Regulatory genome annotation of 33 insect species

Hasiba Asma[1], Ellen Tieke[2], Kevin D. Deem[2§], Jabale Rahmat[2], Tiffany Dong[1], Xinbo Huang[1], Yoshinori Tomoyasu[2] and Marc S. Halfon[1,3,4,5,*]

[1]Program in Genetics, Genomics, and Bioinformatics, University at Buffalo-State University of New York, Buffalo NY 14203

[2]Department of Biology, Miami University, Oxford OH 45056

[3]Department of Biochemistry, University at Buffalo-State University of New York, Buffalo NY 14203

[4]Department of Biomedical Informatics, University at Buffalo-State University of New York, Buffalo NY 14203

[5]Department of Biological Sciences, University at Buffalo-State University of New York, Buffalo NY 14260

*Author for Correspondence at:

Jacobs School of Medicine and Biomedical Sciences
University at Buffalo
955 Main St.
Room 5128
Buffalo, NY 14203

716-829-3126

mshalfon@buffalo.edu

[§]Present address:
Department of Biology
University of Rochester
Rochester, NY 14627

Keywords: Regulatory genomics, arthropods, enhancers, cis-regulation, genome annotation, functional genomics, enhancer prediction

ABSTRACT

Annotation of newly-sequenced genomes frequently includes genes, but rarely covers important non-coding genomic features such as the *cis*-regulatory modules—e.g., enhancers and silencers—that regulate gene expression. Here, we begin to remedy this situation by developing a workflow for rapid initial annotation of insect regulatory sequences, and provide a searchable database resource with enhancer predictions for 33 genomes. Using our previously-developed SCRMshaw computational enhancer prediction method, we predict over 2.8 million regulatory sequences along with the tissues where they are expected to be active, in a set of insect species ranging over 360 million years of evolution. Extensive analysis and validation of the data provides several lines of evidence suggesting that we achieve a high true-positive rate for enhancer prediction. One, we show that our predictions target specific loci, rather than random genomic locations. Two, we predict enhancers in orthologous loci across a diverged set of species to a significantly higher degree than random expectation would allow. Three, we demonstrate that our predictions are highly enriched for regions of accessible chromatin. Four, we achieve a validation rate in excess of 70% using in vivo reporter gene assays. As we continue to annotate both new tissues and new species, our regulatory annotation resource will provide a rich source of data for the research community and will have utility for both small-scale (single gene, single species) and large-scale (many genes, many species) studies of gene regulation. In particular, the ability to search for functionally-related regulatory elements in orthologous loci should greatly facilitate studies of enhancer evolution even among distantly related species.

68 **INTRODUCTION**

69 The past two decades have witnessed an explosive rise in sequenced metazoan genomes, from a
70 mere handful in the first few years of the century to over 8,000 today (ref. 1; accessed 16 January
71 2024). This impressive statistic, however, masks the reality that these genomes exist in various
72 stages of completion. Fewer than 30% of these genomes are assembled at the chromosome level,
73 and only 28% of those have a comprehensive annotation (1). Moreover, almost none of the
74 genome annotations include regulatory sequences (also referred to as *cis*-regulatory modules, or
75 CRMs) such as enhancers and silencers. This is unfortunate, as CRMs comprise a significant
76 percentage of the genome, and knowledge of these sequences is expected to have value
77 comparable to that of knowing the protein-coding genes. Characterizing CRMs is critical for
78 understanding mechanisms of gene regulation and the organization of gene regulatory networks.
79 Moreover, the role of regulatory mutations is increasingly recognized as a driver of both
80 evolution and disease (2-6).

81 One reason for the overall dearth of regulatory annotations is that, historically, large-scale CRM
82 discovery has been difficult and both resource and labor intensive (7). For much of the last four
83 decades, CRMs could only be identified through painstaking, low-throughput experimental
84 assays. Although in recent years high-throughput empirical and computational CRM discovery
85 methods have been developed, the various different methods frequently show limited agreement
86 (8-10), with the result that comprehensive CRM annotation across all cell types and life-cycle
87 stages has remained a challenge for all but the most exhaustively studied model organisms. The
88 problem is particularly acute for the insects. Insects represent a species-rich class—they
89 constitute somewhere between 65%-90% of all animal species (11, 12)—and have major impacts

90 on human health and agriculture. The early radiation of the insects, coupled with typically short
91 generation times, means that most of the relevant biomedically and agriculturally important
92 species share little non-coding sequence conservation with each other or with the principal insect
93 model species, *Drosophila melanogaster*. Thus, common sequence-homology based CRM
94 discovery approaches are of little use in providing regulatory insights into these species, and
95 knowledge transfers poorly from one species to another. Furthermore, many insects have a
96 complex and varied life cycle, making it particularly important—yet onerous—to assay for CRM
97 function at multiple stages.

98 We previously developed a powerful computational method, SCRMshaw ("Supervised *Cis*-
99 Regulatory Module prediction"), for accurate prediction of CRMs, particularly enhancers (13-
100 15). (Although we expect SCRMshaw to be adept at finding multiple CRM types, our validation
101 efforts to date have focused solely on enhancers.) SCRMshaw requires only a sequenced genome
102 and a "training set" of some 15-30 known enhancers that regulate a common pattern of gene
103 expression (e.g., midgut expression) and relies on the idea that enhancers with similar function
104 will have similar sequence characteristics, not possible to detect by eye or by traditional
105 alignment methods, but identifiable using machine-learning. Although not universally true, this
106 assumption is robust enough to allow effective enhancer discovery without requiring knowledge
107 of transcription factor binding sites or of the expression patterns of the genes being regulated.
108 Importantly, we have shown that we can leverage the wealth of existing *D. melanogaster*
109 enhancer data (16) to train models for *cross-species* supervised enhancer discovery in diverged
110 (160-345 million years (MY)) insect species—including flies, mosquitoes, beetles, bees, and
111 wasps—despite a virtually complete lack of observable alignment at the non-coding sequence
112 level (17-20).

113 Here, we use SCRMshaw to undertake an initial regulatory annotation of 33 individual insect
114 genomes, using a collection of 48 training sets composed of experimentally validated *D.*
115 *melanogaster* enhancers. These species are spread across five orders spanning over 360 MY of
116 evolution, and represent roughly 10% of annotated insect species with scaffold-level or better
117 assembly. Annotated predicted enhancers are provided in a searchable database that allows
118 querying by species, tissue/cell type, or potential target gene. A series of simulations as well as *in*
119 *silico* and *in vivo* validation experiments demonstrate the effectiveness of our approach and place
120 an upper bound on false positive prediction rates. Our results represent the first release of a rich
121 insect regulatory genome annotation resource, which will continue to grow and annotate insect
122 regulatory genomes in parallel with the sequencing of new insect genomes.

123

124 **RESULTS**

125 We previously demonstrated that SCRMshaw is remarkably effective at predicting enhancers
126 across the entire ~345 MY range of the holometabolous insects, using training data derived
127 solely from *D. melanogaster* (17, 19). SCRMshaw (Fig. 1A) uses training sets composed of
128 known enhancers defined by a common functional characterization (e.g. "nervous system,"
129 "wing disc") to build a statistical model that captures their short DNA subsequence (kmer) count
130 distribution. This kmer distribution is then compared to that of a set of non-enhancer
131 "background" sequences in a machine-learning framework. The kmers likely serve as proxies for
132 the unknown transcription factor binding sites, but these sites themselves, even when known, are
133 not explicitly used by the algorithm. The trained model is then used to score overlapping

134 sequence windows in the genome, and the highest-scoring windows are output as predicted
135 enhancers (13, 14, 21). When searching the genomes of the mosquitoes *Anopheles gambiae* and
136 *Aedes aegypti,* the red flour beetle *Tribolium castaneum,* the honey bee *Apis mellifera,* and the
137 wasp *Nasonia vitripennis,* SCRMshaw successfully predicted enhancers in a cross-species
138 fashion with an approximately 75% prediction success rate, based on reporter gene assays in
139 xenotransgenic *D. melanogaster* (chosen as a pragmatic transgene host species) and comparison
140 to already-identified enhancers in the other species (17-20). These results suggest that there are
141 significant, albeit hidden, homologies governing the sequence characteristics of insect enhancers,
142 at least for those involved in a substantial number of gene regulatory networks, and motivated us
143 to apply SCRMshaw to a large and diverse set of sequenced insect genomes.

144

145 ***A cross-species SCRMshaw pipeline***

146 To facilitate application of SCRMshaw to large numbers of newly-sequenced genomes, we
147 developed a detailed workflow to ensure proper formatting of input genomes, rapid prediction of
148 tissue-specific enhancer sequences, evaluation of results, and annotation of loci with information
149 drawn from the respective orthologous regions in the richly-investigated *D. melanogaster*
150 genome (Fig. 1B; see Methods). The workflow consists of four major steps:

151 *Preflight*: SCRMshaw requires two input files for each genome: a FASTA-formatted file of the
152 genome sequence itself, and a GFFv3-formatted file of the genome annotation. The genome file
153 is masked for tandem repeats using Tandem Repeat Finder (22). *Preflight* validates the formats

154 of these files and produces a comprehensive log file that highlights any issues along with basic
155 information such as the number of chromosomes/scaffolds and their sizes, data types present in
156 the annotation (e.g., 'gene', 'exon', 'ncRNA', etc.), and average intergenic distances. *Preflight*
157 also provides a sample output of the SCRMshaw-generated 'gene' and 'exon' files. This feature
158 allows users to identify any discrepancies or errors stemming from the input files and to reformat
159 these files as needed before running SCRMshaw. Any minor scaffolds that are not annotated as
160 containing genes are discarded.

161 *SCRMshaw:* SCRMshaw is run as previously described (15), using the "HD" variant (21).

162 *Post-processing:* The raw SCRMshaw output is post-processed to determine the final set of
163 predicted enhancers. The original post-processing procedure described in (21) had a tendency to
164 predict enhancers skewed toward long lengths (median ~ 1100 bp). We have revised that method
165 here (see Methods) to yield predictions of more compact size (median 750 bp), which is more in
166 keeping with empirically characterized enhancer lengths (23).

167 *Orthology mapping:* As a final step, we map all putative target genes in the SCRMshaw output to
168 their *D. melanogaster* orthologs (if an ortholog exists) using a custom pipeline incorporating the
169 *Orthologer* software from the Zdobnov lab (24) (see Methods). We use *D. melanogaster* as it has
170 by far the most comprehensive gene annotation of the insects and thus provides the most detailed
171 functional information for each gene. Mapping the genes from each species to a common
172 ortholog allows us to assess whether we have obtained predicted enhancers in orthologous loci
173 within the various species on which we have run SCRMshaw. The orthology mapping step
174 requires that a set of predicted proteins is present as part of the existing genome annotation.

175

176 ***Annotation of 33 insect genomes***

177 We ran our annotation pipeline on an initial set of 33 genomes (additional genome annotation is
178 ongoing). These initial genomes were chosen based on availability and to sample broadly among
179 the holometabolous insect orders and the Hemiptera (Fig 2A; Table 1). For each genome, we ran
180 SCRMshaw using a collection of 48 training sets (Supplemental Table S1) and all three
181 SCRMshaw scoring methods ("IMM," "hexMCD", "PAC-rc"; (13, 14)). For fifteen species
182 where a protein annotation was available, we assigned *D. melanogaster* orthologs to each
183 predicted locus, with an average of 54% (range 38-82%) of genes in a given species having a *D.*
184 *melanogaster* ortholog (Fig. 2B).

185 Collectively, we predicted a total of 2,873,192 enhancers in these 33 species, with each species
186 having on average approximately 87,000 predictions (Supplemental Table S2) (Fig. 2C). (As
187 some enhancers are predicted by multiple scoring methods, or from more than one training set,
188 the number of unique sequences is lower at 1,164,354; see gray bars in Fig. 2C.) The median
189 length of the predicted enhancers across all species was 750 bp (mean 695, range 490-32500 bp).
190 However, we noted the presence of a small number—2642, < 0.1% of the total—of unusually
191 large regions (> 2000 bp), the bulk of which were confined to just a few genomes (Supplemental
192 Fig. S1). We therefore discarded any predictions with length greater than 1.5 times the
193 interquartile range of the complete prediction set and re-evaluated the size distribution. This
194 resulted in a median size of 740 bp with a mean of 676 bp and a range of 490-1120 bp, indicating
195 that the overall impact of these outlier sequences is minimal (Fig. 2D). Inspection of the

196 excessively large elements revealed that they result from SCRMshaw predictions that lie
197 immediately adjacent to each other (without gaps) and have similar SCRMshaw scores, which
198 thus become merged into one broad predicted element. Preliminary analysis suggests that these
199 regions result from insufficient masking of tandem repeat regions and/or genome assembly
200 errors, although other causes, such as extremely enhancer-dense "superenhancer" regions
201 (reviewed by 25), cannot entirely be ruled out.

202

203 ***Many loci contain multiple enhancer predictions***

204 We have noted in the past that SCRMshaw often predicts multiple enhancers in a single locus
205 (e.g. 26). This is consistent with the concept of "shadow enhancers," sets of multiple redundant
206 or semi-redundant enhancers regulating the same gene (reviewed by 27). On the other hand, if
207 SCRMshaw is predicting enhancers with low specificity (i.e., largely at random), it is instead
208 possible that larger loci may just accumulate a high number of predictions simply due to their
209 greater length.

210 To distinguish between these possibilities, we conducted simulations on three representative
211 genomes (*D. melanogaster, C. pipiens,* and *A. aegypti*), each with a different average intergenic
212 region size chosen to represent small, medium and large genomes respectively (average sizes 610
213 bp, 5,274.5 bp, and 14,641 bp; see Methods). We randomized the location of SCRMshaw
214 predictions across the non-coding component of each genome and compared the randomized
215 results to our SCRMshaw output. We found that, for a subset of loci, the number of real

216 SCRMshaw predictions per locus was consistently higher than the number obtained at random
217 (Supplemental Table S3). For example, using the *mapping1.visceral_mesoderm* training set,
218 4.6% (15/328) of *D. melanogaster* loci containing one or more SCRMshaw predictions had a
219 significantly (P<0.001) larger number of predictions/locus than expected, 1.5% (10/658) of *C.*
220 *pipiens* loci had more predictions than expected, and 1% (3/299) of *A. aegypti* loci had excess
221 predicted enhancers (Supplemental Table S3). Similarly, for the *mapping2.ectoderm* training set,
222 7.6%, 1.5%, and 11% of loci in the three species, respectively, had a significantly (P<0.001)
223 greater than expected number of predictions per locus (Supplemental Table S3). Averaged over
224 all training sets, 3-4% of loci with predictions had significantly more than expected by chance
225 (mean values: 3.0% for *D. melanogaster*, 3.0% for *C. pipiens*, and 3.7% for *A. aegypti*;
226 maximum values: 10.2% for *D. melanogaster,* 9.3% for *C. pipiens*, and 11.6% for *A. aegypti*).
227 Only very few training sets did not have any loci with significantly more enhancers than
228 expected (1 set in *A. aegypti,* 1 in *C. pipiens* and 8 in *D. melanogaster*; Supplemental Table S3).

229 To further ensure that these results were not influenced by locus size, we binned the loci by
230 length and assessed the numbers of real versus simulated predictions/locus for each bin. The
231 binned results were similar to the results using all loci, i.e., the number of SCRMshaw
232 predictions/locus (Fig. 3, blue boxplots) was consistently higher than the number obtained via
233 randomization (Fig. 3, pink boxplots). Results were also similar when comparing predictions
234 only at intergenic versus only at intronic positions (Supplemental Table S3). These results
235 suggest that SCRMshaw is not predicting enhancers randomly throughout the genome, but rather
236 is identifying multiple related "shadow" enhancers in a subset of loci. Consistent with this
237 interpretation, functional tests of the full set of predicted enhancers in several *D. melanogaster*

238 loci has confirmed that many of the predictions act as functionally similar enhancers (T.
239 Williams, personal communication).

240

241 ***SCRMshaw predicts enhancers in orthologous loci across species***

242 A major premise underlying cross-species applications of SCRMshaw is that conserved
243 regulatory strategies should allow for a model trained on enhancers in one species to predict for
244 similar enhancers in another species. We reasoned that at least some of the time, this should lead
245 to identification of enhancers in orthologous loci, as orthologous genes are frequently involved in
246 similar biological pathways and developmental regulatory networks (3). Indeed, we previously
247 showed that SCRMshaw was able to predict enhancers in several orthologous loci, for instance
248 those for the *single-minded* locus in *D. melanogaster, A. gambiae,* and *A. mellifera,* and the
249 *wingless* locus in *D. melanogaster, A. mellifera,* and *T. castaneum* (17). To test whether this is
250 generally true, we used the fifteen species with mapped *D. melanogaster* orthologs (plus *D.*
251 *melanogaster* itself), and all 48 of our training sets, to compare the number of SCRMshaw-based
252 versus random predictions obtained in common (orthologous) loci.

253 We observed a substantial reduction in the number of orthologous loci with predictions in
254 common as we moved from considering a set of ten to the full set of all sixteen species (average
255 number of common loci over all 48 training sets: 70.7 (10 species) > 39.4. > 19.8 > 9.14 > 4.12 >
256 1.6 > 0.6 (16 species) out of an average number of 640 predictions with mapped orthologs) (Fig.
257 4A, Supplemental Table S4a). This rapid decline is likely due to a variety of factors, including

258 differences in taxonomic order (e.g., Diptera vs. Hymenoptera), the quality and degree of
259 ortholog data for each species, and the quality of annotation for each species, all of which likely
260 lead to an underestimation of the true numbers of common loci in our data (see Discussion). We
261 simulated SCRMshaw predictions for all sixteen species by randomizing the SCRMshaw results
262 and compared the number of common loci between the simulated and real data for combinations
263 of ten to sixteen species. The number of common loci among the real SCRMshaw predictions
264 was consistently significantly higher than the number of common loci observed in the simulated
265 data ($P < 0.05$; Fig. 4B and Supplemental Table S4). In particular, when evaluating between 10
266 and 12 species (data for 14-16 species are not reliable due to the very small number of observed
267 loci in common) only a single training set, *adult_PNS*, did not have a significant
268 overrepresentation of common loci (Supplemental Table S4c). We also examined the fold
269 enrichment, i.e., the extent to which the true number was in excess of the simulated number, and
270 found that on average there were greater than 2.4x more predictions in common loci than
271 expected, when considering groups of 10-12 species; almost all training sets had a fold
272 enrichment >1.5x (Fig. 4C, Supplemental Table S4d). These results strongly suggest that
273 SCRMshaw is successfully finding sets of conserved (or convergent) enhancers, as it is
274 predicting sequences in orthologous loci in a training-set specific manner across multiple species
275 significantly more often than can be accounted for by chance.

276

277 ***SCRMshaw predictions correlate with regions of accessible chromatin***

278 Active enhancers are frequently found in regions of accessible chromatin, as assayed by methods
279 such as DNAse-seq, FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements with
280 sequencing), and ATAC-seq (Assay for Transposase-Accessible Chromatin by sequencing)(28-
281 30). We confirmed previously that FAIRE-predicted and SCRMshaw-predicted enhancers in *T.*
282 *castaneum* have a high (>79%) degree of overlap (18). To determine whether a similar
283 correlation exists for other species, we compared our SCRMshaw predictions with available
284 FAIRE-seq and ATAC-seq data for seven species: *D. melanogaster, T. castaneum, A. gambiae,*
285 *Danaus plexippus, Vanessa cardui, Junonia coenia,* and *Heliconius himera*. These comparisons
286 are imperfect, as the tissues used to obtain the chromatin data do not precisely correspond to the
287 training sequences used for SCRMshaw. Nevertheless, in the majority of cases where we were
288 able to establish a rough match between the tissues, we observed significant overlap between the
289 two methods of enhancer detection. For example, in *D. melanogaster*, 40% of SCRMshaw
290 predictions using the *blastoderm.mapping1* training set overlapped ATAC-seq peaks obtained
291 from blastoderm embryos (*P*<4.15e-112, fold enrichment 2.98) (31)(Table 2 row 2). Similarly,
292 35% and 41% of SCRMshaw predictions from the *mappng2.wing* and *haltere_disc* sets
293 overlapped FAIRE data that included wing and haltere cells (32)(*P*<4.6e-117 and 1.5e-102, fold
294 enrichment of 3.29 and 2.98)(Table 2, rows 1, 3). In the mosquito *A. gambiae*, we compared
295 SCRMshaw predictions from the *embryonic_midgut* and *mapping1.salivary* training sets to
296 ATAC-seq data from adult midgut and salivary tissues, observing overlaps of 40% and 37%
297 respectively (*P*<1.5e-102 and 2.15e-93, fold enrichment of 3.45 and 3.22;)(Table 2, rows 9, 10).
298 When comparing our SCRMshaw predictions from the *mapping2.wing* set to ATAC-seq data for
299 larval wing tissue in the butterflies *D. plexippus* and *H. himera*, we observed overlaps of 60%
300 and 68% (*P*<5.59e-19 and *P*<9.92e-25, fold enrichment of 1.99, 1.66;)(Table 2, rows 11,

301 12)(33). The butterflies *J. coenia* and *V. cardui* are exceptions; intriguingly, they show a
302 depletion in SCRMshaw predictions compared to expectation (Table 2, rows 13, 14). Whether
303 this is due to a mismatch in the data used for comparison, the state of the genome assemblies for
304 these two draft genomes, or some other failure of SCRMshaw to perform strongly on these
305 species remains to be determined. Overall, the highly significant overlap we observe between
306 SCRMshaw predictions and open chromatin regions in most species and tissues provides further
307 strong evidence that SCRMshaw is effectively predicting enhancers across a broad range of
308 genomes.

309

310 ***Reporter gene analysis demonstrates that SCRMshaw predictions are functional enhancers***

311 As a concrete test of our ability to use SCRMshaw to predict functional enhancers, we assayed a
312 subset of SCRMshaw predictions by reporter gene analysis in transgenic *D. melanogaster*. Our
313 previous studies have shown a high success rate for such assays, ranging from 65% through
314 >80%, depending on the species and training sets used (13, 14, 17, 19, 20).

315 We focused our *in vivo* validation experiments on SCRMshaw predictions made using the
316 *mapping2.wing, haltere_disc,* and *disc.mapping2* training sets and a set of four species we had
317 previously shown to be amenable to SCRMshaw prediction*: D. melanogaster, A. aegypti, T.*
318 *castaneum,* and *A. mellifera*. The imaginal discs are well-established tissues for investigating
319 gene regulation in *D. melanogaster*, and the chosen training sets all gave significant results in the
320 open-chromatin comparisons discussed above. We selected six sets of putative enhancers for

321 testing (Table 3, Table 4). For the first three sets, we selected sequences where we had
322 predictions in each of the orthologous loci for *D. melanogaster, T. castaneum,* and *A. mellifera*
323 (*A. aegypti* was not considered for these sets), and where the *D. melanogaster* prediction mapped
324 near a gene expressed in the wing imaginal discs (Fig. 5). These predictions were conducted
325 using SCRMshaw's IMM scoring method only, with post-processing performed using the
326 original method described in (21), and the Amel_4.5 version of the *A. mellifera* genome. For the
327 second three sets, we chose sequences where we had predictions in orthologous loci for at least
328 three of the four species and where the *D. melanogaster* sequence had previously been tested in a
329 reporter gene assay and was known to be active in the relevant imaginal discs (Fig. 5). Imposing
330 this latter criterion allowed us to leverage existing knowledge and reduce the necessary amount
331 of *in vivo* testing for each set of predictions, enabling us to test a larger set of sequences overall.
332 For this second set of three, we used all three SCRMshaw scoring methods with a revised post-
333 processing algorithm (see Methods), and the Amel_Hav3.1 *A. mellifera* genome. For all six sets,
334 predictions were chosen for testing based on high SCRMshaw scores and overlap with open-
335 chromatin data (where available). We also considered the position of each prediction within the
336 locus (i.e., first intron, downstream intergenic region, etc.), favoring sequences where position
337 was maintained among the orthologs. Selected sequences were cloned into a cross-species
338 compatible reporter vector (Deem et al., *submitted*)(18), and reporter gene activity was
339 visualized either directly or by using the lineage tracing system G-TRACE (34). In the latter,
340 enhancer activity is visualized through two reporters; the first reporter visualizes the direct
341 enhancer activity while expression of the second reporter is induced and maintained in all cells
342 that descend from a cell in which the enhancer is initially active, even if activity subsequently
343 shuts off.

344

345 *ex*

346 The *expanded (ex*) gene plays a crucial role in tissue growth control, including wings (35-37).
347 There was a single prediction in the *D. melanogaster ex* locus, *Dm_ex_20p0*, which falls within
348 an open chromatin region of the third intron (Supplemental Fig. S2), and which comprises an
349 untested subsequence of a longer sequence that acts as an imaginal disc enhancer (Fig. 5B; (37)).
350 This sequence drove reporter activity in the wing, leg, and antennal discs (Fig. 6A). The *T.*
351 *castaneum* genome also had only one prediction for the *ex* locus, *Tc_ex_9p0*, which overlaps
352 well with a FAIRE-seq peak (Supplemental Fig. S2) within the second intron. Although no
353 imaginal disc activity was observed (Fig. 6B), *Tc_ex_9p0*-driven reporter gene expression was
354 observed during late pupal stages in the legs (Supplemental Fig. S4G). The *A. mellifera* genome
355 (v4.5) had one prediction, *Am_ex_20p3*, within the fourth intron of the *ex* locus (Supplemental
356 Fig. S2; however, note that subsequent prediction using the updated Amel_Hav3.1 genome and
357 revised SCRMshaw post-processing yielded additional predictions in this locus). *Am_ex_20p3*
358 drove active but variable expression in both the pouch and notum regions of the wing imaginal
359 disc (Fig. 6C). Although often significantly limited to a small number of cells, the pouch
360 expression of *Am_ex_20p3* was similar in pattern to that seen with *Dm_ex_20p0* (cf. Fig. 6A).
361 Weak and inconsistent expression in the pouch region of the haltere discs was also observed. In
362 addition, *Am_ex_20p3* drove expression in the leg discs (Fig. 6C).

363

364 *klu*

365 *Klumpfuss (klu*) encodes a zinc finger protein important for proper tissue specification and
366 differentiation (38). The *D. melanogaster* genome had three predictions for the *klu* locus
367 (Supplemental Fig. S2). We selected *Dm_klu_16p1*, which overlaps a region of open chromatin
368 within the second intron present in most imaginal discs (most distinct in the T3 leg disc;
369 Supplemental Fig. S2). *Dm_klu_16p1* drove active expression in the notum regions of the wing
370 and haltere discs and in the leg discs (Fig. 6D). *Tc_klu_8p6* was the only prediction for the *T.*
371 *castaneum klu* locus. It falls within the second intron, similar to *D. melanogaster's klu* prediction
372 *Dm_klu_161p1* (Supplemental Fig. S2), but lacked enhancer activity (Fig. 6E). *A. mellifera* had
373 only one prediction for the *klu* locus (*Am_klu_20p2*) (Supplemental Fig. S2). We included this
374 prediction for validation even though its location in the third intron does not match with that in
375 the other two species. *Am_klu_20p2* drove weak expression in the most proximal part of the leg
376 discs (Fig. 6F, arrows), but no activity was observed in the other imaginal discs. (Note that
377 additional predictions were subsequently produced with our revised post-processing algorithm
378 and the updated Amel_Hav3.1 genome, for both the *T. castaneum* and *A. mellifera klu* loci).

379

380 *ush*

381 Our final choice from our first set for *in vivo* validation was *u-shaped* (*ush*), which encodes a
382 transcription factor with described imaginal disc expression (39-41). The *D. melanogaster*
383 genome had two predictions for the *ush* locus (three when using the updated post-processing

384 algorithm) (Supplemental Fig. S2). We tested *Dm_ush_16p4*, but did not observe any larval disc
385 activity (Fig. 6G). *Tc_ush_6p8* and *Am_ush_20p8* were the only predictions for *T. castaneum* and
386 *A. mellifera*, respectively, both located in the second intron (again, additional predictions are
387 found using updated methods and genome versions). *Tc_ush_6p8* had expression in the wing disc
388 as well as weak activity in the peripodial membrane of the eye-antennal disc (Fig. 6H,
389 Supplemental Fig. S4K). *Am_ush_20p8* displayed active expression in the peripodial membrane
390 surrounding the eye disc, as well as in a single cell, or small subset of cells, located at the center
391 of each leg disc (Fig. 6I, arrows).

392

393 *hth*

394 SCRMshaw predicted ten enhancers in the locus of the *D. melanogaster* Hox cofactor
395 *homothorax* (*hth*). One of the two top-scoring predictions, *Dm_hth_30p1*, located in the fifth
396 intron, overlapped the known *hth_GMR46D04* enhancer, which has activity in all of the larval
397 imaginal discs (Fig. 5F, Supplemental Fig. S3)(42). *A. aegypti* had a single prediction in the
398 orthologous *hth* locus, *Aa_hth_35p9*, located in the first intron (Supplemental Fig. S3). This
399 sequence failed to display activity in imaginal discs (Fig. 7A).

400 *T. castaneum* had 23 *hth*-locus predictions. We selected a high-scoring (albeit not the highest-
401 scoring) sequence, *Tc_hth_15p5*, due to the similarity of its position to that of the *D.*
402 *melanogaster* enhancer—in the fifth intron—and overlapping FAIRE peak (Supplemental Fig.
403 S3)(18). The *Tc_hth_15p5* reporter showed activity in the presumptive notum region of the wing

404 disc and in the proximal region of the leg discs, resembling both the endogenous *D.*
405 *melanogaster hth* expression pattern and that of the *GMR46D04* enhancer (Fig. 7B, cf. Fig. 5F).
406 No enhancer activity was observed in the haltere or eye-antennal discs (Fig. 7B). In the pupal
407 stage, *Tc_hth_15p5* exhibited active expression along the thorax, corresponding to adult *hth*
408 expression in *D. melanogaster* (Supplemental Fig. S4I, arrows) (43). No *hth* predictions were
409 obtained for *A. mellifera*.

410

411 *Ubx*

412 The classic homeotic gene *Ultrabithorax* (*Ubx*) regulates tissue identity in the thoracic and
413 abdominal segments (44). Of six SCRMshaw predictions in the *D. melanogaster Ubx* locus, we
414 noted that one, *Dm_Ubx_36p1*, overlaps a cluster of known enhancers in the third intron centered
415 on *Ubx_GMR39A02* and *Ubx_abx6.8* (Supplemental Fig. S3). *Ubx_GMR39A02* mirrors the
416 haltere activity of *Ubx*, but also displays ectopic activity in the pouch and notum regions of the
417 wing disc (Fig. 5H)(42). Similarly, *Ubx_abx6.8* also drives both native and ectopic expression in
418 the imaginal discs (45).

419 There were eight predictions in the *Ubx* locus of the *A. aegypti* genome (Supplemental Fig. S3).
420 We selected sequence *Aa_Ubx_26p0*, as it was both the highest scoring prediction and was in the
421 third intron, similar to its putative *D. melanogaster* counterparts. Although direct reporter
422 expression from *Aa_Ubx_26p0* was too weak to observe directly, use of the lineage-tracing G-
423 TRACE system confirmed widespread activity in all leg discs (Fig. 7C).

424 We chose two *T. castaneum* sequences for testing, out of ten predictions: *Tc_Ubx_17p4*, which
425 overlapped well with an accessible chromatin region in the third thoracic epidermal tissue and
426 the central nervous system (18), and *Tc_Ubx_19p9*, which had the highest local prediction score
427 but only overlapped with a chromatin region accessible predominantly during early
428 embryogenesis (Supplemental Fig. S3). *Tc_Ubx_17p4* displayed activity in all three leg discs
429 (Fig. 7D). The expression in T2 and T3 leg discs corresponds to endogenous *Ubx* expression,
430 whereas T1 leg disc expression appears to be ectopic. *Tc_Ubx_19p9,* predicted using the
431 "haltere" training set, did not drive clear expression in the haltere disc but did drive G-TRACE
432 expression in the wing disc in the adepithelial adult myoblast cells, as well as direct reporter
433 expression in the peripodial membrane of the eye disc. Very limited expression was also
434 observed in the leg discs when using G-TRACE (Fig. 7E, Supplemental Fig. S4J).

435 The *A. mellifera* genome had eleven predictions at the *Ubx* locus (Supplemental Fig. S3).
436 *Am_Ubx_37p2* was chosen based on a high SCRMshaw score, although it falls within the fourth,
437 rather than the third, intron. *Am_Ubx_0p39*, on the other hand, was in the corresponding third
438 intron location. *Am_Ubx_0p39* had activity in the pouch region of the wing and haltere discs, as
439 well as in the proximal region of leg discs (Fig. 7F). *Am_Ubx_37p2* drove expression in specific
440 portions of the wing and haltere discs (Fig. 7G). Although *Ubx* is not expressed in the *D.*
441 *melanogaster* wing disc (i.e., the forewing of the fly), wing activity has been observed previously
442 with tested *Ubx* enhancer fragments (e.g. 42). Moreover, *Ubx* is expressed in both the forewing
443 and hindwing discs in honeybees, making the reporter expression we observed driven by the two
444 predicted *A. mellifera* enhancers consistent with their potential native activities (46).

445

446 *psq*

447 We chose *pipsqueak* (*psq*), a transcription factor involved in Polycomb group gene silencing
448 (47), and its orthologous loci as the final targets for enhancer validation. There were four
449 predicted *psq* enhancers in *D. melanogaster*, one of which, *Dm_psq_25p6,* overlaps known
450 enhancer *psq_GMR41E12* (Supplemental Fig. S3). This enhancer is located within the second
451 *psq* intron and drives expression in all imaginal discs (Fig. 5J)(42). The *A. aegypti* genome had
452 four predictions for the *psq* locus (Supplemental Fig. S3); we chose *Aa_psq_21p5*, located in the
453 second intron and with the highest local SCRMshaw score, for validation (Supplemental Fig.
454 S3). However, *Aa_psq_21p5* did not have observable imaginal disc expression (Fig. 7H).

455 *T. castaneum* had only two predictions for the *psq* locus, both within the third intron
456 (Supplemental Fig. S3). *Tc_psq_19p7*, which was chosen for validation due to its higher score,
457 drove expression in the eye and leg discs (Fig. 7I).

458 From the *A. mellifera* genome, we selected the only prediction for the *psq* locus, *Am_psq_29p2*
459 (Supplemental Fig. S3). *Am_psq_29p2* reporter activity was negative in all tissues assayed (Fig.
460 7J).

461

462 *Embryonic activity:*

463 Although our SCRMshaw predictions were targeted toward imaginal disc activity, activity was
464 also observed in embryos for many of the reporter lines (Supplemental Fig. S5). Analysis of this
465 activity was complicated by the fact that our reporter vectors, while not having any basal activity
466 in imaginal discs (Supplemental Fig. S4, A-E), displayed reporter gene expression in several
467 tissues including hemocytes, caudal visceral mesoderm, and the proventriculus, even in the
468 absence of a putative enhancer sequence (Supplemental Fig. S5A-D). In those lines that had
469 reporter gene expression in other tissues (Supplemental Fig. S5I-S, X, Y), most of that
470 expression was not clearly associated with the expected endogenous expression of the predicted
471 target gene, although the complex embryonic expression patterns of these genes makes a
472 definitive assessment difficult. Moreover, for most of the species, we do not currently know the
473 expression patterns of either the gene in its native species (as opposed to the expression of its
474 *Drosophila* ortholog), or the expression patterns of other nearby potential target genes. Further
475 analysis, including additional control experiments, use of different reporter vectors, and
476 assessment of gene expression patterns in each of the relevant species will be necessary before
477 drawing final conclusions as to embryonic enhancer activity.

478

479 ***An insect regulatory annotation resource***

480 Taken together, the results from our simulations, our comparisons to open chromatin regions, and
481 our *in vivo* validation experiments demonstrate that SCRMshaw is remarkably effective at
482 predicting regulatory sequences across a wide range of insect species. To facilitate access to our
483 SCRMshaw-based regulatory annotations, we created a database with the results from our
484 predictions using all training sets and all completed species. This database, which is freely

accessible as part of the REDfly insect regulatory annotation site (16), contains processed final prediction data and can be searched and filtered by gene, *D. melanogaster* ortholog, training set, enhancer location, and various other criteria. We will continue to add to this database as additional species and training sets are run through our annotation pipeline.

DISCUSSION

The resource we introduce here is part of an ongoing effort to provide an initial regulatory annotation for all sequenced insects. These annotations are not complete, as we currently lack well-curated training sets for many tissues, including key embryonic tissues such as the central nervous system and many non-embryonic tissues. We aim to generate the necessary additional training sets over time, and add these new annotations to those presented here, along with comprehensive annotation for additional species. As a predictive pipeline, SCRMshaw is subject to the usual tradeoffs of sensitivity versus specificity in generating results. Sensitivity is difficult to assess, as there is no "complete" known set of enhancers for any organism, and it is currently impossible to make an accurate estimate of what the yield should be for any of our training models. Similarly, without comprehensive *in vivo* testing using multiple conditions and methods, an accurate false-positive rate cannot be computed. However, we presented here several lines of evidence demonstrating that true positives significantly outweigh false positives: we non-randomly predict multiple enhancers for specific subsets of loci; we predict enhancers in orthologous loci at a rate significantly higher than random expectation; our predictions are highly

505 enriched for regions of accessible chromatin; and we achieve a high rate of validation using *in*
506 *vivo* reporter gene assays.

507

508 ***Validation success rates***

509 The results from the *in vivo* validation experiments are summarized in Table 5. Overall, 17/22
510 (77%) of tested sequences revealed imaginal disc activity, consistent with our previous
511 SCRMshaw success rates. 59% of these (10/17), or 45% of the total (10/22), had the correct
512 target specificity of wing and/or haltere discs. This is again consistent with previous SCRMshaw
513 experience, which shows that functional enhancers are predicted at a higher success rate than
514 enhancers with specific targeted activity (13, 14).

515 These numbers may in fact underestimate the success rate of our predictions, for several reasons.
516 One, all of the testing was performed in transgenic *D. melanogaster*, despite the putative
517 enhancers being from three additional species. Reduced efficiency of certain transcription factor
518 or cofactor binding, reduced enhancer-promoter compatibility, or other species-specific
519 differences may lead to elevated false negative results. Two, predictions for *ex, klu*, and *ush* were
520 made using a less-effective SCRMshaw post-processing algorithm and, for *A. mellifera*, a less-
521 complete genome build. This may have led to suboptimal candidate enhancer selection. Three, it
522 is possible that some of our predicted enhancers act as silencers rather than enhancers. Silencers,
523 which attenuate rather than promote gene expression, are less well understood than enhancers,
524 but in at least some instances have identical sequence characteristics (and in fact can act

525 simultaneously as enhancers in some tissues and silencers in others) (48, 49). Our reporter gene
526 assay was not designed to detect silencers, which would therefore appear as false-positive
527 predictions. Indeed, an intriguing possibility is that some of our sequences that had reporter gene
528 activity in non-targeted discs (e.g., leg discs) act as enhancers in those tissues but as silencers in
529 the targeted wing discs. Alternatively, identified enhancers that drove expression in multiple
530 imaginal discs may simply represent pleiotropic enhancers. Enhancer pleiotropy is not
531 uncommon (50, 51), and moreover, there is overlap in the enhancer content of our training sets
532 for the various discs. There are many gene expression and developmental similarities among the
533 discs, and likely significant regulatory overlap; chromatin profiling experiments have found that
534 wing, leg, and eye discs share the majority of their regions of accessible chromatin (18, 32).
535 Additional experiments will be necessary to distinguish between pleiotropic enhancers, dual-
536 function silencers/enhancers, and other possibilities.

537

538 ***Redundant enhancers***

539 Almost all of our training sets predicted multiple enhancers in certain loci at rates exceeding
540 random expectation, consistent with the noted prevalence of redundant or "shadow" enhancers in
541 numerous plant and animal species (reviewed by 27). The few training sets that did not may
542 reflect types of tissues or regulatory processes for which having shadow enhancers is unusual, or
543 may be indicative of poorly-performing training sets. Indeed, the sole training set that did not
544 suggest the presence of shadow enhancers in either mosquito species, *adult_PNS*, also performed

545 poorly by other metrics, such as failing to produce predictions in common over multiple species
546 and scoring "poor" using our *pCRM_eval* training set evaluation tool (21).

547 Although the evolutionary origins of shadow enhancers are not well understood, several studies
548 suggest that at least one important contribution of shadow enhancers is to provide phenotypic
549 robustness during development, particularly during stress conditions (52-57). In some cases, a
550 mechanistic basis for this robustness has been suggested by the finding that different groups of
551 transcription factors act on individual members of a shadow enhancer set (58, 59). Further
552 support for this idea can be found in the fact that shadow enhancers appear to have limited
553 sequence similarity (58, 60). However, our successes with SCRMshaw have demonstrated that
554 enhancers can have little overt sequence similarity but still rely on the same underlying
555 subsequence (and presumably, transcription factor binding) model. As the potential shadow
556 enhancers we identify are predicted using the same training set, they are likely to be functioning
557 using similar, rather than independent, regulatory mechanisms. Thus, shadow enhancers appear
558 to come in at least two flavors: those that use similar mechanisms, such as those identified here,
559 and those that use different mechanisms, such as the *Krüppel* enhancers studied by (58).

560

561 ***Enhancer evolution at large divergence distances***

562 SCRMshaw's ability to find putatively "orthologous" enhancers—i.e., enhancers in orthologous
563 loci predicted using the same regulatory model—not only substantiates the non-random nature of
564 our prediction approach, but also opens the door to exciting studies of enhancer evolution over

565 large divergence ranges. We previously illustrated the power of such an approach with an
566 analysis of a small number of enhancers in just 2-3 highly diverged species (17). However, the
567 greatly increased number of species for which we now have regulatory predictions should allow
568 us to follow the evolution of specific enhancers over their entire divergence range. Although the
569 number of common loci tails off rapidly as the number of species considered increases, we
570 suspect that this small (albeit statistically significant) number understates the true results. For
571 one, only a limited number of species have been evaluated for common predictions so far, and
572 these are not evenly distributed along the phylogenetic spectrum of sequenced species. Also, our
573 analysis depends wholly on the presence of recognized *D. melanogaster* orthologous genes to
574 define the common loci. This in turn is dependent on several factors, including the sensitivity
575 and accuracy of our ortholog-calling pipeline, the reliability of the protein annotations we use in
576 that pipeline, and on accurate target gene assignments. Each of these has known sources of error.
577 For example, our ortholog-calling pipeline currently does not disambiguate multiple paralogs
578 (see Methods), and ortholog detection has been shown to be sensitive to the method used for
579 generating protein annotation, in particular when different annotation methods have been applied
580 to different species in the comparison (61). Moreover, our target gene assignments are presently
581 based solely on closest-gene relationships, a method which is known to mis-assign a fair number
582 of enhancer-gene target pairings (e.g. 62, 63, 64) and which does not take into account
583 complexities in genome architecture such as nested genes, long non-coding RNAs, promoter
584 competition, and the like (which are common in *D. melanogaster*, e.g. 65, 66). Developing
585 effective and scalable solutions to these issues will be an important goal for future work.

586 As the number of species in our prediction database with fully mapped orthologs grows, it will
587 become possible to ask increasingly sophisticated questions about the nature of enhancer

588 evolution. For instance, we will be able to determine whether the numbers of putatively
589 orthologous enhancer predictions follow phylogenetic relationships and degree of sequence
590 divergence, and whether certain loci are only found in common for specific groups of species.
591 Also in need of further investigation will be to determine whether sets of common enhancers are
592 restricted to certain functional Gene Ontology categories, and how this might vary with
593 phylogenetic grouping.

594

595 ***Ongoing regulatory annotation***

596 While effective, SCRMshaw still has various limitations and aspects in need of improvement.
597 Errors in genome assembly and insufficient repeat masking both appear to contribute to overly
598 long predictions (multiple kilobases) that are unlikely to represent individual single enhancers
599 (HA and MSH, unpublished observations). Poor assembly, while not having major effects on
600 SCRMshaw overall, can significantly affect results at specific loci, as can inaccurate gene
601 annotation (21). Also requiring further investigation is how to best combine and weight the
602 scores from the three individual SCRMshaw scoring methods of IMM, hexMCD, and PAC-rc.
603 Individual SCRMshaw predictions should therefore be treated as just that—predictions—and
604 appropriate validation experiments are recommended for any sequences of interest. The
605 regulatory annotations presented here represent initial "1.0" versions of the regulatory genome.
606 Like all genome annotations, these will continue to be revised as updated models become
607 available, including addition of new training sets, confidence scores based on validation
608 experiments, and improved genome builds. Nevertheless, the enrichment scores from our *in*

609 *silico* experiments and the true-positive rates from our *in vivo* reporter gene experiments suggest
610 that overall, true positive rates for most training sets are between 50-85%, with most likely
611 having success rates exceeding 70%. Note that even this lower bound of a 50% true positive rate
612 means that one out of every two predictions is correct—a more than acceptable rate to encourage
613 follow-up experiments for predicted enhancers in organisms of interest. The catalog of predicted
614 enhancers we introduce here, spanning 33 insect species and growing, should thus be useful
615 resource for both large-scale and small-scale studies of the insect regulatory genome.

616

617 **METHODS**

618 **Datasets**

619 Sequence and annotation for *D. melanogaster* were obtained from FlyBase (67). For other
620 species, wherever possible, genome sequence (FASTA) and annotation (GFF) files were
621 downloaded from NCBI at www.ncbi.nlm.nih.gov/datasets. Otherwise, the genome sequences
622 and annotations were obtained from the primary literature or directly from the data generators
623 (see Table 1 for references). Version numbers for genomes and annotations are provided in Table
624 1.

625

626 **SCRMshaw**

627 A detailed SCRMshaw protocol can be found at Asma et al., 2024 (*in preparation*). Genome
628 files were masked using Tandem Repeat Finder (22) using parameters 2 7 7 80 10 50 500 -m -h.
629 Genome and annotation files were then assessed using our *preflight* script. Any sequence
630 scaffolds not containing annotated genes were removed before passing the genome sequence to
631 the main SCRMshaw program. SCRMshaw was run using the "HD" default settings as described
632 in Asma and Halfon (21). SCRMshaw can be downloaded at
633 https://github.com/HalfonLab/SCRMshaw_HD.

634

635 **Postprocessing**

636 The top 5000 hits were extracted using scripts *Generate_top_N_SCRM-hits.pl* and
637 *concatenatenatingOffsetResults.sh* and passed to *postProcessingScrmshawPipeline.py* with *-num*
638 = "5000" and *-topN* = "Median." Post-processing was performed essentially as previously
639 described (15), with the following modifications (Supplemental Fig. S6): before evaluating each
640 10 bp region, scores from each of the 25 individual SCRMshaw instances were assessed and any
641 500 bp window whose score was below the value of the 5000th ranked score was eliminated by
642 having its score reset to zero (Supplemental Fig. S6B). The "elbow" point of the SCRMshaw
643 score curve of the 5000 top scores from each instance was then determined (Fig. S6C), and any
644 scores below the elbow point were reset to zero (Supplemental Fig. S6D; gray boxes). Only after
645 these two rounds of score evaluation were all windows grouped together (Fig. S6E, F) and
646 subjected to peak calling on 10 bp intervals (Fig. S6G). Final "top predictions" were then any
647 peaks with an amplitude above the selected amplitude threshold (elbow point of amplitude curve,

648 represented by a red dot in Supplemental Fig. S6H), following the peak-calling step. See (21)
649 and Asma et al. 2024 (*in preparation*) for details.

650 All scripts are available at https://github.com/HalfonLab/.

651

652 **Orthology mapping**

653 The final SCRMshaw output from the above steps was used as input to our orthology mapping
654 pipeline. For each species, a FASTA-formatted file of all annotated proteins was downloaded
655 from NCBI (www.ncbi.nlm.nih.gov/datasets). The annotated proteins from *D. melanogaster* plus
656 each individual other species were used as input to *Orthologer* (24) to obtain the *Drosophila*
657 ortholog for each protein (when existing). Our approach was designed to be minimally restrictive
658 in that we did not enforce a one-to-one ortholog mapping; in cases of likely paralogs, we
659 considered all of the paralogs as a potential result. Details on the orthology mapping protocol can
660 be found in Asma et al. 2024 (*in preparation*).

661

662 **Evaluating the number of predictions per locus**

663 For each training set, BEDTools "merge" was used to remove any overlapping predictions (68).
664 The results were then permuted 1000 times using BEDTools "shuffle", with coding regions

665 excluded. BEDTools "closest" was used to assign upstream and downstream flanking genes to
666 each of the permuted predictions, with the following parameters: the '-io' flag was enabled to
667 ignore any overlaps between predictions and genes, and '-D ref -id' and '-D ref -iu' were used to
668 obtain the closest 5' and 3' genes (with respect to chromosome coordinates), respectively. A
669 custom Python script, *checkSameLocus_ForSimulations.py* (located at
670 https://github.com/HalfonLab/UtilityPrograms/blob/master/checkSameLocus_ForSimulations.py
671) was used to calculate the number of predictions per locus for both the real and permuted results.
672 For this purpose, "locus" was defined as the entire region between the left and right flanking
673 genes for each SCRMshaw prediction; that is, for each prediction we take the region between the
674 3' end of the upstream gene and the transcription start site of the downstream gene. If a
675 prediction is within an intron, we define the locus as the entire span of the enclosing gene. If a
676 prediction overlaps two genes, the locus is considered to be the entire span of the two genes.
677 Average intergenic region sizes for the genomes were estimated using the distances between
678 neighboring genes, discarding any nested or overlapping gene pairs, as provided by the
679 SCRMshaw *preflight* script.

680 Significance was assessed by calculating the empirical *P*-value, defined as the number of real
681 loci with a number of predictions greater than the maximum number obtained from the 1000
682 simulations, for each locus containing at least one SCRMshaw prediction.

683

684 **Evaluating the number of predictions in common across species**

685 To determine the expected number of common predictions—i.e., predictions in orthologous
686 loci—across species, we utilized SCRMshaw predictions for 16 species. Each set of predictions
687 was sorted, merged, permuted, and mapped to new loci as described above for "Evaluating the
688 number of predictions per locus." The permuted predictions were used as input for the script
689 "*checkSameLocus_ForSimulations_crossSpecies.py*" (located at
690 https://github.com/HalfonLab/UtilityPrograms/blob/master/checkSameLocus_ForSimulations_cr
691 ossSpecies.py). This script identifies the *Drosophila* orthologs of the nearest flanking genes for
692 each species and calculates the number of common loci flanking the simulated SCRMshaw
693 predictions for 5, 10, 11, 12, 13, 14, 15, and 16 species. For each species, the process was
694 repeated for a total of 360 permutations. The mean and standard deviation of the permuted
695 results were then used to calculate a *z*-score for each training set. We considered training sets
696 with z-score ≥ 1.645 to be significant ($P<0.05$, not corrected for multiple testing).

697

698 **Overlap between SCRMshaw predictions and open chromatin regions**

699 Open chromatin data were obtained from the following sources:

700 *D. melanogaster* : Data for *D. melanogaster* were downloaded from GEO and consisted of
701 ATAC-seq and FAIRE-seq data from accessions GSE101827 , GSE38727 and GSE118240.
702 These assays were performed using blastoderm embryos, eye-antennal discs, wing discs, haltere
703 discs, leg discs, third instar central nervous system, and wing, leg, and haltere pharate
704 appendages (31, 32, 69).

705 *T. castaneum*: FAIRE-seq data for three stages of embryogenesis, larval central nervous system,
706 and larval second and third thoracic epidermal tissues were downloaded from GEO
707 (GSE104495)(18). The FAIRE profiles were remapped to the version 5.2 of the *T. castaneum*
708 genome (Tcas5.2) for this study. The remapped FAIRE profiles are available on iBeetle-Base
709 (https://ibeetle-base.uni-goettingen.de/)(70).

710 *A. gambiae*: ATAC-seq data for adult midgut and salivary gland were downloaded from GEO
711 (GSE152924)(71).

712 *D. plexippus, J. coenia, H. himera, and V. cardui*: ATAC-seq data for larval forewing and
713 hindwing tissues at stage M5 were provided by Anyi Mazo-Vargas and Robert Reed (33).

714 The overlap between SCRMshaw predictions and open chromatin regions was determined using
715 BEDTools "intersect" with parameters -wa -u -f 0.1 such that sequences needed to overlap at
716 least 10% of their length and were considered a single overlap in the event that more than one
717 open chromatin peak overlapped a prediction.

718 To assess significance, the SCRMshaw predictions were permuted 500 times using BEDTools
719 "shuffle" and the overlaps with open chromatin regions assessed as above. The mean and
720 standard deviation of the permuted results were then used to calculate a *z*-score for each training
721 set.

722 We also calculated a "fold enrichment" score by dividing the observed number of overlapping
723 regions by the expected number, to provide a sense of effect size in addition to statistical
724 significance.

725

726 **Reporter constructs and transgenic *Drosophila***

727 Sequences for reporter gene analysis, including attL1 and attL2 sites, were synthesized *de novo*
728 and cloned into pUC57 Kan-r (GenScript, Piscataway NJ) as entry vectors suitable for Gateway
729 cloning (72). Gateway LR recombination was then used to move the sequences into
730 piggyPhiGUGd (*hth, Ubx,* and *psq* lines) and piggyPhiGUGd-TomatoI (*ex, klu,* and *ush* lines)
731 (Deem et al., *submitted*). Transgenic flies were generated by BestGene (Chino Hills, CA) using
732 PhiC31 recombination and the attP2 third chromosome insertion site. piggyPhiGUGd lines were
733 subsequently crossed to G-TRACE for visualizing enhancer activities (34).

734 For each construct, imaginal discs from at least six larvae were dissected and mounted for direct
735 fluorescence visualization using a Zeiss Axio Imager M2 microscope with ApoTome 2. Embryos
736 were fixed and stained using standard *Drosophila* methods using anti-dsRed (Clontech) and
737 visualized using the ABC-HRP kit (VectorLabs, Newark CA).

738

739 **Database implementation**

The SCRMshaw results database is implemented as part of REDfly (16), a MariaDB-based database hosted on a private OpenStack cloud infrastructure maintained by the University at Buffalo Center for Computational Research. As part of an ongoing transition of REDfly to a more modern software architecture, backend functions are implemented in Node.JS and Python, while frontend components utilize React.JS and Next.js. GraphQL is used as the query language. REDfly is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives License v4 International (CC BY-NC-ND 4.0) and its underlying source code under a GNU General Public License v3 (GNU GPL 3.0).

Image credits

The insect silhouettes in Table 5 were obtained from The Noun Project (thenounproject.com), artist Georgiana Ionescu, under a Creative Commons CC-BY 3.0 license.

ACKNOWLEDGMENTS

We thank Bob Reed, Anyi Mazo-Vargas, and Tom Williams for sharing data and results, Jack Leatherbarrow for technical support, members of the Halfon and Tomoyasu labs for helpful discussion and advice, and Tom Williams for comments on the manuscript. SCRMshaw analyses were run using the resources of the University at Buffalo Center for Computational Research. The Center for Bioinformatics and Functional Genomics at Miami University provided instrumentation and technical support.

Funding:

Miami University Faculty Research Grants Program (CFR) (to Y.T.), the National Science Foundation (NSF) (grant IOS1557936 to Y.T.), National Institutes of Health (NIH) (grant U24 GM142435 to M.S.H.), and the U.S. Department of Agriculture (USDA) (grant 2018-08230 to M.S.H. and Y.T.).

772 **FIGURE LEGENDS**

773 **Figure 1: The SCRMshaw method and analysis pipeline.** (A) Supervised motif-blind CRM
774 discovery (SCRMshaw). (a) SCRMshaw uses a training set of known *D. melanogaster* enhancers
775 ("training sequences"), drawn from REDfly, that are defined by common functional
776 characterization, and a 10-fold larger background set of similarly sized common functional
777 characterization, non-enhancer sequences ("background sequences"). (b) The short DNA
778 subsequence (*kmer*) count distributions of these sequences are then used to train a statistical
779 model. The trained model (c) is used to score overlapping windows in the "target genome." (d)
780 High-scoring regions are predicted to be functional regulatory sequences (asterisks). Figure
781 adapted from (73). (B) The workflow used for the regulatory genome annotation described in this
782 paper. The left side shows pre-processing steps, the right side, post-processing. Input to
783 SCRMshaw consists of the genome sequence and gene annotation. A protein sequence
784 annotation is supplied later for the orthology mapping step. Final results are made available as
785 part of the REDfly regulatory annotation knowledgebase.

786 **Figure 2: Annotation of 33 insect genomes.** (A) Genomes from five insect orders were
787 annotated in this study (more are ongoing). (B) Percentage of genes with *Drosophila* orthologs as
788 mapped via our orthology pipeline (see Methods), for the 15 mapped species. For complete
789 species names, see Table 1. (C) Total SCRMshaw predictions for each species. For each species,
790 the lefthand column shows cumulative results for each SCRMshaw sub-method summed over
791 each of the 48 training sets. The righthand column shows the number of unique predictions after
792 merging overlapping predictions from both sub-methods and training sets. Species are displayed
793 alphabetically by taxonomic order. (D) Size distribution of SCRMshaw predictions, prior to

794 merging overlapping predictions but after removing outlier predictions > 2 kb in length. Species
795 are ordered identically to panel C.

796 **Figure 3: SCRMshaw makes multiple predictions per locus.** The number of SCRMshaw
797 predictions per locus (y-axis) are shown as boxplots for loci falling within the given size ranges
798 (x-axis). Black boxes cover the 25th-75th percentiles, bars indicate median values and dots
799 indicate values exceeding 1.5 times the interquartile range (boxes are not visible for all bins due
800 to very low degrees of variation). Values in pink represent expected values drawn from
801 randomization, while values in blue represent observed values from SCRMshaw. All values are
802 from results with the training set "mapping1.visceral_mesoderm"; results from other training sets
803 were similar (see Supplemental Table S3). Shown are results from the genomes of (A) *D.*
804 *melanogaster*, (B) *C. pipiens*, and (C) *A. aegypti* representing small, medium, and large
805 genomes, respectively.

806 **Figure 4: SCRMshaw predicts CRMs in orthologous loci across species.** (A) The number of
807 loci in common that contain at least one SCRMshaw prediction, for 10 or more species. (B) *z*-
808 scores demonstrating that the number of loci in common with one or more SCRMshaw
809 predictions is significantly higher than expectation, based on 360 randomizations. The small
810 number of common predictions for 14-16 species make these statistics unreliable. Dotted lines
811 indicate *z*-score values representing significance at the (unadjusted) $P<0.005$ and $P<0.05$ levels.
812 (C) Fold enrichment values illustrating the excess of loci in common with one or more
813 SCRMshaw predictions compared to expectation. Dotted line shows 1.5x enrichment.

814 **Figure 5: Previously described gene expression and enhancer activity for select *D.***
815 ***melanogaster* sequences predicted by SCRMshaw.** The lefthand column shows native *D.*
816 *melanogaster* gene expression in imaginal discs (green), while the righthand column shows
817 described enhancer activity (magenta). Gray shading indicates that expression has not been
818 described. Moving clockwise from the left side of each panel are the wing, haltere, leg, and eye-
819 antennal discs. The enhancers whose activities are described in the table are: (B) *ex_BCDE* (37),
820 (F) *hth_GMR46D04* (42), (H) *Ubx_GMR39A02* (42), (J) *psq_GMR41E12* (42).

821 **Figure 6: Reporter gene expression for tested *ex*, *klu*, and *ush* predicted enhancer**
822 **sequences.** Each row shows expression for the indicated construct in (i) wing discs, (ii) haltere
823 discs, (iii) T1 (prothoracic) leg discs, (iv) T2 (mesothoracic) leg discs, (v) T3 (metathoracic) leg
824 discs, and (vi) eye-antennal discs (with eye portion to the left). Enhancer activities were
825 visualized by UAS-tdTomato that was included in the reporter construct. Scale bar is 50 µm for
826 each column.

827 **Figure 7: Reporter gene expression for tested *hth*, *Ubx*, and *psq* predicted enhancer**
828 **sequences.** Each row shows expression for the indicated construct in (i) wing discs, (ii) haltere
829 discs, (iii) T1 (prothoracic) leg discs, (iv) T2 (mesothoracic) leg discs, (v) T3 (metathoracic) leg
830 discs, and (vi) eye-antennal discs (with eye portion to the left). Enhancer activities were detected
831 by the G-TRACE system; magenta represents direct enhancer activity detected by dsRed
832 expression, while green indicates lineage-based GFP expression. Scale bar is 50 µm for each
833 column.

834

841

842 **Table 1:** Species used in this study

843

Scientific name	common name	Order	Assembly version/ Annotation version	link/URL for assembly information
Acyrthosiphon pisum	pea aphid	Hemiptera	racon and v3_wdel	Courtesy Jennifer Brisson (University of Rochester)
Aedes aegypti	yellow fever mosquito	Diptera	L5.2	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7159/
Agrilus planipennis	emerald ash borer	Coleoptera	Apla_2.0	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/224129/
Anopheles gambiae	African malaria mosquito	Diptera	P4.9	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7165/
Aphis gossypii	cotton aphid/melon aphid	Hemiptera	ASM401081v2	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/80765/
Apis mellifera	honey bee	Hymenoptera	HAv3.1	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7460/
Atta cephalotes	leafcutter ant	Hymenoptera	A.ceph_1.0	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/12957/
Atta colombica	leafcutter ant	Hymenoptera	Acol1.0	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/520822/
Bombyx mori	silkworm	Lepidoptera	ASM15162v1	http://ensembl.lepbase.org/Bombyx_mori_asm15162v1/Info/Index
Cimex lectularius	bed bug	Hemiptera	Clec_2.1	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/79782/
Culex pipiens pallens	Northern house mosquito	Diptera	TS_Cpip_V1	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/42434/
Culex quinquefasciatus	Southern house mosquito	Diptera	VIPSU_Cqui_1.0_pri_paternal	https://academic.oup.com/gbe/article/13/3/evab005/6121099
Danaus plexippus	monarch butterfly	Lepidoptera	v3	http://ensembl.lepbase.org/Danaus_plexippus_v3/Info/Index
Drosophila ananassae	fruit fly	Diptera	caf1	http://ftp.flybase.net/genomes/Drosophila_ananassae/
Drosophila melanogaster	fruit fly	Diptera	r6 1.8	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7227/
Drosophila mojavensis	fruit fly	Diptera	caf1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000005175.2/
Drosophila pseudoobscura	fruit fly	Diptera	r3	http://ftp.flybase.net/genomes/Drosophila_pseudoobscura/dpse_r3.03_FB2015_01/

Formica exsecta	wood ant	Hymenoptera	ASM365146v1	https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003651465.1/
Heliconius erato	red postman butterfly	Lepidoptera	v1	http://ensembl.lepbase.org/Heliconius_erato_lativitta_v1/Info/Index
Heliconius melpomene	postman butterfly	Lepidoptera	Hmel2	http://ensembl.lepbase.org/Heliconius_melpomene_melpomene_hmel2/Info/Index
Heliconius himera	false postman butterfly	Lepidoptera	Hed.V1	Courtesy Robert Reed (Cornell University)
Junonia coenia	common buckeye butterfly	Lepidoptera	JC v1.0	Courtesy Robert Reed (Cornell University)
Leptinotarsa decemlineata	Colorado potato beetle	Coleoptera	Ldec_2.0	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7539/
Manduca sexta	tobacco hornworm	Lepidoptera	v1.0	http://ensembl.lepbase.org/Manduca_sexta_msex1/Info/Index
Nezara viridula	southern green stink bug	Hemiptera		https://www.ncbi.nlm.nih.gov/datasets/taxonomy/85310/
Nasonia vitripennis	jewel wasp	Hymenoptera	Psr_1.1	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7425/
Onthophagus taurus	dung beetle	Coleoptera	Otau_2.0	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/166361/
Pseudoatta argentina	leafcutter ant	Hymenoptera	ASM1760752v1	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/621737/
Stomoxys calcitrans	stable fly	Diptera	Stomoxys_calcitrans-1.0.1	https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-021-00975-9?sap-outbound-id=DB0B95C63ABC56834FFE23B3C176FAE498F3D66E&utm_source=hybris-campaign&utm_medium=email&utm_campaign=000_MKQ5584_0000012067_BMCF_AWA_MK01_GL_12915_eToc_Apr21_CTRL&utm_content=EN_internal_24671_20210419&mkt-key=42010A0557EB1EDA9CFC80777B53EE6F
Tribolium castaneum	red flour beetle	Coleoptera	r5.2	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7070/
Trichoplusia ni	cabbage looper	Lepidoptera	tn1	https://www.ncbi.nlm.nih.gov/datasets/taxonomy/7111/
Vanessa cardui	painted lady butterfly	Lepidoptera	Vcar_v1	Courtesy Robert Reed (Cornell University)
Nebria riversi	ground beetle	Coleoptera	v1	https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13409

844

845

846 **Table 2:** Overlap of SCRMshaw predictions with FAIRE-seq and ATAC-seq peaks

Species	Training Set	Profiled Tissue	Overlap (real)	Overlap (random)	s.d.[a]	Z-score	FE[b]
D. melanogaster	haltere_disc	wing, leg, and haltere third instar discs and pharate appendages; eye-antennal disc; third instar CNS	41.82%	14.02%	10.65	21.53	2.98
D. melanogaster	blastoderm.mapping1	blastoderm	40.99%	13.86%	12.35	22.56	2.96
D. melanogaster	mapping2.wing	wing, leg, and haltere third instar discs and pharate appendages; eye-antennal disc; third instar CNS	35.14%	10.69%	10.19	23.01	3.29
T. castaneum	mapping2.wing	embryo, larval thoracic epidermis, larval brain	85.58%	26.18%	10.98	28.88	3.27
T. castaneum	mapping2.ectoderm	embryo, larval thoracic epidermis, larval brain	81.68%	25.48%	12.48	26.29	3.21
T. castaneum	mapping1.ventral_ectoderm	embryo, larval thoracic epidermis, larval brain	81.17%	24.76%	12.56	33.87	3.28
T. castaneum	mapping1.dorsal_ectoderm	embryo, larval thoracic epidermis, larval brain	71.15%	24.69%	12.64	24.07	2.88
T. castaneum	mapping1.ectoderm	embryo, larval thoracic epidermis, larval brain	69.53%	24.66%	11.92	22.35	2.82
A. gambiae	embryonic_midgut	adult midgut, salivary_gland	40.20%	11.66%	7.90	21.56	3.45
A. gambiae	mapping1.salivary_gland	adult midgut, salivary_gland	36.76%	11.43%	8.52	20.55	3.22
D. plexippus	mapping2.wing	larval forewing, hindwing, head	60.43%	30.35%	6.30	8.93	1.99
H. himera	mapping2.wing	larval forewing, hindwing, head	68.34%	41.10%	8.97	10.27	1.66
J. coenia	mapping2.wing	larval forewing, hindwing, head	3.06%	34.91%	8.52	-12.22	0.09
V. cardui	mapping2.wing	larval forewing, hindwing, head	15.46%	36.01%	8.01	-7.13	0.43

847 [a]s.d., standard deviation
848 [b]FE, fold enrichment
849

850 **Table 3:** Gene loci chosen for *in vivo* validation

851

852

D. melanogaster			*T. castaneum*		*A. mellifera*	*A. aegypti*
name	**symbol**	**FlyBaseID**	**RefSeq**	**iBeetleBase**	**RefSeq**	**VectorBase**
expanded	*ex*	FBgn0004583	*gene-LOC657053*	TC012545	*gene-LOC551519*	AAEL001437
u-shaped	*ush*	FBgn0003963	*gene-LOC659918*	TC013689	*gene-LOC100577801*	AAEL020615
klumpfuss	*klu*	FBgn0013469	*gene-LOC103312803*	TC002783	*gene-LOC100577692*	AAEL013544
homothorax	*hth*	FBgn0001235	*gene-Hth*	TC008629	*gene-LOC552079*	AAEL011643
pipsqueak	*psq*	FBgn0263102	*gene-LOC660343*	TC003349	*gene-psq*	AAEL021255
Ultrabithorax	*Ubx*	FBgn0003944	*gene-Ubx*	TC000903	*gene-ubx*	AAEL014032

853

854

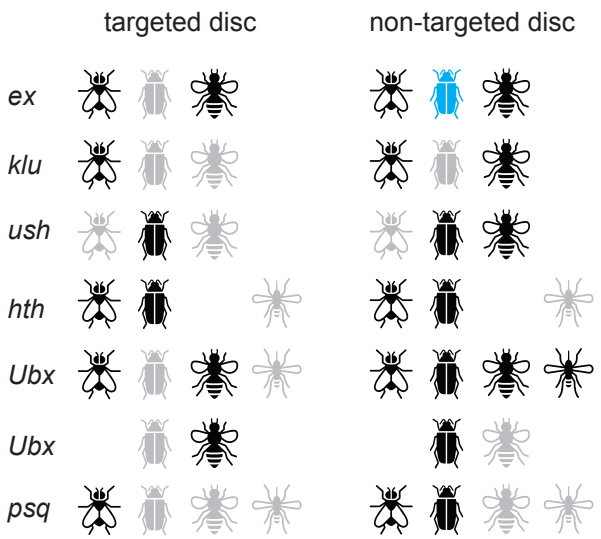
855 **Table 4:** SCRMshaw predictions chosen for *in vivo* validation

856

	Coordinates	Max. score	Training set(s)	Method
Set 1				
T. castaneum				
Tc_ex_9p0	NC_007424.3:11221530..11222170	9.04	mapping2.wing	imm
Tc_ush_6p8	NC_007420.3:5968840..5969370	6.84	haltere_disc	imm
Tc_klu_8p6	NC_007418.3:10416700..10417300	8.59	mapping2.wing	imm
D. melanogaster				
Dm_ex_20p0	2L:442110..442810	20.04	mapping2.wing	imm
Dm_klu_16p1	3L:10991040..10991700	16.06	mapping2.wing	imm
Dm_ush_16p4	2L:531250..532060	16.40	mapping2.wing	imm
A. mellifera				
Am_ex_20p3	NC_007075.3:7545750..7546440	20.36	mapping2.wing	imm
Am_klu_20p2	NC_007070.3:720220..720870	20.25	mapping2.wing	imm
Am_ush_20p8	NC_007080.3:10609550..10610200	20.80	haltere_disc	imm
Set 2				
T. castaneum				
Tc_hth_15p5	NC_007422.5:13408990..13409850	15.52	mapping2.wing	hexmcd,imm,pac
Tc_psq_19p7	NC_007418.3:1805000..1806250	19.71	haltere_disc, mapping2.wing	hexmcd,imm,pac
Tc_Ubx_19p9	NC_007417.3:8137250..8138000	19.32	haltere_disc	hexmcd
Tc_Ubx_17p4	NC_007417.3:8153250..8154030	19.95	mapping2.wing	hexmcd,imm
A. aegypti				
Aa_hth_35p9	1:149733960..149734710	35.96	mapping2.wing	imm
Aa_psq_21p5	2:228190750..228191640	21.50	disc.mapping2	hexmcd,imm
Aa_Ubx_26p0	1:309747490..309748390	26.08	mapping2.wing	imm
A. mellifera				
Am_psq_29p2	NC_007078.3:10712000..10712750	29.26	mapping2.wing	hexmcd
Am_Ubx_37p2	NC_007085.3:2921250..2922250	37.18	haltere_disc, mapping2.wing	hexmcd,imm
Am_Ubx_0p39	NC_007085.3:2967750..2968250	0.39	mapping2.wing	pac

857

Table 5: Summary of *in vivo* validation results



Results are shown for *D. melanogaster, T. castaneum, A. mellifera,* and *A. aegypti.* Black, positive for expression; grey, negative for expression; blue, expression observed in pupae but not in larvae.

883 **REFERENCES**

884 1. NCBI. NCBI Datasets: Genome 2024 [Available from:
885 https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=33208].

886 2. Carroll SB, Grenier JK, Weatherbee SD. From DNA to Diversity. Molecular Genetics and
887 the Evolution of Animal Design. 2nd ed. Malden, MA: Blackwell Publishing; 2005.

888 3. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of
889 morphological evolution. Cell. 2008;134(1):25-36.

890 4. Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment
891 strategies. Trends Mol Med. 2021;27(11):1060-73.

892 5. Rickels R, Shilatifard A. Enhancer Logic and Mechanics in Development and Disease.
893 Trends Cell Biol. 2018;28(8):608-30.

894 6. Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nature structural &
895 molecular biology. 2014;21(3):210-9.

896 7. Suryamohan K, Halfon MS. Identifying transcriptional cis-regulatory modules in animal
897 genomes. Wiley Interdisciplinary Reviews: Developmental Biology. 2015;4(2):59-84.

898 8. Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ
899 significantly in genomic distribution, evolution, and function. BMC Genomics.
900 2019;20(1):511.

901 9. Halfon MS. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and
902 Other Biases. Trends Genet. 2019;35(2):93-103.

903 10. Lindhorst D, Halfon MS. Reporter gene assays and chromatin-level assays define
904 substantially non-overlapping sets of enhancer sequences. BMC Genomics. 2023;24(1):17.

905 11. IUCN. The IUCN list of threatened species 2022 [Available from:
906 https://www.iucnredlist.org.

907 12. Royal Entomological Society. Understanding Insects: Facts and figures St. Albans, UK2023
908 [Available from: https://www.royensoc.co.uk/understanding-insects/facts-and-figures/.

909 13. Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, Robinson GE, et al.
910 Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse.
911 Dev Cell. 2009;17(4):568-79.

912 14. Kazemian M, Zhu Q, Halfon MS, Sinha S. Improved accuracy of supervised CRM discovery
913 with interpolated Markov models and cross-species comparison. Nucleic Acids Res.
914 2011;39(22):9463-72.

915 15. Kazemian M, Halfon MS. CRM Discovery Beyond Model Insects. Methods Mol Biol.
916 2019;1858:117-39.

917 16. Keränen SVE, Villahoz-Baleta A, Bruno AE, Halfon MS. REDfly: An Integrated
918 Knowledgebase for Insect Regulatory Genomics. Insects. 2022;13(7).

919 17. Kazemian M, Suryamohan K, Chen JY, Zhang Y, Samee MA, Halfon MS, et al. Evidence for
920 deep regulatory similarities in early developmental programs across highly diverged insects.
921 Genome biology and evolution. 2014;6(9):2301-20.

922 18. Lai YT, Deem KD, Borras-Castells F, Sambrani N, Rudolf H, Suryamohan K, et al. Enhancer
923 identification and activity evaluation in the red flour beetle, Tribolium castaneum.
924 Development. 2018;145(7).

925 19. Schember I, Halfon MS. Identification of new Anopheles gambiae transcriptional enhancers
926 using a cross-species prediction approach. Insect molecular biology. 2021;30(4):410-9.

927 20. Suryamohan K, Hanson C, Andrews E, Sinha S, Scheel MD, Halfon MS. Redeployment of a
928 conserved gene regulatory network during Aedes aegypti development. Dev Biol.
929 2016;416(2):402-13.

930 21. Asma H, Halfon MS. Computational enhancer prediction: evaluation and improvements.
931 BMC bioinformatics. 2019;20(1):174.

932 22. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res.
933 1999;27(2):573-80.

934 23. Li L, Zhu Q, He X, Sinha S, Halfon M. Large-scale analysis of transcriptional cis-regulatory
935 modules reveals both common features and distinct subclasses. Genome Biology.
936 2007;8(6):R101.

937 24. Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, et al.
938 OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity.
939 Nucleic Acids Res. 2023;51(D1):D445-D51.

940 25. Grosveld F, van Staalduinen J, Stadhouders R. Transcriptional Regulation by
941 (Super)Enhancers: From Discovery to Mechanisms. Annu Rev Genomics Hum Genet.
942 2021;22:127-46.

943 26. Weinstein ML, Jaenke CM, Asma H, Spangler M, Kohnen KA, Konys CC, et al. A novel role
944 for trithorax in the gene regulatory network for a rapidly evolving fruit fly pigmentation trait.
945 PLoS Genet. 2023;19(2):e1010653.

946 27. Kvon EZ, Waymack R, Gad M, Wunderlich Z. Enhancer redundancy in development and
947 disease. Nat Rev Genet. 2021;22(5):324-36.

948 28. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution
949 mapping and characterization of open chromatin across the genome. Cell. 2008;132(2):311-
950 22.

951 29. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native
952 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding
953 proteins and nucleosome position. Nature methods. 2013;10(12):1213-8.

954 30. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted
955 Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.
956 Genome Res. 2007;17(6):877-85.

957 31. Bozek M, Cortini R, Storti AE, Unnerstall U, Gaul U, Gompel N. ATAC-seq reveals regional
958 differences in enhancer accessibility during the establishment of spatial coordinates in the
959 Drosophila blastoderm. Genome Research. 2019;29(5):771-83.

960 32. McKay DJ, Lieb JD. A Common Set of DNA Regulatory Elements Shapes Drosophila
961 Appendages. Developmental Cell. 2013;27(3):306-18.

962 33. Mazo-Vargas A, Langmuller AM, Wilder A, van der Burg KRL, Lewis JJ, Messer PW, et al.
963 Deep cis-regulatory homology of the butterfly wing pattern ground plan. Science.
964 2022;378(6617):304-8.

965 34. Evans CJ, Olson JM, Ngo KT, Kim E, Lee NE, Kuoy E, et al. G-TRACE: rapid Gal4-based
966 cell lineage analysis in Drosophila. Nature methods. 2009;6(8):603-5.

967 35. Boedigheimer M, Laughon A. Expanded: a gene involved in the control of cell proliferation
968 in imaginal discs. Development. 1993;118(4):1291-301.

969 36. Wang LH, Baker NE. Salvador-Warts-Hippo pathway in a developmental checkpoint
970 monitoring helix-loop-helix proteins. Dev Cell. 2015;32(2):191-202.

971 37. Wang LH, Baker NE. Spatial regulation of expanded transcription in the Drosophila wing
972 imaginal disc. PLoS One. 2018;13(7):e0201317.

973 38. Klein T, Campos-Ortega JA. klumpfuss, a Drosophila gene encoding a member of the EGR
974 family of transcription factors, is involved in bristle and leg development. Development.
975 1997;124(16):3123-34.

976 39. Buchberger E, Bilen A, Ayaz S, Salamanca D, Matas de Las Heras C, Niksic A, et al.
977 Variation in Pleiotropic Hub Gene Expression Is Associated with Interspecific Differences in
978 Head Shape and Eye Size in Drosophila. Mol Biol Evol. 2021;38(5):1924-42.

979 40. Cubadda Y, Heitzler P, Ray RP, Bourouis M, Ramain P, Gelbart W, et al. u-shaped encodes a
980 zinc finger protein that regulates the proneural genes achaete and scute during the formation
981 of bristles in Drosophila. Genes Dev. 1997;11(22):3083-95.

982 41. Tomoyasu Y, Ueno N, Nakamura M. The decapentaplegic morphogen gradient regulates the
983 notal wingless expression through induction of pannier and u-shaped in Drosophila. Mech
984 Dev. 2000;96(1):37-49.

985 42. Jory A, Estella C, Giorgianni MW, Slattery M, Laverty TR, Rubin GM, et al. A survey of
986 6,300 genomic fragments for cis-regulatory activity in the imaginal discs of Drosophila
987 melanogaster. Cell reports. 2012;2(4):1014-24.

988 43. Aldaz S, Morata G, Azpiazu N. Patterning function of homothorax/extradenticle in the thorax
989 of Drosophila. Development. 2005;132(3):439-46.

990 44. Lewis EB. A gene complex controlling segmentation in Drosophila. Nature.
991 1978;276(5688):565-70.

992 45. Simon J, Peifer M, Bender W, O'Connor M. Regulatory elements of the bithorax complex
993 that control expression along the anterior-posterior axis. EMBO J. 1990;9(12):3945-56.

994 46. Prasad N, Tarikere S, Khanale D, Habib F, Shashidhara LS. A comparative genomic analysis
995 of targets of Hox protein Ultrabithorax amongst distant insect species. Scientific reports.
996 2016;6:27885.

997 47. Huang DH, Chang YL, Yang CC, Pan IC, King B. pipsqueak encodes a factor essential for
998 sequence-specific targeting of a polycomb group protein complex. Mol Cell Biol.
999 2002;22(17):6261-71.

1000 48. Halfon MS. Silencers, Enhancers, and the Multifunctional Regulatory Genome. Trends
1001 Genet. 2020;36(3):149-51.

1002 49. Segert JA, Gisselbrecht SS, Bulyk ML. Transcriptional Silencers: Driving Gene Expression
1003 with the Brakes On. Trends Genet. 2021;37(6):514-27.

1004 50. Laiker I, Frankel N. Pleiotropic Enhancers are Ubiquitous Regulatory Elements in the
1005 Human Genome. Genome biology and evolution. 2022;14(6).

1006 51. Sabaris G, Laiker I, Preger-Ben Noon E, Frankel N. Actors with Multiple Roles: Pleiotropic
1007 Enhancers and the Paradigm of Enhancer Modularity. Trends Genet. 2019;35(6):423-33.

1008 52. Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. Phenotypic robustness
1009 conferred by apparently redundant transcriptional enhancers. Nature. 2010;466(7305):490-3.

1010 53. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow enhancers foster robustness of
1011 Drosophila gastrulation. Curr Biol. 2010;20(17):1562-7.

1012 54. Perry MW, Boettiger AN, Levine M. Multiple enhancers ensure precision of gap gene-
1013 expression patterns in the Drosophila embryo. Proc Natl Acad Sci U S A.
1014 2011;108(33):13570-5.

1015 55. Osterwalder M, Barozzi I, Tissieres V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al.
1016 Enhancer redundancy provides phenotypic robustness in mammalian development. Nature.
1017 2018;554(7691):239-43.

1018 56. Antosova B, Smolikova J, Klimova L, Lachova J, Bendova M, Kozmikova I, et al. The Gene
1019 Regulatory Network of Lens Induction Is Wired through Meis-Dependent Shadow Enhancers
1020 of Pax6. PLoS Genet. 2016;12(12):e1006441.

1021 57. Sagai T, Amano T, Maeno A, Kiyonari H, Seo H, Cho SW, et al. SHH signaling directed by
1022 two oral epithelium-specific enhancers controls tooth and oral development. Scientific
1023 reports. 2017;7(1):13004.

1024 58. Waymack R, Fletcher A, Enciso G, Wunderlich Z. Shadow enhancers can suppress input
1025 transcription factor noise through distinct regulatory logic. eLife. 2020;9.

1026 59. Cannavo E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, et al. Shadow
1027 Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr Biol.
1028 2016;26(1):38-51.

1029 60. Barth NKH, Li L, Taher L. Independent Transposon Exaptation Is a Widespread Mechanism
1030 of Redundant Enhancer Evolution in the Mammalian Genome. Genome biology and
1031 evolution. 2020;12(3):1-17.

1032 61. Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative
1033 analysis inflates the apparent number of lineage-specific genes. Curr Biol. 2022;32(12):2632-
1034 9 e2.

1035 62. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene
1036 promoters. Nature. 2012;489(7414):109-13.

1037 63. Hafez D, Karabacak A, Krueger S, Hwang YC, Wang LS, Zinzen RP, et al. McEnhancer:
1038 predicting gene expression via semi-supervised assignment of enhancers to target genes.
1039 Genome Biol. 2017;18(1):199.

1040 64. Chua EHZ, Yasar S, Harmston N. The importance of considering regulatory domains in
1041 genome-wide analyses - the nearest gene is often wrong! Biol Open. 2022;11(4).

1042 65. Crosby MA, Gramates LS, Dos Santos G, Matthews BB, St Pierre SE, Zhou P, et al. Gene
1043 Model Annotations for Drosophila melanogaster: The Rule-Benders. G3. 2015;5(8):1737-49.

1044 66. Matthews BB, Dos Santos G, Crosby MA, Emmert DB, St Pierre SE, Gramates LS, et al.
1045 Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data.
1046 G3. 2015;5(8):1721-36.

1047 67. Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, Dos Santos G, et al. FlyBase: a
1048 guided tour of highlighted features. Genetics. 2022;220(4).

1049 68. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
1050 Bioinformatics. 2010;26(6):841-2.

1051 69. Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, et al. The
1052 transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by
1053 displacing nucleosomes. Nat Genet. 2018;50(7):1011-20.

1054 70. Donitz J, Gerischer L, Hahnke S, Pfeiffer S, Bucher G. Expanded and updated data and a
1055 query pipeline for iBeetle-Base. Nucleic Acids Res. 2018;46(D1):D831-D5.

1056 71. Ruiz JL, Ranford-Cartwright LC, Gomez-Diaz E. The regulatory genome of the malaria
1057 vector Anopheles gambiae: integrating chromatin accessibility and gene expression. NAR
1058 Genom Bioinform. 2021;3(1):lqaa113.

1059 72. Katzen F. Gateway((R)) recombinational cloning: a biological operating system. Expert Opin
1060 Drug Discov. 2007;2(4):571-89.

1061 73. Asma H, Halfon MS. Annotating the Insect Regulatory Genome. Insects. 2021;12(7):591.

1062

1063

1064 **SUPPORTING INFORMATION:**

1065 **S1 Text:** Legends for Supplemental Figures

1066 **S1 Figure:** Lengths of predicted enhancers, including long outliers

1067 **S2 Figure:** Open chromatin data for predictions in the *ex, klu,* and *ush* loci

1068 **S3 Figure:** Open chromatin data for predictions in the *hth, Ubx,* and *psq* loci

1069 **S4 Figure:** Additional expression observed in selected transgenic reporter lines

1070 **S5 Figure:** Reporter gene expression observed in embryos

1071 **S6 Figure:** Revised post-processing method used for SCRMshaw

1072 **S1 Table:** SCRMshaw training sets used in this study

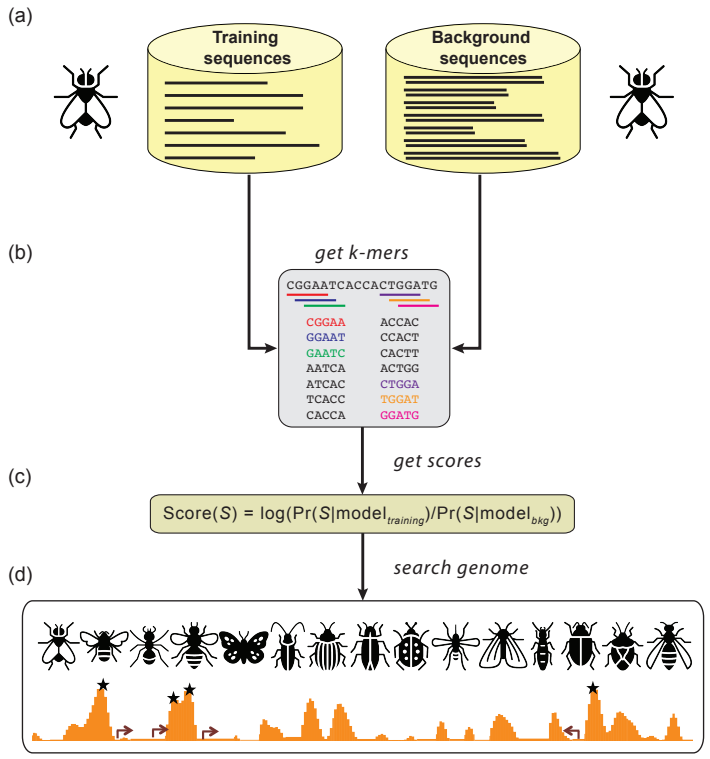
1073 **S2 Table:** Number of predicted enhancers for each species, by method

1074 **S3 Table:** Real and simulated predictions per locus

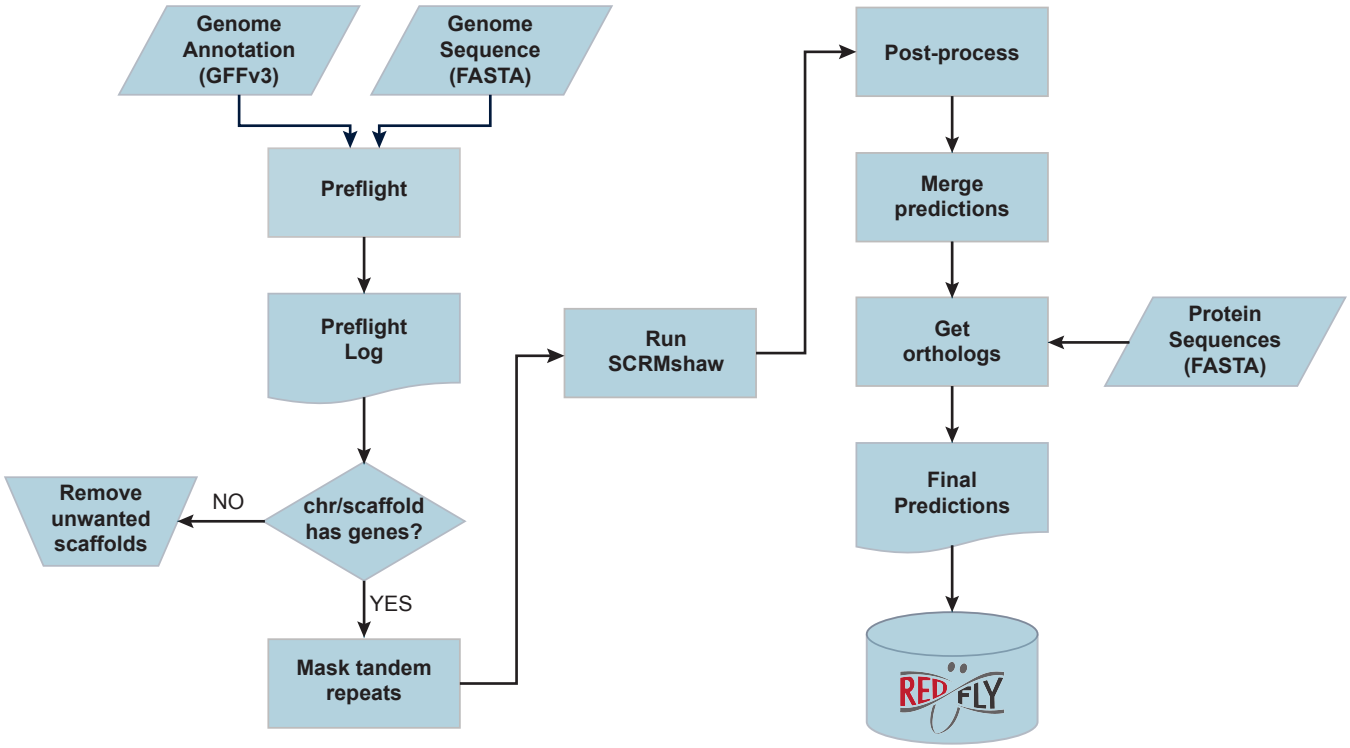
1075 **S4 Table:** Real and simulated counts of predictions in orthologous loci

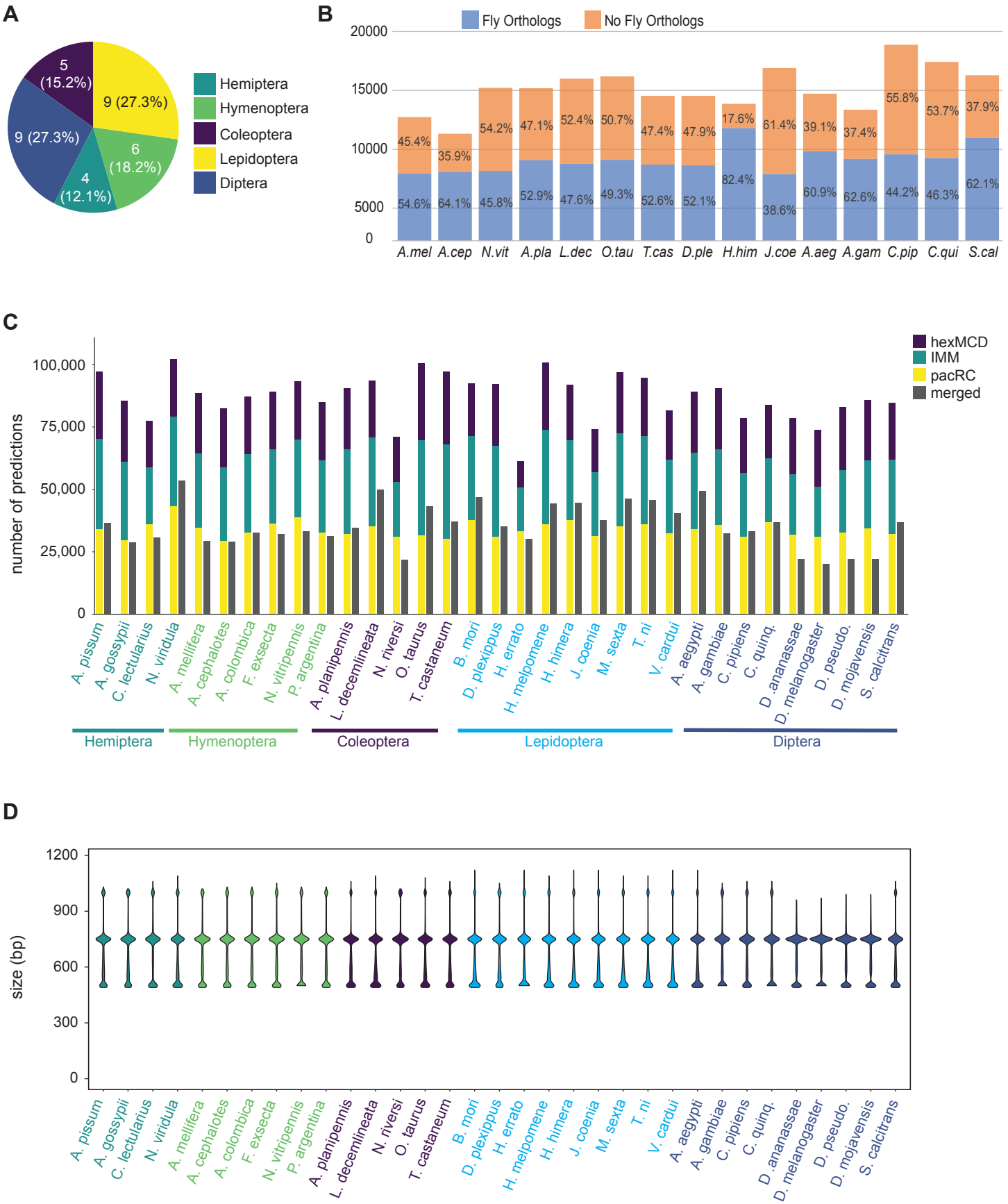
1076 **S5 Table:** Primers and sequences used for *in vivo* validation

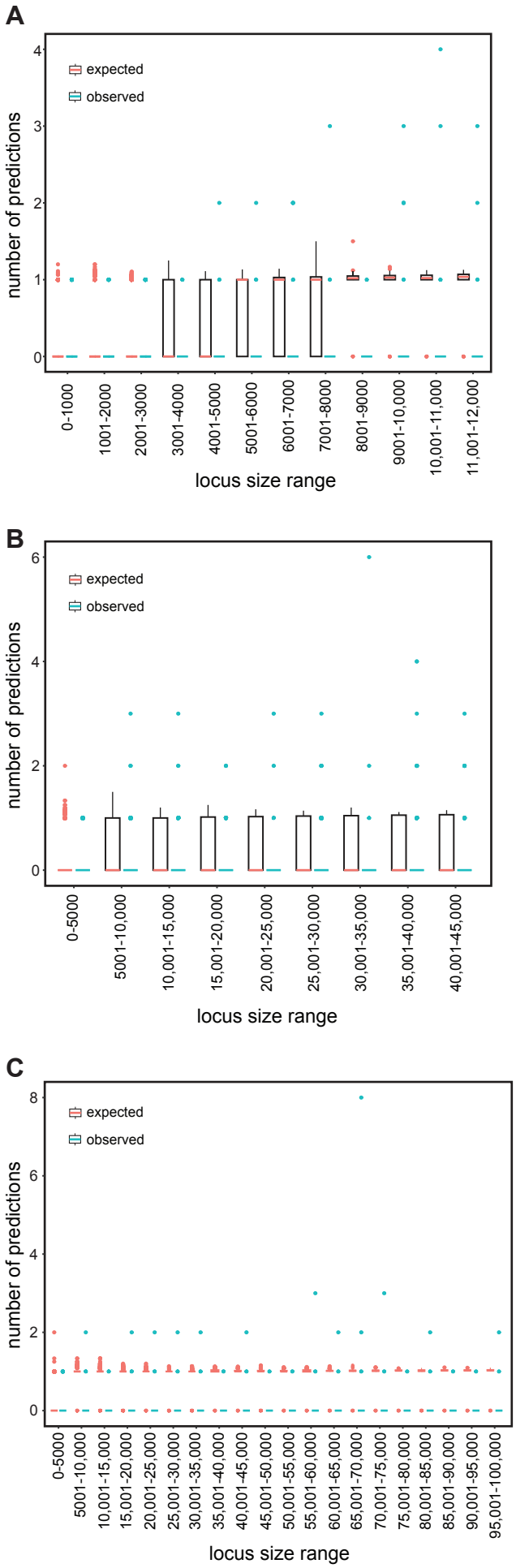
A


(a)
Training sequences
Background sequences
(b)
get k-mers
CGGAATCACCACTGGATG
CGGAA
GGAAT
GAATC
AATCA
ATCAC
TCACC
CACCA
ACCAC
CCACT
CACTT
ACTGG
CTGGA
TGGAT
GGATG
(c)
get scores
Score(S) = log(Pr(S|model_training)/Pr(S|model_bkg))
(d)
search genome

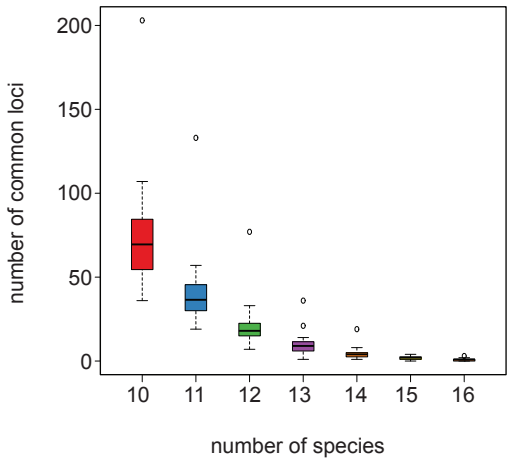
B


Genome Annotation (GFFv3)
Genome Sequence (FASTA)
Preflight
Preflight Log
chr/scaffold has genes?
NO
Remove unwanted scaffolds
YES
Mask tandem repeats
Run SCRMshaw
Post-process
Merge predictions
Get orthologs
Protein Sequences (FASTA)
Final Predictions
RED FLY

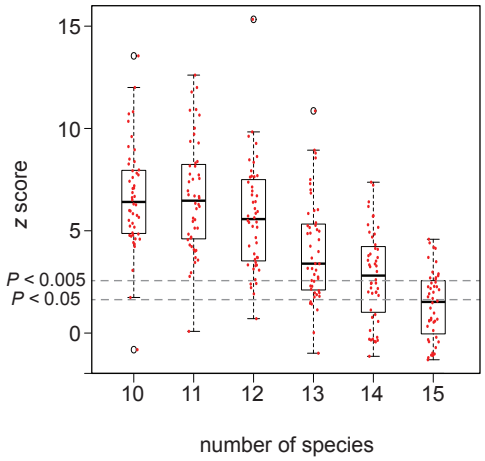

A
5 (15.2%)
9 (27.3%)
9 (27.3%)
6 (18.2%)
4 (12.1%)
Hemiptera
Hymenoptera
Coleoptera
Lepidoptera
Diptera
B
Fly Orthologs
No Fly Orthologs
20000
15000
10000
5000
0
45.4%
35.9%
54.2%
47.1%
52.4%
50.7%
47.4%
47.9%
17.6%
61.4%
39.1%
37.4%
55.8%
53.7%
37.9%
54.6%
64.1%
45.8%
52.9%
47.6%
49.3%
52.6%
52.1%
82.4%
38.6%
60.9%
62.6%
44.2%
46.3%
62.1%
A.mel
A.cep
N.vit
A.pla
L.dec
O.tau
T.cas
D.ple
H.him
J.coe
A.aeg
A.gam
C.pip
C.qui
S.cal
C
hexMCD
IMM
pacRC
merged
number of predictions
100,000
75,000
50,000
25,000
0
A. pissum
A. gossypii
C. lectularius
N. viridula
A. mellifera
A. cephalotes
A. colombica
F. exsecta
N. vitripennis
P. argentina
A. planipennis
L. decemlineata
N. riversi
O. taurus
T. castaneum
B. mori
D. plexippus
H. errato
H. melpomene
H. himera
J. coenia
M. sexta
T. ni
V. cardui
A. aegypti
A. gambiae
C. pipiens
C. quinq.
D .ananassae
D. melanogaster
D. pseudo.
D. mojavensis
S. calcitrans
Hemiptera
Hymenoptera
Coleoptera
Lepidoptera
Diptera
D
size (bp)
1200
900
600
300
0
A. pissum
A. gossypii
C. lectularius
N. viridula
A. mellifera
A. cephalotes
A. colombica
F. exsecta
N. vitripennis
P. argentina
A. planipennis
L. decemlineata
N. riversi
O. taurus
T. castaneum
B. mori
D. plexippus
H. errato
H. melpomene
H. himera
J. coenia
M. sexta
T. ni
V. cardui
A. aegypti
A. gambiae
C. pipiens
C. quinq.
D .ananassae
D. melanogaster
D. pseudo.
D. mojavensis
S. calcitrans


A
expected
observed
number of predictions
0
1
2
3
4
0-1000
1001-2000
2001-3000
3001-4000
4001-5000
5001-6000
6001-7000
7001-8000
8001-9000
9001-10,000
10,001-11,000
11,001-12,000
locus size range
B
expected
observed
number of predictions
0
2
4
6
0-5000
5001-10,000
10,001-15,000
15,001-20,000
20,001-25,000
25,001-30,000
30,001-35,000
35,001-40,000
40,001-45,000
locus size range
C
expected
observed
number of predictions
0
2
4
6
8
0-5000
5001-10,000
10,001-15,000
15,001-20,000
20,001-25,000
25,001-30,000
30,001-35,000
35,001-40,000
40,001-45,000
45,001-50,000
50,001-55,000
55,001-60,000
60,001-65,000
65,001-70,000
70,001-75,000
75,001-80,000
80,001-85,000
85,001-90,000
90,001-95,000
95,001-100,000
locus size range

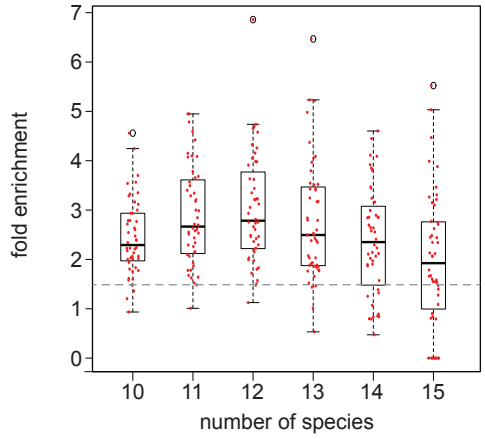
A

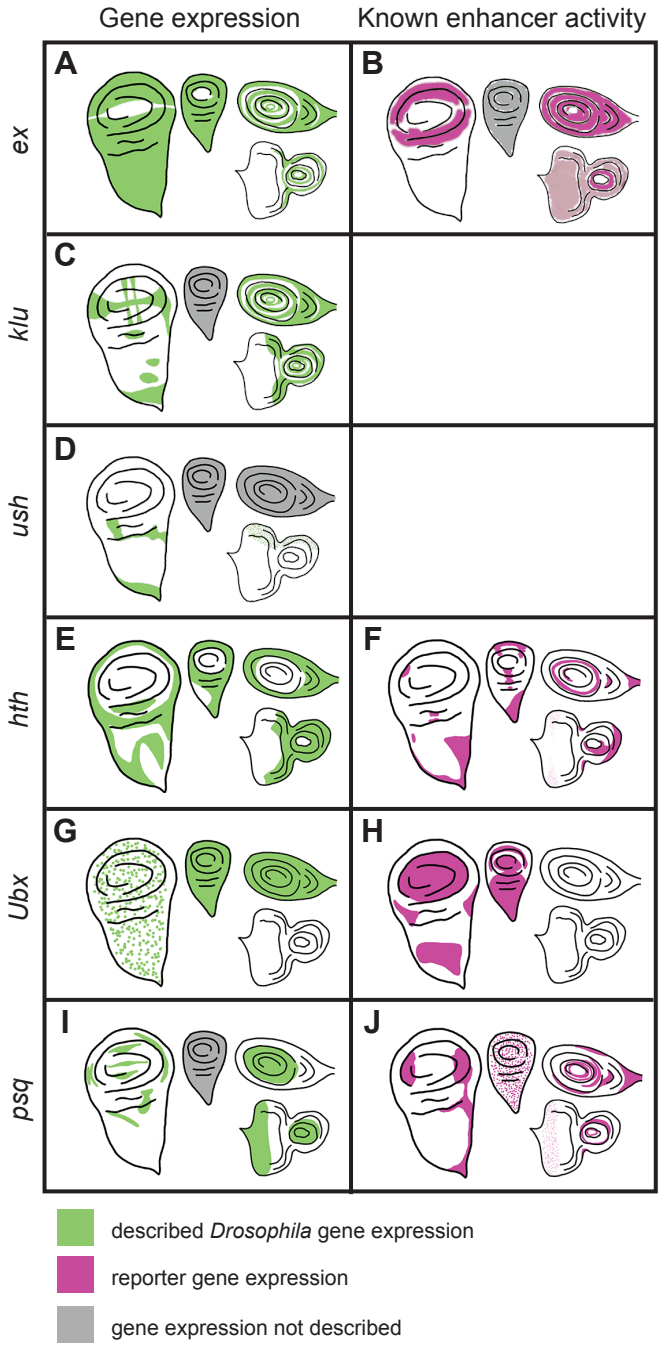

number of common loci
200
150
100
50
0
10
11
12
13
14
15
16
number of species

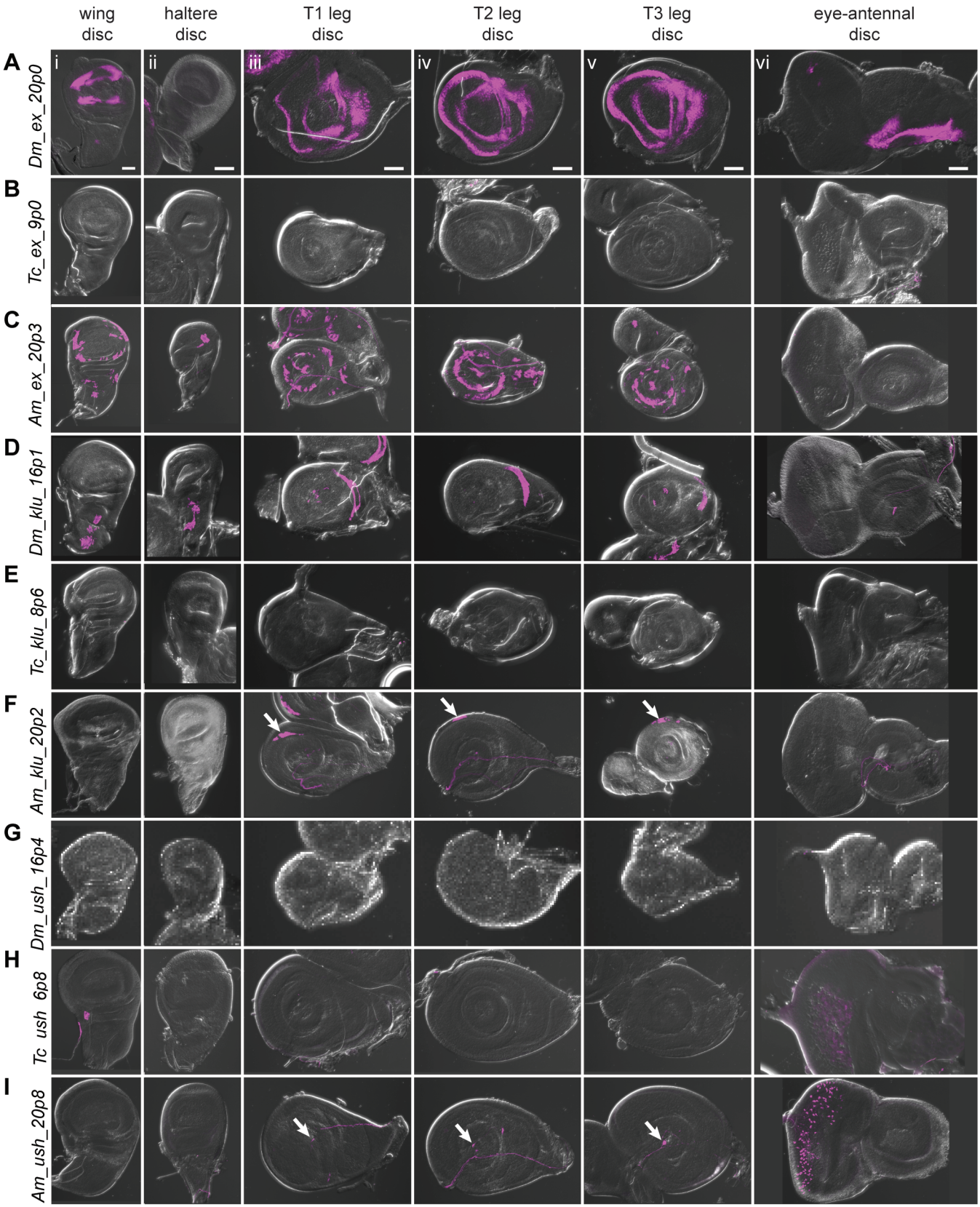
B

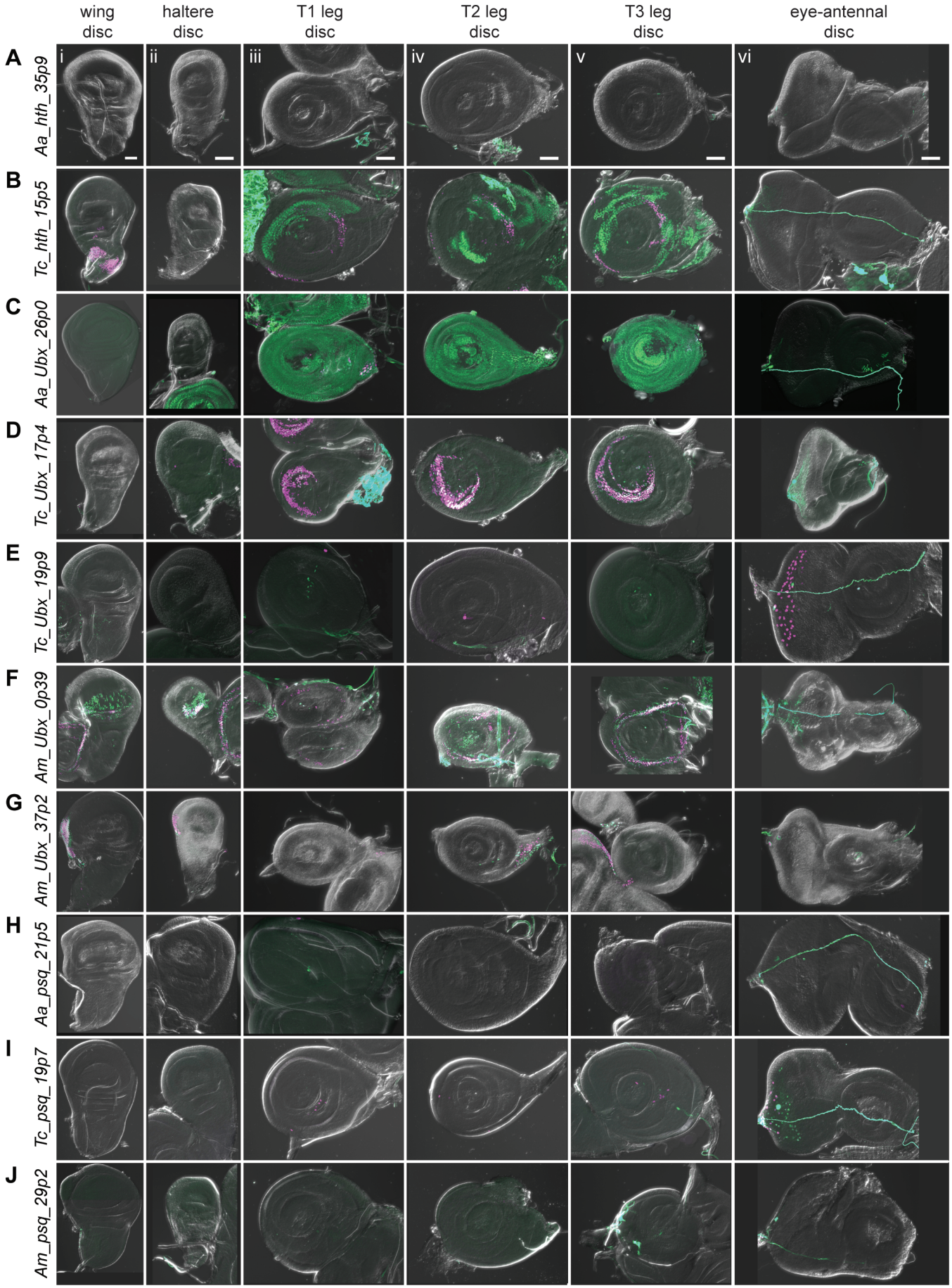

z score
15
10
5
0
P < 0.005
P < 0.05
10
11
12
13
14
15
number of species

C


fold enrichment
7
6
5
4
3
2
1
0
10
11
12
13
14
15
number of species


Gene expression
Known enhancer activity
ex
klu
ush
hth
Ubx
psq
A
B
C
D
E
F
G
H
I
J
described Drosophila gene expression
reporter gene expression
gene expression not described


wing disc
haltere disc
T1 leg disc
T2 leg disc
T3 leg disc
eye-antennal disc
A
Dm_ex_20p0
i
ii
iii
iv
v
vi
B
Tc_ex_9p0
C
Am_ex_20p3
D
Dm_klu_16p1
E
Tc_klu_8p6
F
Am_klu_20p2
G
Dm_ush_16p4
H
Tc_ush_6p8
I
Am_ush_20p8


wing disc
haltere disc
T1 leg disc
T2 leg disc
T3 leg disc
eye-antennal disc
A
Aa_hth_35p9
i
ii
iii
iv
v
vi
B
Tc_hth_15p5
C
Aa_Ubx_26p0
D
Tc_Ubx_17p4
E
Tc_Ubx_19p9
F
Am_Ubx_0p39
G
Am_Ubx_37p2
H
Aa_psq_21p5
I
Tc_psq_19p7
J
Am_psq_29p2