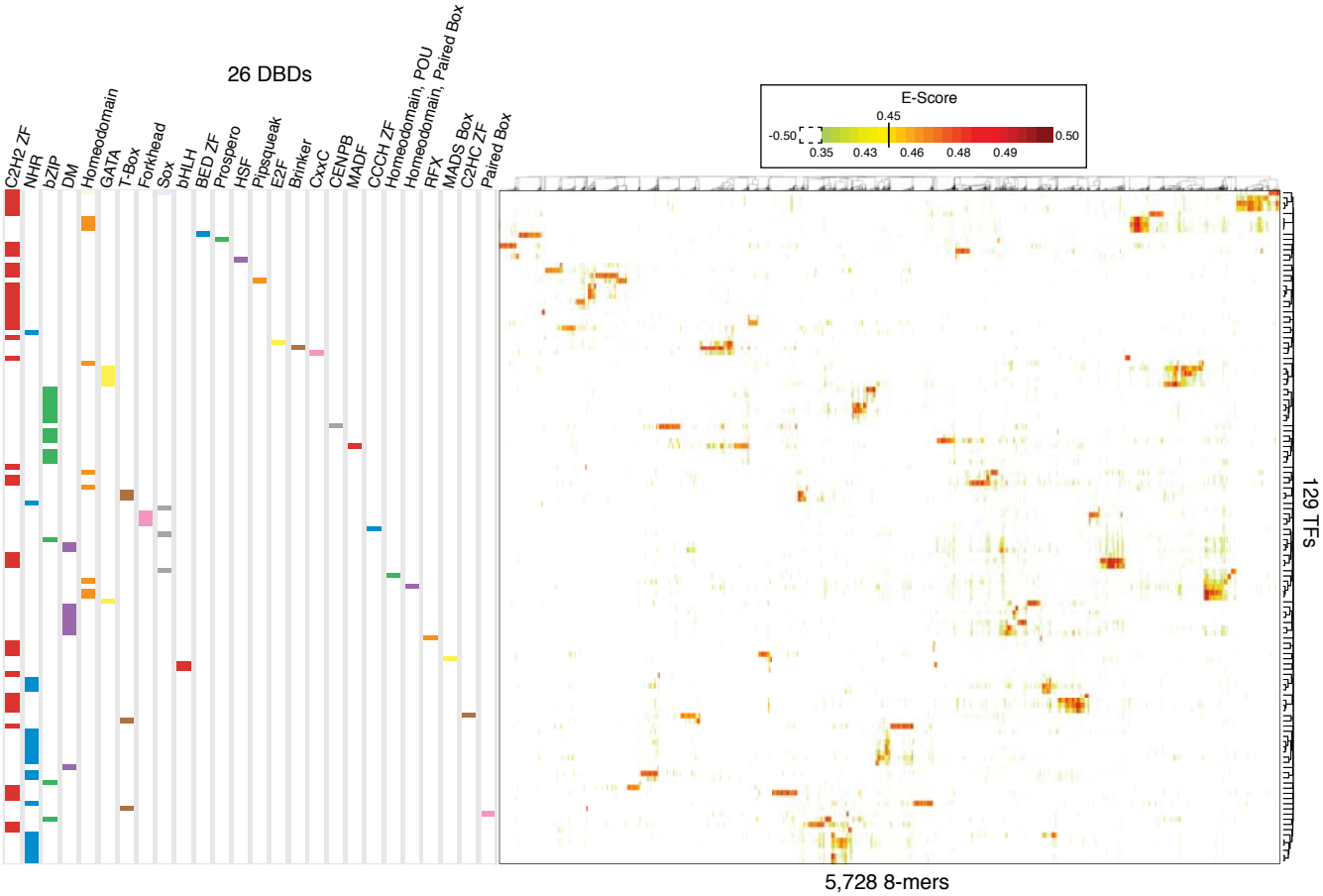
1015

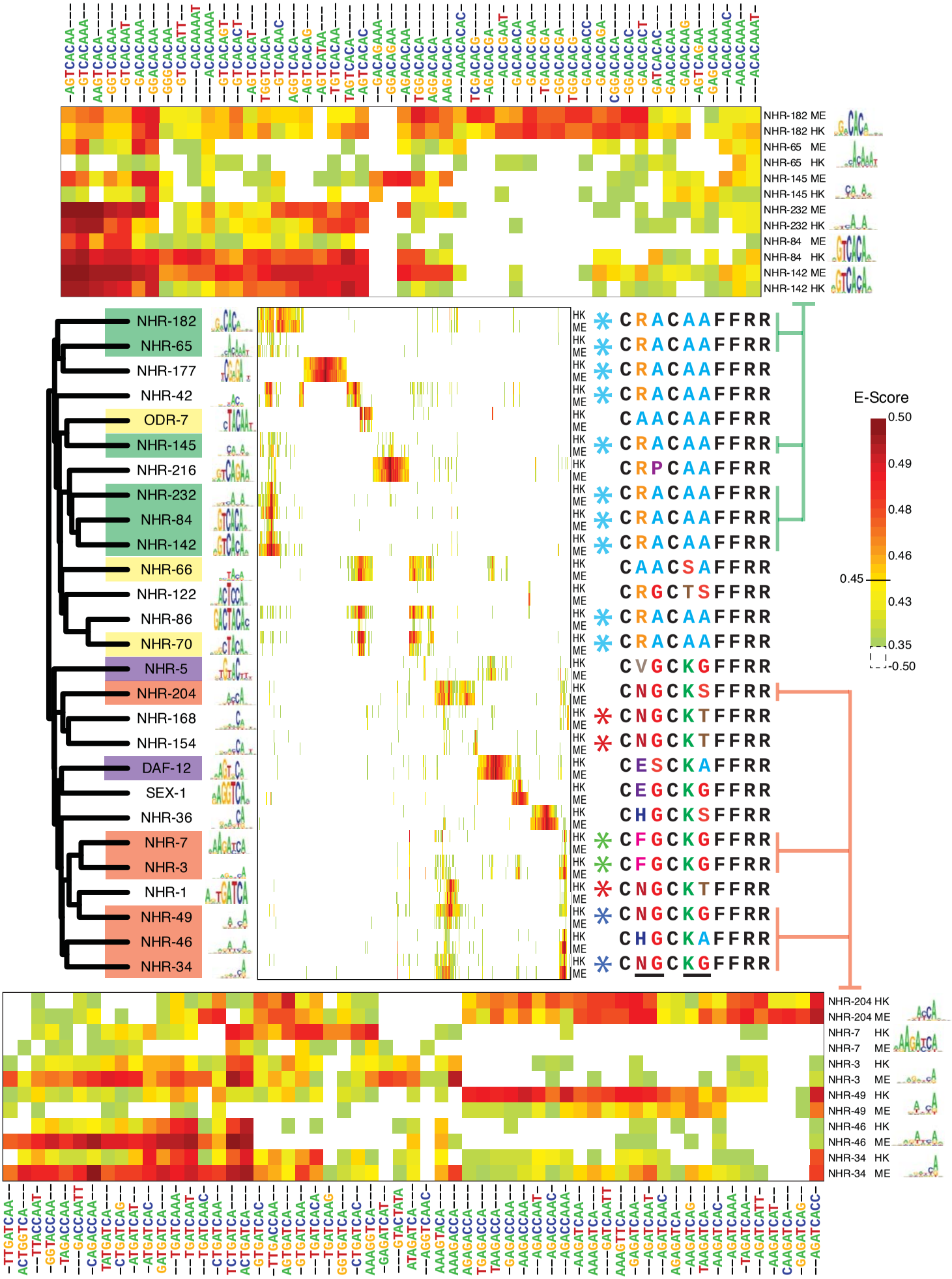


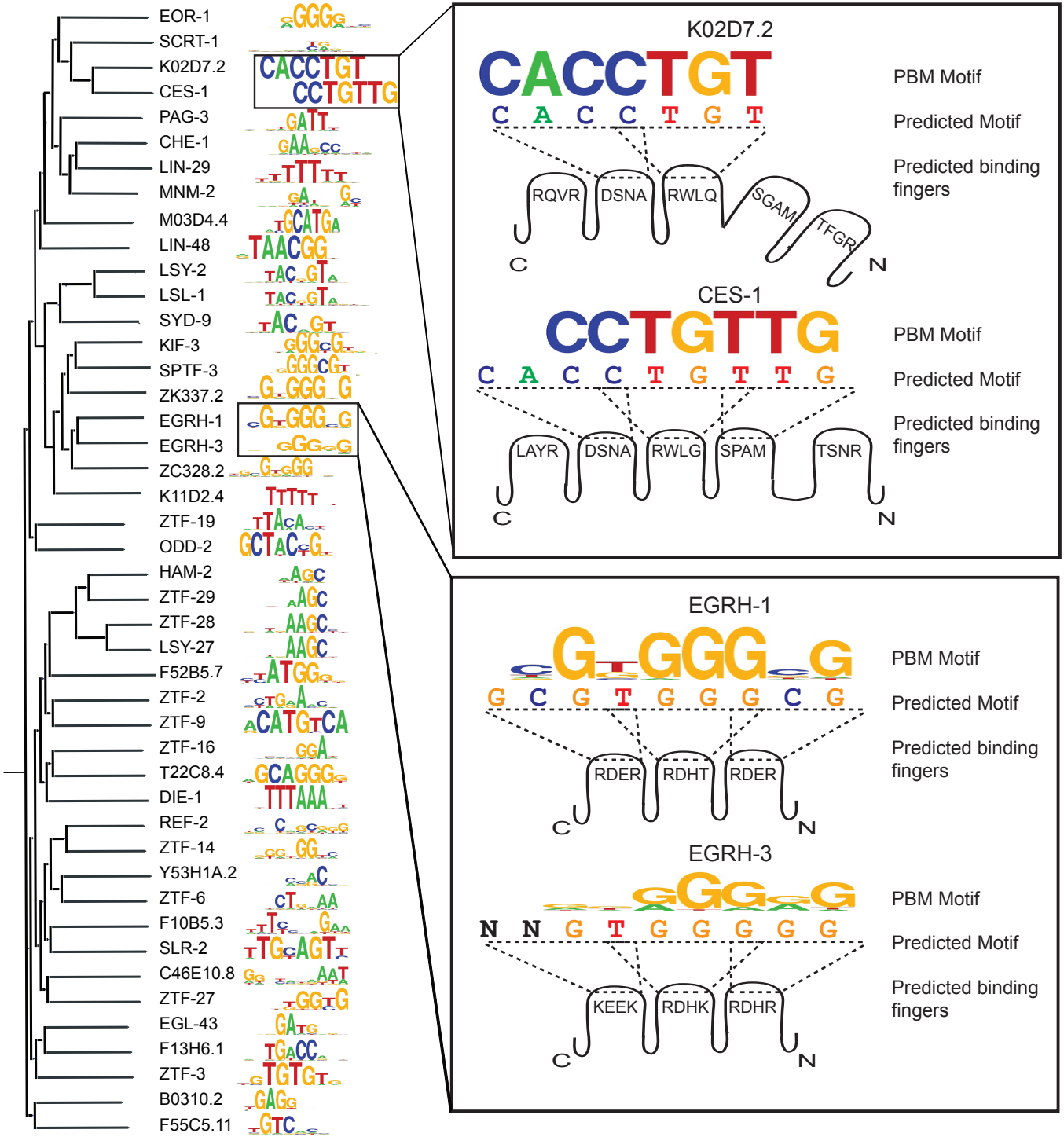
Number of TFs with

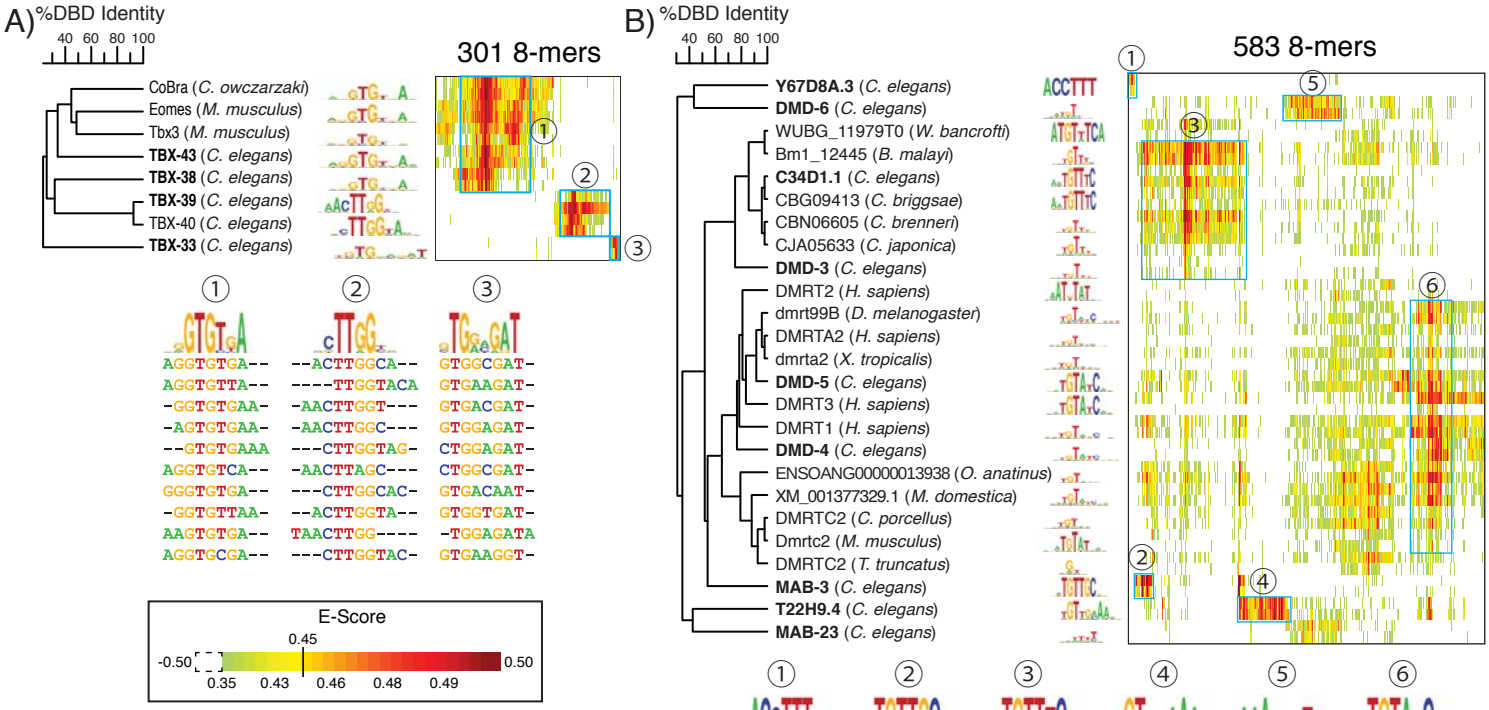
Canonical DBD	763
PBM Motifs	129
Previous Motifs	71
Predicted Motifs	97
Experimental and Predicted Motifs	292

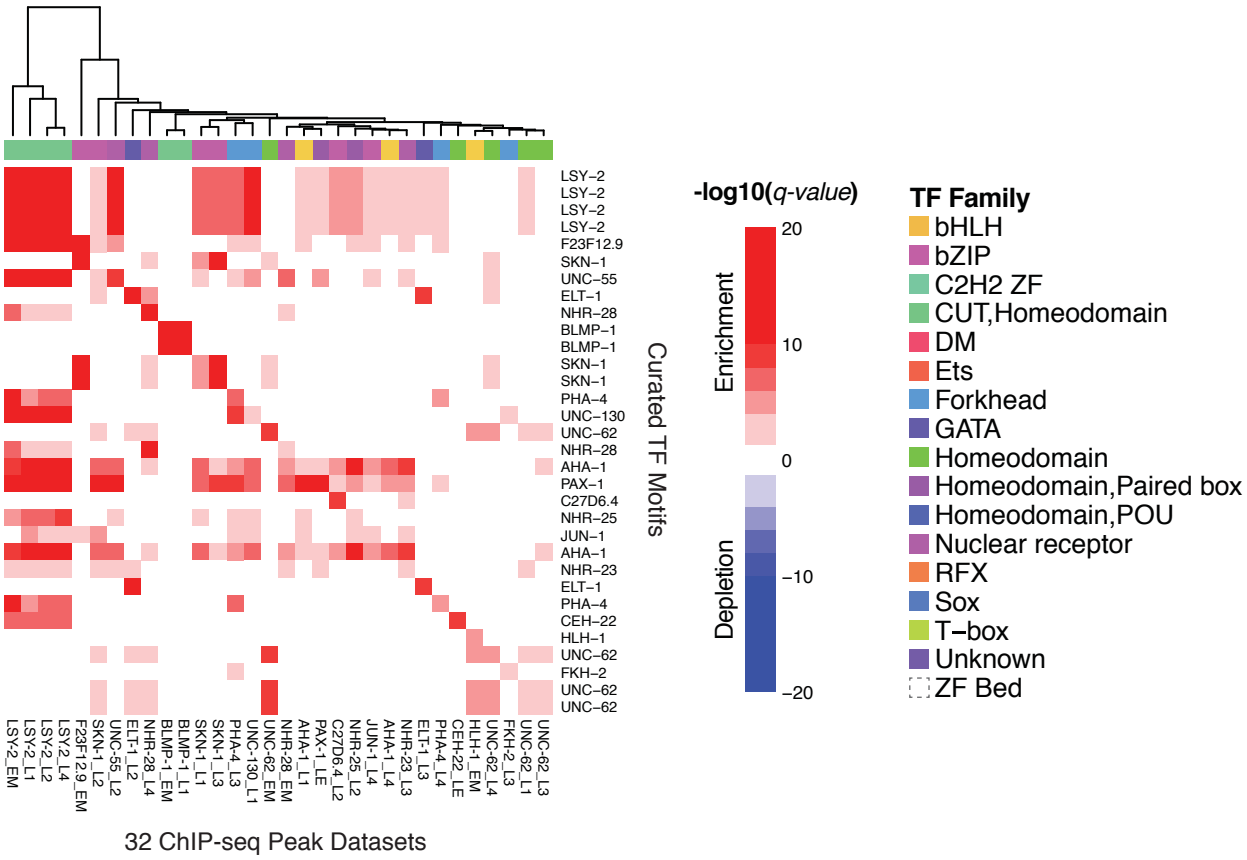

A)
bZIP
B)
NHR
C)
C2H2 ZF
8-mer overlap
DBD %ID
D)
Curated bZIP Motifs
Y61A9LA.9 (E, PBM)
R07H5.10 (E, PBM)
D1005.3 (E, PBM)
T27F2.4 (E, PBM)
F17A9.3 (E, PBM)
F23F12.9 (E, PBM)
ATF−2 (E, PBM)
CES−2 (E, PBM)
Y51H4A.4 (E, PBM)
C/EBP Family TFs
XBP−1 (E, PBM)
ATF−6 (E, PBM)
ZIP−3 (E, PBM)
ATF−7 (E, PBM)
ATF−5 (E, PBM)
W08E12.1 (E, PBM)
CRH−1 (P, Transfac)
C27D6.4 (E, PBM)
SKN−1 (E, PBM)
JUN−1 (E, PBM)
CREB Family TFs
C/EBP Group 1
C/EBP Group 2
XBP-1 Motifs
CREB Family (half-site)
CREB Family Long Motifs 1
SKN-1 Motifs
CREB Family Long Motifs 2
Motif Correlation
1
0.8
0
Y61A9LA.9 (E, PBM)
R07H5.10 (E, PBM)
D1005.3 (E, PBM)
T27F2.4 (E, PBM)
F17A9.3 (E, PBM), CES−2 (P, PBM)
F17A9.3 (P, Transfac)
F17A9.3 (P, SELEX)
F17A9.3 (P, HOCOMOCO)
F17A9.3 (P, SELEX)
F17A9.3 (P, SELEX)
CES−2 (E, Transfac), F17A9.3 (P, Transfac)
F23F12.9 (E, PBM)
ATF−2 (E, PBM)
CES−2 (E, PBM), F17A9.3 (P, PBM)
Y51H4A.4 (E, PBM)
XBP−1 (E, PBM)
ATF−6 (E, PBM)
CRH−1 (P, PBM)
CRH−1 (P, PBM)
CRH−1 (P, PBM)
ZIP−3 (E, PBM)
ATF−7 (E, PBM)
ATF−5 (E, PBM)
CRH−1 (P, PBM)
W08E12.1 (E, PBM)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
CRH−1 (P, HOCOMOCO)
CRH−1 (P, Transfac)
CRH−1 (P, HOCOMOCO)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
CRH−1 (P, HOCOMOCO)
CRH−1 (P, Transfac)
CRH−1 (P, Transfac)
C27D6.4 (P, B1H)
C27D6.4 (E, PBM)
SKN−1 (E, ChIP−SEQ)
SKN−1 (E, Transfac)
SKN−1 (E, PBM)
SKN−1 (E, Transfac)
JUN−1 (E, PBM)
CRH−1 (P, Transfac)

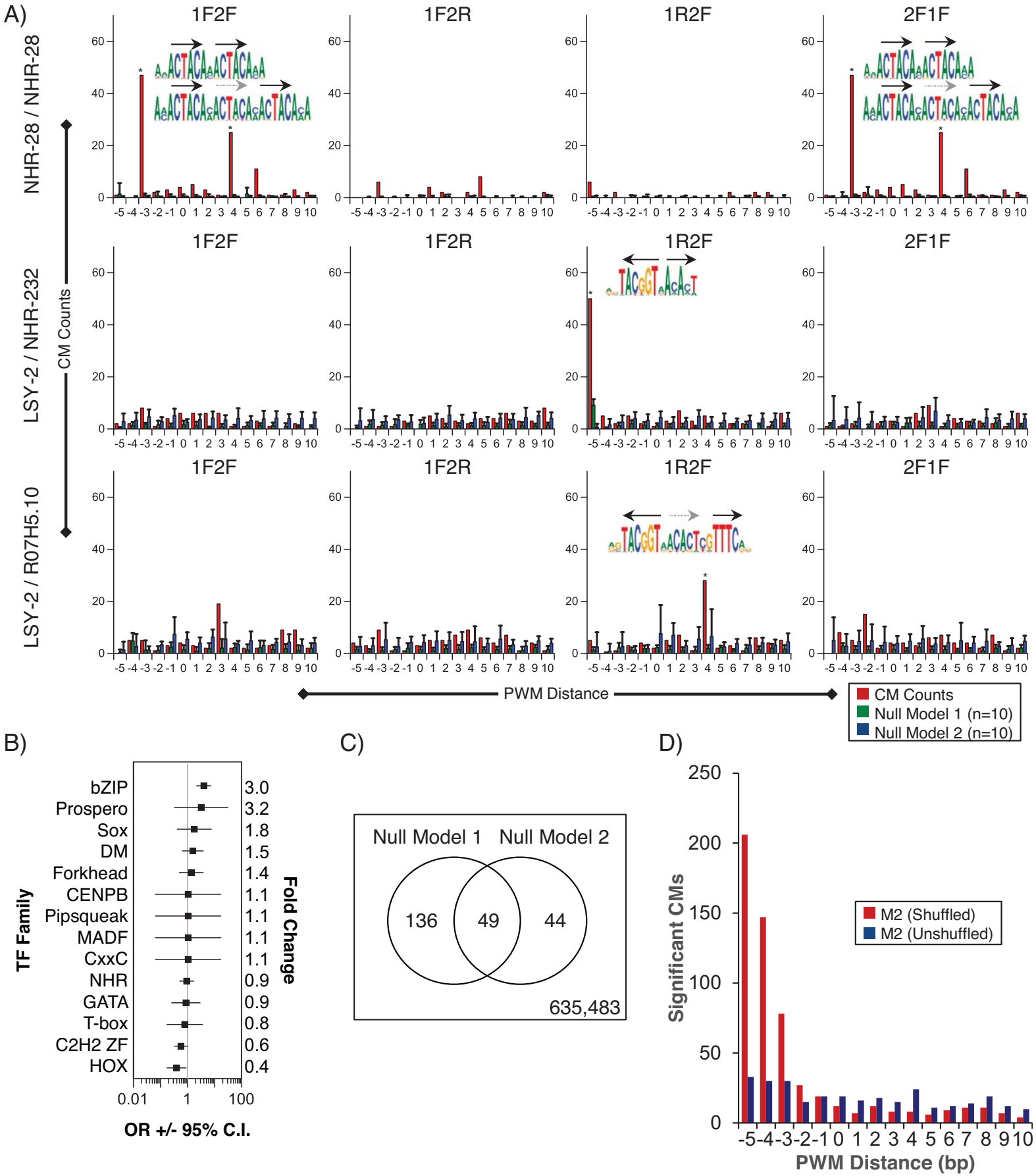

26 DBDs
C2H2 ZF
NHR
bZIP
DM
Homeodomain
GATA
T-Box
Forkhead
Sox
bHLH
BED ZF
Prospero
HSF
Pipsqueak
E2F
Brinker
CxxC
CENPB
MADF
CCCH ZF
Homeodomain, POU
Homeodomain, Paired Box
RFX
MADS Box
C2HC ZF
Paired Box
E-Score
-0.50
0.35
0.43
0.45
0.46
0.48
0.49
0.50
129 TFs
5,728 8-mers

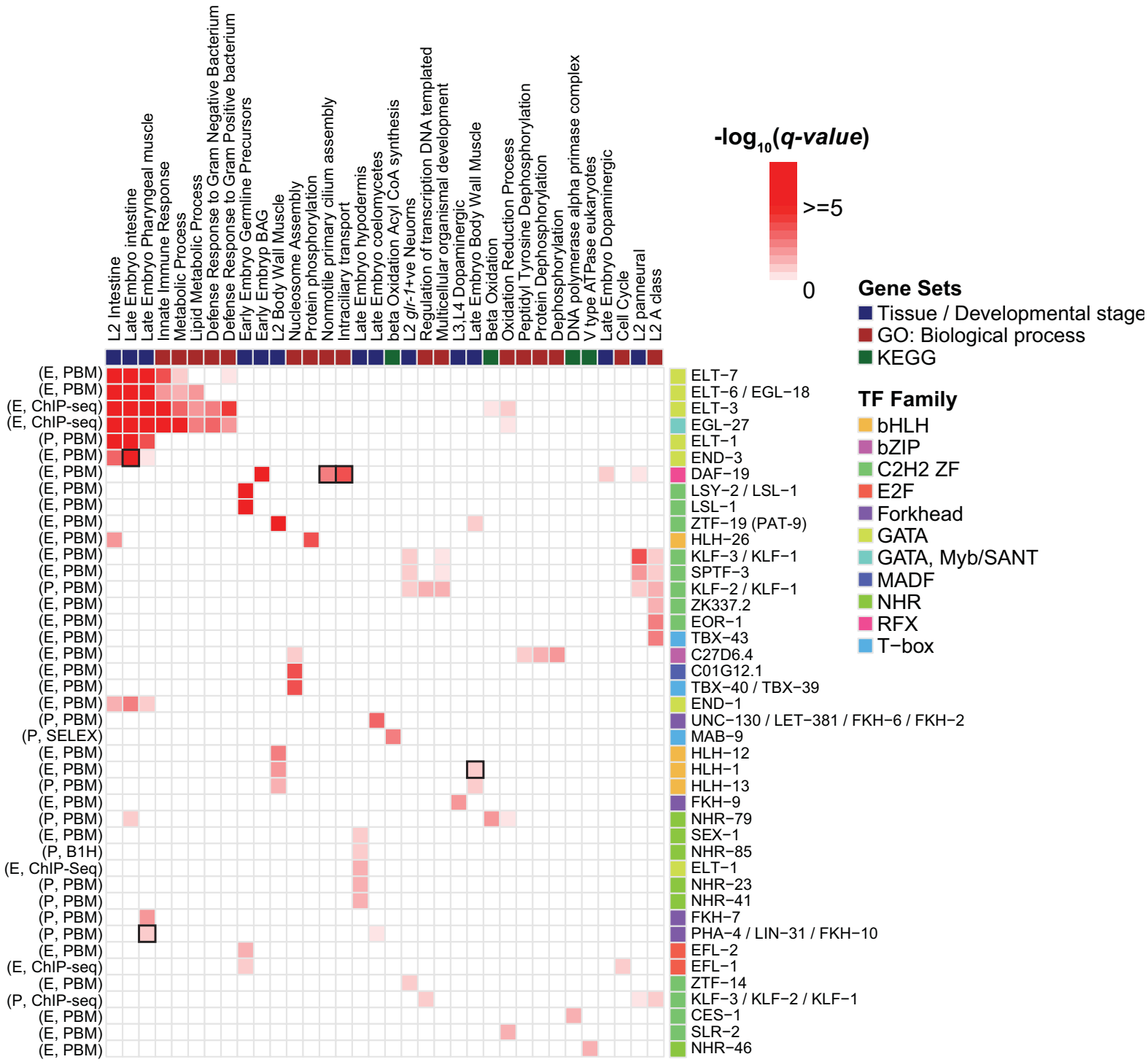

NHR-182 ME
NHR-182 HK
NHR-65 ME
NHR-65 HK
NHR-145 ME
NHR-145 HK
NHR-232 ME
NHR-232 HK
NHR-84 ME
NHR-84 HK
NHR-142 ME
NHR-142 HK
NHR-182
NHR-65
NHR-177
NHR-42
ODR-7
NHR-145
NHR-216
NHR-232
NHR-84
NHR-142
NHR-66
NHR-122
NHR-86
NHR-70
NHR-5
NHR-204
NHR-168
NHR-154
DAF-12
SEX-1
NHR-36
NHR-7
NHR-3
NHR-1
NHR-49
NHR-46
NHR-34
E-Score
0.50
0.49
0.48
0.46
0.45
0.43
0.35
-0.50
NHR-204 HK
NHR-204 ME
NHR-7 HK
NHR-7 ME
NHR-3 HK
NHR-3 ME
NHR-49 HK
NHR-49 ME
NHR-46 HK
NHR-46 ME
NHR-34 HK
NHR-34 ME


EOR-1
SCRT-1
K02D7.2
CES-1
PAG-3
CHE-1
LIN-29
MNM-2
M03D4.4
LIN-48
LSY-2
LSL-1
SYD-9
KIF-3
SPTF-3
ZK337.2
EGRH-1
EGRH-3
ZC328.2
K11D2.4
ZTF-19
ODD-2
HAM-2
ZTF-29
ZTF-28
LSY-27
F52B5.7
ZTF-2
ZTF-9
ZTF-16
T22C8.4
DIE-1
REF-2
ZTF-14
Y53H1A.2
ZTF-6
F10B5.3
SLR-2
C46E10.8
ZTF-27
EGL-43
F13H6.1
ZTF-3
B0310.2
F55C5.11
K02D7.2
PBM Motif
Predicted Motif
Predicted binding fingers
RQVR
DSNA
RWLQ
SGAM
TFGR
C
N
CES-1
PBM Motif
Predicted Motif
Predicted binding fingers
LAYR
DSNA
RWLG
SPAM
TSNR
C
N
EGRH-1
PBM Motif
Predicted Motif
Predicted binding fingers
RDER
RDHT
RDER
C
N
EGRH-3
PBM Motif
Predicted Motif
Predicted binding fingers
KEEK
RDHK
RDHR
C
N


A)
%DBD Identity
40 60 80 100
CoBra (C. owczarzaki)
Eomes (M. musculus)
Tbx3 (M. musculus)
TBX-43 (C. elegans)
TBX-38 (C. elegans)
TBX-39 (C. elegans)
TBX-40 (C. elegans)
TBX-33 (C. elegans)
301 8-mers
1
2
3
AGGTGTGA-- --ACTTGGCA-- GTGGCGAT-
AGGTGTTA-- ----TTGGTACA GTGAAGAT-
-GGTGTGAA- -AACTTGGT--- GTGACGAT-
-AGTGTGAA- -AACTTGGC--- GTGGAGAT-
--GTGTGAAA ---CTTGGTAG- CTGGAGAT-
AGGTGTCA-- -AACTTAGC--- CTGGCGAT-
GGGTGTGA-- ---CTTGGCAC- GTGACAAT-
-GGTGTTAA- --ACTTGGTA-- GTGGTGAT-
AAGTGTGA-- TAACTTGG---- -TGGAGATA
AGGTGCGA-- ---CTTGGTAC- GTGAAGGT-
E-Score
0.45
-0.50
0.50
0.35 0.43 0.46 0.48 0.49
B)
%DBD Identity
40 60 80 100
Y67D8A.3 (C. elegans)
DMD-6 (C. elegans)
WUBG_11979T0 (W. bancrofti)
Bm1_12445 (B. malayi)
C34D1.1 (C. elegans)
CBG09413 (C. briggsae)
CBN06605 (C. brenneri)
CJA05633 (C. japonica)
DMD-3 (C. elegans)
DMRT2 (H. sapiens)
dmrt99B (D. melanogaster)
DMRTA2 (H. sapiens)
dmrta2 (X. tropicalis)
DMD-5 (C. elegans)
DMRT3 (H. sapiens)
DMRT1 (H. sapiens)
DMD-4 (C. elegans)
ENSOANG00000013938 (O. anatinus)
XM_001377329.1 (M. domestica)
DMRTC2 (C. porcellus)
Dmrtc2 (M. musculus)
DMRTC2 (T. truncatus)
MAB-3 (C. elegans)
T22H9.4 (C. elegans)
MAB-23 (C. elegans)
583 8-mers
1
2
3
4
5
6
TGACCTTT-- -ATGTTGCG- -ATGTTTCA- TGTTGAAA- AAAAGTTA -ATGTATCA-
-GACCTTTC- --TGTTGCGA AATGTTTC-- -GTTGAAAT AAAATATT --TGTATCAA
-GACCTTTT- AATGTTGC-- -ATGTTTCT- TGTTAAAA- AAAAAGTT AATGTATC--
CAACCTTT-- GATGTTGC-- TATGTTTC-- -GTTGAAAA AAATATTT --TGTAACAA
CGACCTTT-- TATGTTGC-- -ATGTTGCA- TGTCGAAA- AAATATTA -TTGTAACA-
--ACCTTTTA -TTGTTGCG- --TGTTTCAA -GTTAAAAA AAAATTTT --TGTAACAT
TAACCTTT-- ATTGTTGC-- --TGTTTCAC -GTTAAAAT AAAAATTT -ATGTAACA-
-AACCTTTC- --TGTTGCGT --TGTTTCAT TGTTGAAT- AAATATAT -ATGTATCT-
--ACCTTTCC --TGTTGCGG ATTGTTTC-- TGTAGAAA- ATAAATTT ATTGTATC--
CGACTTTT-- --TGTTGCCA -TTGTTTCA- CGTTGCAA- TAAAATTT AATGTAAC--


LSY-2
LSY-2
LSY-2
LSY-2
F23F12.9
SKN-1
UNC-55
ELT-1
NHR-28
BLMP-1
BLMP-1
SKN-1
SKN-1
PHA-4
UNC-130
UNC-62
NHR-28
AHA-1
PAX-1
C27D6.4
NHR-25
JUN-1
AHA-1
NHR-23
ELT-1
PHA-4
CEH-22
HLH-1
UNC-62
FKH-2
UNC-62
UNC-62
LSY-2_EM
LSY-2_L1
LSY-2_L2
LSY-2_L4
F23F12.9_EM
SKN-1_L2
UNC-55_L2
ELT-1_L2
NHR-28_L4
BLMP-1_EM
BLMP-1_L1
SKN-1_L1
SKN-1_L3
PHA-4_L3
UNC-130_L1
UNC-62_EM
NHR-28_EM
AHA-1_L1
PAX-1_LE
C27D6.4_L2
NHR-25_L2
JUN-1_L4
AHA-1_L4
NHR-23_L3
ELT-1_L3
PHA-4_L4
CEH-22_LE
HLH-1_EM
UNC-62_L4
FKH-2_L3
UNC-62_L1
UNC-62_L3
32 ChIP-seq Peak Datasets
Curated TF Motifs
-log10(q-value)
Enrichment
Depletion
20
10
0
−10
−20
TF Family
bHLH
bZIP
C2H2 ZF
CUT,Homeodomain
DM
Ets
Forkhead
GATA
Homeodomain
Homeodomain,Paired box
Homeodomain,POU
Nuclear receptor
RFX
Sox
T−box
Unknown
ZF Bed


A)
1F2F
1F2R
1R2F
2F1F
NHR-28 / NHR-28
LSY-2 / NHR-232
LSY-2 / R07H5.10
CM Counts
PWM Distance
CM Counts
Null Model 1 (n=10)
Null Model 2 (n=10)
B)
bZIP 3.0
Prospero 3.2
Sox 1.8
DM 1.5
Forkhead 1.4
CENPB 1.1
Pipsqueak 1.1
MADF 1.1
CxxC 1.1
NHR 0.9
GATA 0.9
T-box 0.8
C2H2 ZF 0.6
HOX 0.4
TF Family
Fold Change
OR +/- 95% C.I.
C)
Null Model 1
Null Model 2
136
49
44
635,483
D)
Significant CMs
M2 (Shuffled)
M2 (Unshuffled)
PWM Distance (bp)


L2 Intestine
Late Embryo intestine
Late Embryo Pharyngeal muscle
Innate Immune Response
Metabolic Process
Lipid Metabolic Process
Defense Response to Gram Negative Bacterium
Defense Response to Gram Positive bacterium
Early Embryo Germline Precursors
Early Embryo BAG
L2 Body Wall Muscle
Nucleosome Assembly
Protein phosphorylation
Nonmotile primary cilium assembly
Intraciliary transport
Late Embryo hypodermis
Late Embryo coelomycetes
beta Oxidation Acyl CoA synthesis
L2 glr-1+ve Neuorns
Regulation of transcription DNA templated
Multicellular organismal development
L3,L4 Dopaminergic
Late Embryo Body Wall Muscle
Beta Oxidation
Oxidation Reduction Process
Peptidyl Tyrosine Dephosphorylation
Protein Dephosphorylation
Dephosphorylation
DNA polymerase alpha primase complex
V type ATPase eukaryotes
Late Embryo Dopaminergic
Cell Cycle
L2 panneural
L2 A class
-log10(q-value)
>=5
0
Gene Sets
Tissue / Developmental stage
GO: Biological process
KEGG
TF Family
bHLH
bZIP
C2H2 ZF
E2F
Forkhead
GATA
GATA, Myb/SANT
MADF
NHR
RFX
T−box
(E, PBM) ELT−7
(E, PBM) ELT−6 / EGL−18
(E, ChIP-seq) ELT−3
(E, ChIP-seq) EGL−27
(P, PBM) ELT−1
(E, PBM) END−3
(E, PBM) DAF−19
(E, PBM) LSY−2 / LSL−1
(E, PBM) LSL−1
(E, PBM) ZTF−19 (PAT-9)
(E, PBM) HLH−26
(E, PBM) KLF−3 / KLF−1
(E, PBM) SPTF−3
(P, PBM) KLF−2 / KLF−1
(E, PBM) ZK337.2
(E, PBM) EOR−1
(E, PBM) TBX−43
(E, PBM) C27D6.4
(E, PBM) C01G12.1
(E, PBM) TBX−40 / TBX−39
(E, PBM) END−1
(P, PBM) UNC−130 / LET−381 / FKH−6 / FKH−2
(P, SELEX) MAB−9
(E, PBM) HLH−12
(E, PBM) HLH−1
(P, PBM) HLH−13
(E, PBM) FKH−9
(P, PBM) NHR−79
(E, PBM) SEX−1
(P, B1H) NHR−85
(E, ChIP-Seq) ELT−1
(P, PBM) NHR−23
(P, PBM) NHR−41
(P, PBM) FKH−7
(P, PBM) PHA−4 / LIN−31 / FKH−10
(E, PBM) EFL−2
(E, ChIP-seq) EFL−1
(E, PBM) ZTF−14
(P, ChIP-seq) KLF−3 / KLF−2 / KLF−1
(E, PBM) CES−1
(E, PBM) SLR−2
(E, PBM) NHR−46