Chromosomes and Gene Expression
Genetics and Genomics

Regulatory genome annotation of 33 insect species

Reviewed Preprint
v2 • September 18, 2024
Revised by authors

Reviewed Preprint
v1 • June 7, 2024

Hasiba Asma, Ellen Tieke, Kevin D Deem, Jabale Rahmat, Tiffany Dong, Xinbo Huang, Yoshinori Tomoyasu, Marc S Halfon

Program in Genetics, Genomics, and Bioinformatics, University at Buffalo-State University of New York, Buffalo NY 14203 • Department of Biology, Miami University, Oxford OH 45056 • Department of Biochemistry, University at Buffalo-State University of New York, Buffalo NY 14203 • Department of Biomedical Informatics, University at Buffalo-State University of New York, Buffalo NY 14203 • Department of Biological Sciences, University at Buffalo-State University of New York, Buffalo NY 14260

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.

eLife assessment

In the revised version of this **important** study, the authors present a **convincing** pipeline for insect genome regulatory annotation across 33 insect genomes spanning 5 orders. Despite technical limitations in the field owing to the lack of comprehensive knowledge of enhancer content in any system, the authors employ several independent downstream analyses to support the validity of their enhancer predictions for a subset of these genomes. Taken together, the revised results suggest that this prediction pipeline may have uses in identifying functional enhancers across large phylogenetic distances. Reviewers note caveats that an experimental validation is not yet available in the field to validate a large class of newly identified enhancers across such evolutionary distances, and other pipelines might be of use to compare. This work will be of interest to the computational genomics, evolutionary biology, and gene regulation fields.

https://doi.org/10.7554/eLife.96738.2.sa3

Introduction

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

One reason for the overall dearth of regulatory annotations is that, historically, large-scale CRM discovery has been difficult and both resource and labor intensive (7). For much of the last four decades, CRMs could only be identified through painstaking, low-throughput experimental assays. Although in recent years high-throughput empirical and computational CRM discovery methods have been developed, the various different methods frequently show limited agreement (8–10), with the result that comprehensive CRM annotation across all cell types and life-cycle stages has remained a challenge for all but the most exhaustively studied model organisms. The problem is particularly acute for the insects. Insects represent a species-rich class—they constitute somewhere between 65%-90% of all animal species (11, 12)—and have major impacts on human health and agriculture. The early radiation of the insects, coupled with typically short generation times, means that most of the relevant biomedically and agriculturally important species share little non-coding sequence conservation with each other or with the principal insect model species, *Drosophila melanogaster*. Thus, common sequence-homology based CRM discovery approaches are of little use in providing regulatory insights into these species, and knowledge transfers poorly from one species to another. Furthermore, many insects have a complex and varied life cycle, making it particularly important—yet onerous—to assay for CRM function at multiple stages.

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

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

Results

We previously demonstrated that SCRMshaw is remarkably effective at predicting enhancers across the entire ~345 MY range of the holometabolous insects, using training data derived solely from *D. melanogaster* (17, 19). SCRMshaw (**Fig. 1A**) uses training sets composed of known enhancers defined by a common functional characterization (e.g. "nervous system," "wing disc") to build a statistical model that captures their short DNA subsequence (*kmer*) count distribution. This *kmer* distribution is then compared to that of a set of non-enhancer "background" sequences in a machine-learning framework. The *kmers* likely serve as proxies for the unknown transcription factor binding sites, but these sites themselves, even when known, are not explicitly used by the algorithm. The trained model is then used to score overlapping sequence windows in the genome, and the highest-scoring windows are output as predicted enhancers (13, 14, 21). When searching the genomes of the mosquitoes *Anopheles gambiae* and *Aedes aegypti*, the red flour beetle *Tribolium castaneum*, the honey bee *Apis mellifera*, and the wasp *Nasonia vitripennis*, SCRMshaw successfully predicted enhancers in a cross-species fashion with an approximately 75% prediction success rate, based on reporter gene assays in xenotransgenic *D. melanogaster* (chosen as a pragmatic transgene host species) and comparison to already-identified enhancers in the other species (17–20). These results suggest that there are significant, albeit hidden, homologies governing the sequence characteristics of insect enhancers, at least for those involved in a substantial number of gene regulatory networks, and motivated us to apply SCRMshaw to a large and diverse set of sequenced insect genomes.

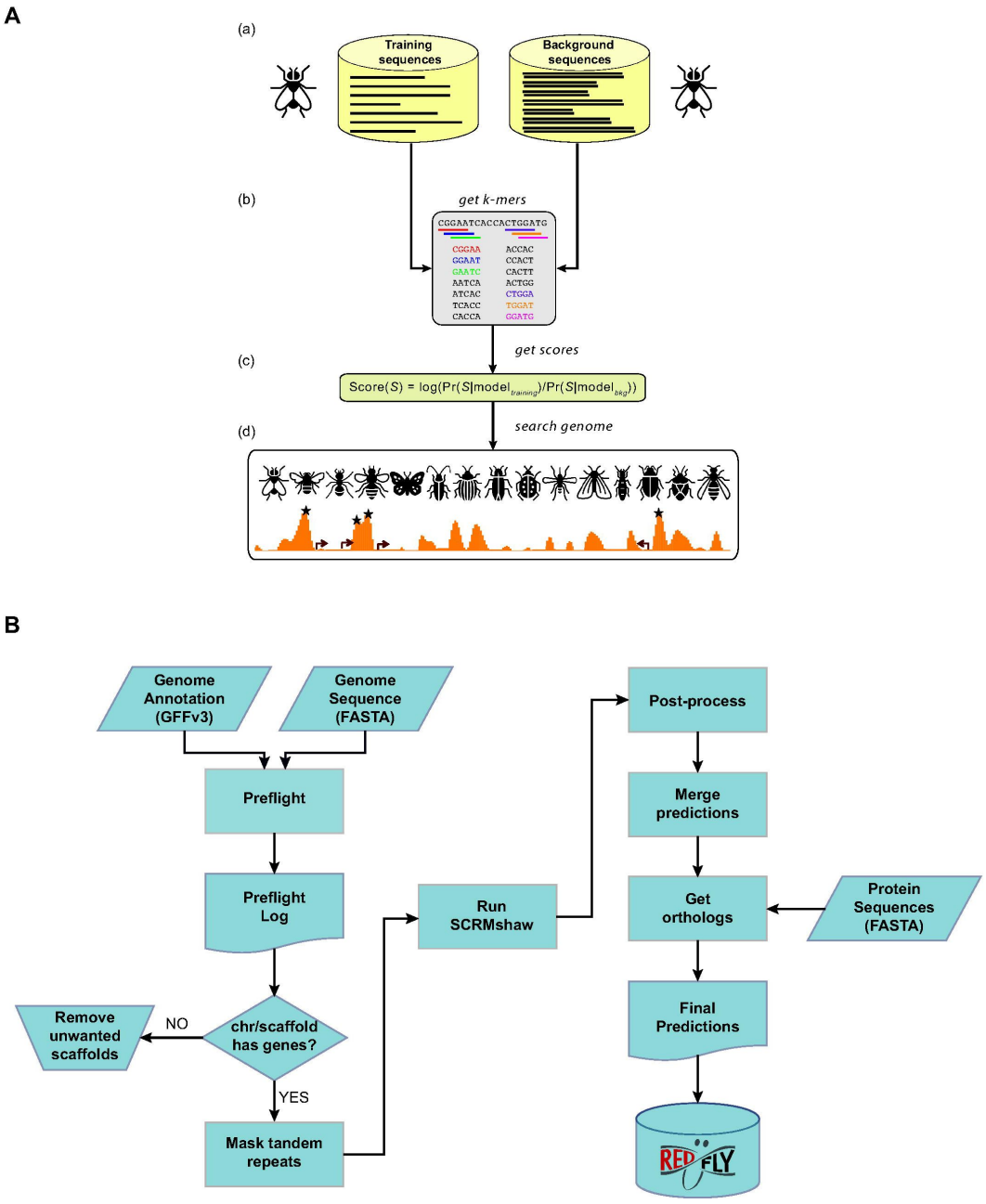


Figure 1

The SCRMshaw method and analysis pipeline.

(A) Supervised motif-blind CRM discovery (SCRMshaw). (a) SCRMshaw uses a training set of known *D. melanogaster* enhancers ("training sequences"), drawn from REDfly, that are defined by common functional characterization, and a 10 fold larger background set of similarly sized common functional characterization, non enhancer sequences ("background sequences"). (b) The short DNA subsequence (*kmer*) count distributions of these sequences are then used to train a statistical model. Note that although the pictured example shows 5-mers, *kmers* of different sizes are used for some of the underlying statistical models (see Methods). The trained model (c) is used to score overlapping windows in the "target genome." (d) High-scoring regions are predicted to be functional regulatory sequences (asterisks). Figure adapted from (81). (B) The workflow used for the regulatory genome annotation described in this paper. The left side shows pre-processing steps, the right side, post-processing. Input to SCRMshaw consists of the genome sequence and gene annotation. A protein sequence annotation is supplied later for the orthology mapping step. Final results are made available as part of the REDfly regulatory annotation knowledgebase.

A cross-species SCRMshaw pipeline

To facilitate application of SCRMshaw to large numbers of newly-sequenced genomes, we developed a detailed workflow to ensure proper formatting of input genomes, rapid prediction of tissue-specific enhancer sequences, evaluation of results, and annotation of loci with information drawn from the respective orthologous regions in the richly-investigated *D. melanogaster* genome (**Fig. 1B**; see Methods). The workflow consists of four major steps: *Preflight*: SCRMshaw requires two input files for each genome: a FASTA-formatted file of the genome sequence itself, and a GFFv3-formatted file of the genome annotation. The genome file is masked for tandem repeats using Tandem Repeat Finder (22). *Preflight* validates the formats of these files and produces a comprehensive log file that highlights any issues along with basic information such as the number of chromosomes/scaffolds and their sizes, data types present in the annotation (e.g., 'gene', 'exon', 'ncRNA', etc.), and average intergenic distances. *Preflight* also provides a sample output of the SCRMshaw-generated 'gene' and 'exon' files. This feature allows users to identify any discrepancies or errors stemming from the input files and to reformat these files as needed before running SCRMshaw. Any minor scaffolds that are annotated as not containing genes are discarded.

SCRMshaw

SCRMshaw is run as previously described (15), using the "HD" variant (21). SCRMshaw_HD scans a genome with a 500 bp sliding window offset in 10 bp increments, using a set of three statistical models that compare the *kmer* composition of a set of training enhancers from *Drosophila melanogaster* to randomly selected *D. melanogaster* non-coding sequences (see Methods).

Post-processing

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

Orthology mapping

Each SCRMshaw prediction is assigned its closest flanking genes as putative enhancer target genes, to aid in interpretation of the SCRMshaw output. There are clear shortcomings to this approach, as many enhancer targets are not the closest genes, leading to mis-assignments (e.g. 24, 25-27). However, potentially more accurate methods for target-gene assignment rely on gene expression data, epigenetic data, and/or chromatin conformation data that are frequently not available for the species we are studying and thus difficult to incorporate into our prediction pipeline (27–30). Putative target genes are then mapped to their *D. melanogaster* orthologs (if an ortholog exists) using the *Orthologer* software from the Zdobnov lab (31) (see Methods). We use *D. melanogaster* as it has by far the most comprehensive gene annotation of the insects and thus provides the most detailed functional information for each gene. Mapping the genes from each species to a common ortholog helps us to assess whether we have obtained predicted enhancers in orthologous loci within the various species on which we have run SCRMshaw. The orthology mapping step requires that a set of predicted proteins is present as part of the existing genome annotation. Note that neither this step nor the earlier target-gene assignment step affects the enhancer predictions themselves, which are generated prior to target-gene assignment and orthology mapping and are considered equally valid regardless of putative target genes and whether or not orthologs can be matched to them.

Annotation of 33 insect genomes

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

Collectively, we predicted a total of 2,873,192 enhancers in these 33 species, with each species having on average approximately 87,000 predictions (Supplemental Table S2) (**Fig. 2C**). (As some enhancers are predicted by multiple scoring methods, or from more than one training set, the number of unique sequences is lower at 1,164,354; see gray bars in **Fig. 2C** and Supplemental Table S2.) The median length of the predicted enhancers across all species was 750 bp (mean 695, range 490-32500 bp). However, we noted the presence of a small number— 2642, < 0.1% of the total—of unusually large regions (> 2000 bp), the bulk of which were confined to just a few genomes (Supplemental Fig. S1). We therefore discarded any predictions with length greater than 1.5 times the interquartile range of the complete prediction set and re-evaluated the size distribution. This resulted in a median size of 740 bp with a mean of 676 bp and a range of 490-1120 bp, indicating that the overall impact of these outlier sequences is minimal (**Fig. 2D**). Inspection of the excessively large elements revealed that they result from SCRMshaw predictions that lie immediately adjacent to each other (without gaps) and have similar SCRMshaw scores, which thus become merged into one broad predicted element. Preliminary analysis suggests that these regions result from insufficient masking of tandem repeat regions and/or genome assembly errors, although other causes, such as extremely enhancer-dense "superenhancer" regions (reviewed by 32), cannot entirely be ruled out.

Many loci contain multiple enhancer predictions

We have noted in the past that SCRMshaw often predicts multiple enhancers in a single locus (e.g. 33). This is consistent with the concept of "shadow enhancers," sets of multiple redundant or semi-redundant enhancers regulating the same gene (reviewed by 34). On the other hand, if SCRMshaw is predicting enhancers with low specificity (i.e., largely at random), it is instead possible that larger loci may just accumulate a high number of predictions simply due to their greater length.

To distinguish between these possibilities, we conducted simulations on three representative genomes (*D. melanogaster*, *C. pipiens,* and *A. aegypti*), each with a different average intergenic region size chosen to represent small, medium and large genomes respectively (average sizes 610 bp, 5,274.5 bp, and 14,641 bp; see Methods). We randomized the location of SCRMshaw predictions across the non-coding component of each genome and compared the randomized results to our SCRMshaw output. We found that, for a subset of loci, the number of real SCRMshaw predictions per locus was consistently higher than the number obtained at random (Supplemental Table S3). For example, using the *mapping1.visceral_mesoderm* training set, 4.6% (15/328) of *D. melanogaster* loci containing one or more SCRMshaw predictions had a significantly (P<0.001) larger number of predictions/locus than expected, 1.5% (10/658) of *C. pipiens* loci had more predictions than expected, and 1% (3/299) of *A. aegypti* loci had excess predicted enhancers (Supplemental Table S3). Similarly, for the *mapping2.ectoderm* training set, 7.6%, 1.5%, and 11% of loci in the three species, respectively, had a significantly (P<0.001) greater than expected number of predictions per locus (Supplemental Table S3). Averaged over all training sets, 3-4% of loci with predictions had significantly more than expected by chance (mean values: 3.0% for *D. melanogaster*, 3.0% for *C. pipiens*, and 3.7% for *A. aegypti*; maximum values: 10.2% for *D. melanogaster*, 9.3% for *C. pipiens*,

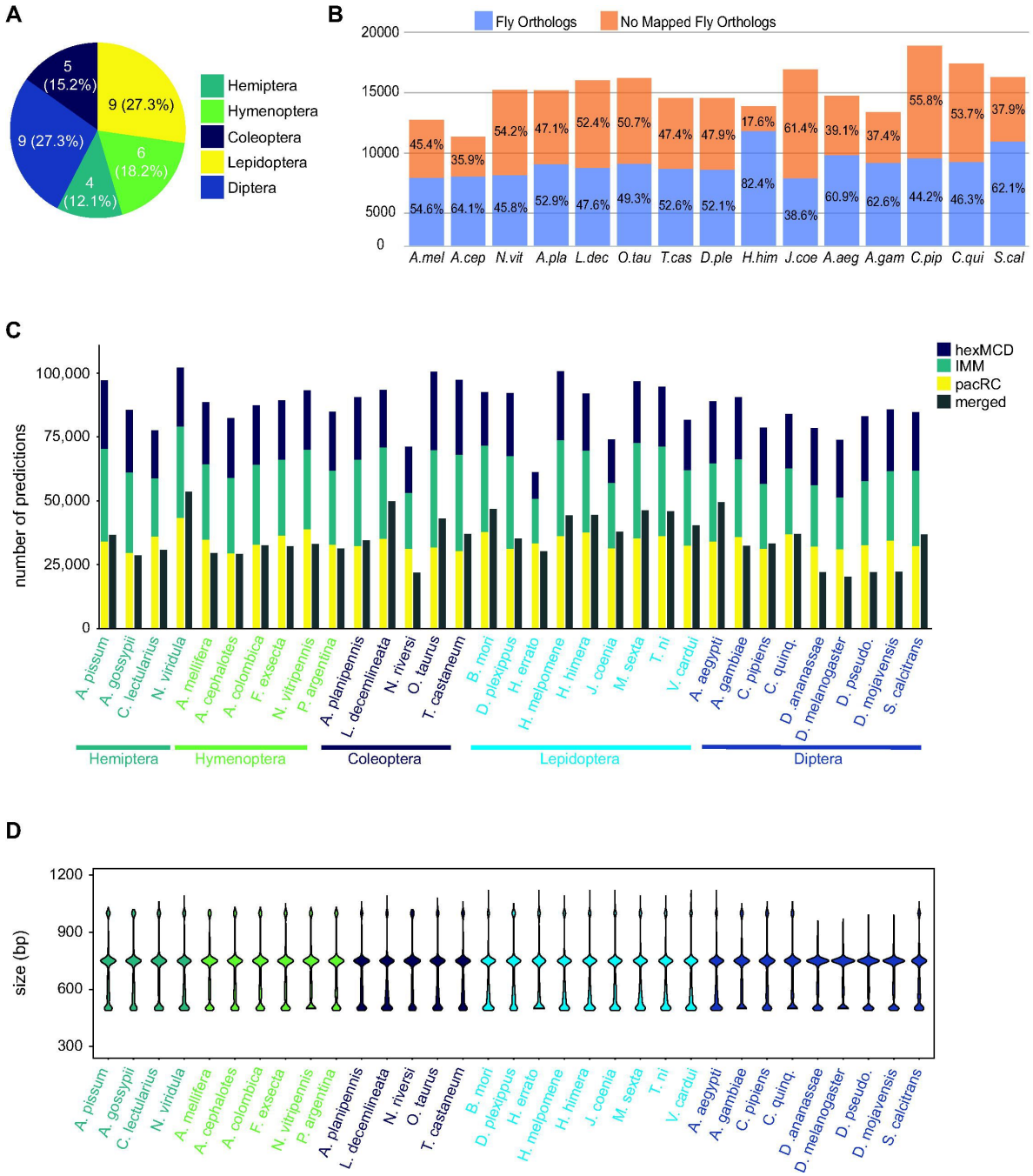


Figure 2

Annotation of 33 insect genomes.

(A) Genomes from five insect orders were annotated in this study (more are ongoing). (B) Percentage of genes with *Drosophila* orthologs as mapped via our orthology pipeline (see Methods), for the 15 mapped species. “No Mapped Fly Orthologs” indicates that our orthology mapping pipeline did not identify clear *D. melanogaster* orthologs. For any given gene, this could reflect either a true lack of a respective ortholog, or failure of our procedure to accurately identify an existing ortholog. For complete species names, see **Table 1**. (C) Total SCRMshaw predictions for each species. For each species, the lefthand column shows cumulative results for each SCRMshaw sub-method summed over each of the 48 training sets. The righthand column shows the number of unique predictions after merging overlapping predictions from both sub-methods and training sets. Species are displayed alphabetically by taxonomic order. (See also Supplemental Table S2.) (D) Size distribution of SCRMshaw predictions, prior to merging overlapping predictions but after removing outlier predictions > 2 kb in length. Species are ordered identically to panel C.

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	ASM365146	https://www.ncbi.nlm.nih.gov/datasets/gen

Table 1

Species used in this study

			v1	ome/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

Table 1 (continued)

and 11.6% for *A. aegypti*). Only very few training sets did not have any loci with significantly more enhancers than expected (1 set in *A. aegypti,* 1 in *C. pipiens* and 8 in *D. melanogaster*; Supplemental Table S3).

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

SCRMshaw predicts enhancers in orthologous loci across species

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

We observed a substantial reduction in the number of orthologous loci with predictions in common as we moved from considering a set of ten to the full set of all sixteen species (average number of common loci over all 48 training sets: 70.7 (10 species) > 39.4. > 19.8 > 9.14 > 4.12 > 1.6 > 0.6 (16 species) out of an average number of 640 predictions with mapped orthologs) (**Fig. 4A**, Supplemental Table S4a). This rapid decline is likely due to a variety of factors, including differences in taxonomic order (e.g., Diptera vs. Hymenoptera), the quality and degree of ortholog data for each species, and the quality of annotation for each species, all of which likely lead to an underestimation of the true numbers of common loci in our data (see Discussion). We simulated SCRMshaw predictions for all sixteen species by randomizing the SCRMshaw results and compared the number of common loci between the simulated and real data for combinations of ten to sixteen species. The number of common loci among the real SCRMshaw predictions was consistently significantly higher than the number of common loci observed in the simulated data ($P< 0.05$; **Fig. 4B** and Supplemental Table S4). In particular, when evaluating between 10 and 12 species (data for 14-16 species are not reliable due to the very small number of observed loci in common) only a single training set, *adult_PNS*, did not have a significant overrepresentation of common loci (Supplemental Table S4c). We also examined the fold enrichment, i.e., the extent to which the true number was in excess of the simulated number, and found that on average there were greater than 2.4x more predictions in common loci than expected, when considering groups of 10-12 species; almost all training sets had a fold enrichment >1.5x (**Fig. 4C**, Supplemental Table S4d). These results strongly suggest that SCRMshaw is successfully finding sets of conserved (or convergent) enhancers, as it is predicting sequences in orthologous loci in a training-set specific manner across multiple species significantly more often than can be accounted for by chance.

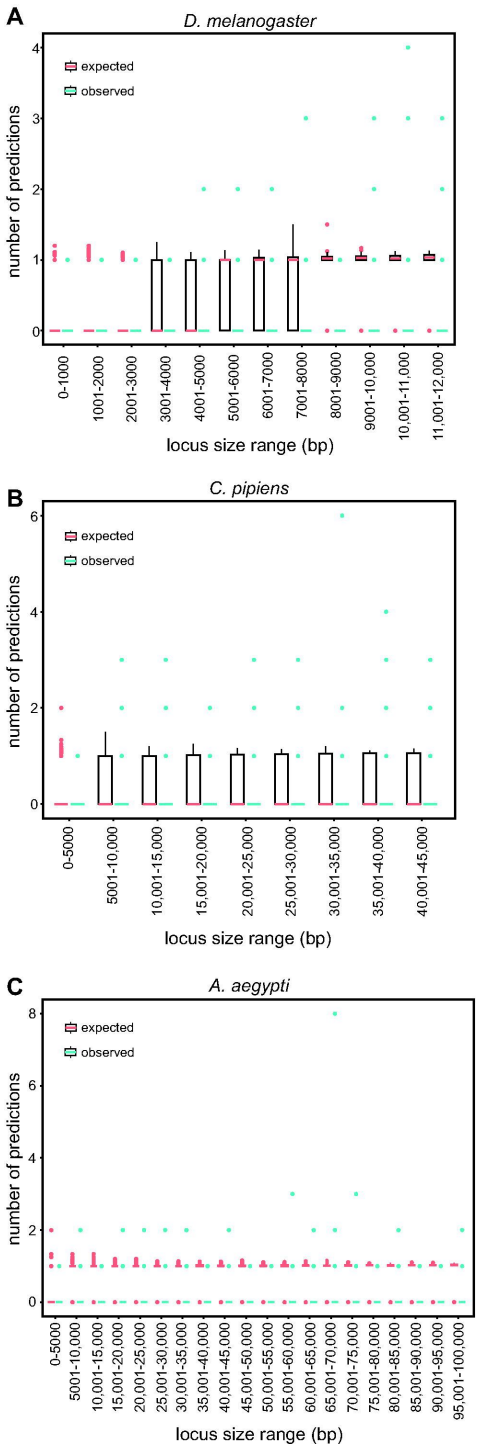


Figure 3

SCRMshaw makes multiple predictions per locus.

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

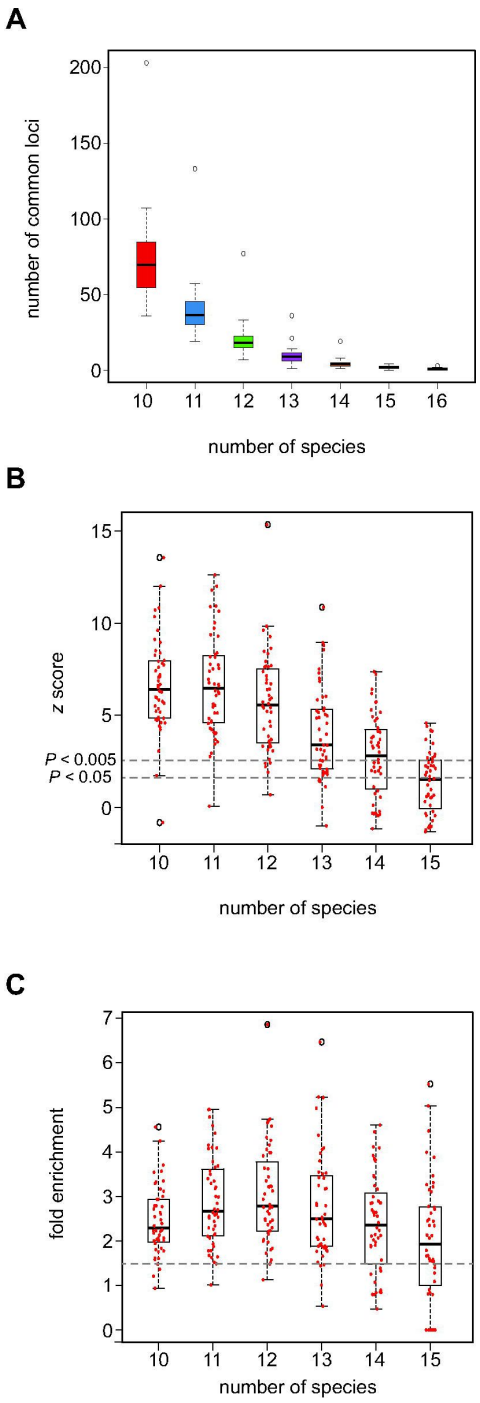


Figure 4

SCRMshaw predicts CRMs in orthologous loci across species.

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

SCRMshaw predictions correlate with regions of accessible chromatin

Active enhancers are frequently found in regions of accessible chromatin, as assayed by methods such as DNAse-seq, FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements with sequencing), and ATAC-seq (Assay for Transposase-Accessible Chromatin by sequencing)([35–37]). We confirmed previously that FAIRE-predicted and SCRMshaw-predicted enhancers in *T. castaneum* have a high (>79%) degree of overlap ([18]). To determine whether a similar correlation exists for other species, we compared our SCRMshaw predictions with available FAIRE-seq and ATAC-seq data for seven species: *D. melanogaster*, *T. castaneum*, *A. gambiae*, *Danaus plexippus, Vanessa cardui, Junonia coenia,* and *Heliconius himera*. These comparisons are imperfect, as the tissues used to obtain the chromatin data do not precisely correspond to the training sequences used for SCRMshaw, and the data were obtained using a variety of methods. Nevertheless, in the majority of cases where we were able to establish a rough match between the tissues, we observed significant overlap between the two methods of enhancer detection. For example, in *D. melanogaster*, 40% of SCRMshaw predictions using the *blastoderm.mapping1* training set overlapped ATAC-seq peaks obtained from blastoderm embryos (*P*<4.15e-112, fold enrichment 2.98) ([38])(**Table 2** row 2). Similarly, 35% and 41% of SCRMshaw predictions from the *mappng2.wing* and *haltere_disc* sets overlapped FAIRE data that included wing and haltere cells ([39])(*P*<4.6e-117 and 1.5e-102, fold enrichment of 3.29 and 2.98)(**Table 2**, rows 1, 3). In the mosquito *A. gambiae*, we compared SCRMshaw predictions from the *embryonic_midgut* and *mapping1.salivary* training sets to ATAC-seq data from adult midgut and salivary tissues, observing overlaps of 40% and 37% respectively (*P*<1.5e-102 and 2.15e-93, fold enrichment of 3.45 and 3.22;) (**Table 2**, rows 9, 10). When comparing our SCRMshaw predictions from the *mapping2.wing* set to ATAC-seq data for larval wing tissue in the butterflies *D. plexippus* and *H. himera*, we observed overlaps of 60% and 68% (*P*<5.59e-19 and *P*<9.92e-25, fold enrichment of 1.99, 1.66;)(**Table 2**, rows 11, 12)([40]). The butterflies *J. coenia* and *V. cardui* are exceptions; intriguingly, they show a depletion in SCRMshaw predictions compared to expectation (**Table 2**, rows 13, 14). Whether this is due to a mismatch in the data used for comparison, the state of the genome assemblies for these two draft genomes, or some other failure of SCRMshaw to perform strongly on these species remains to be determined. Overall, the highly significant overlap we observe between SCRMshaw predictions and open chromatin regions in most species and tissues provides further strong evidence that SCRMshaw is effectively predicting enhancers across a broad range of genomes.

Reporter gene analysis demonstrates that SCRMshaw predictions are functional enhancers

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

We focused our *in vivo* validation experiments on SCRMshaw predictions made using the *mapping2.wing, haltere_disc,* and *disc.mapping2* training sets and a set of four species we had previously shown to be amenable to SCRMshaw prediction*: D. melanogaster, A. aegypti, T. castaneum,* and *A. mellifera*. The imaginal discs are well-established tissues for investigating gene regulation in *D. melanogaster* (although see Discussion for a caveat on using imaginal discs for evaluating enhancer activities in a cross-species setting), and the chosen training sets all gave significant results in the open-chromatin comparisons discussed above. We selected six sets of putative enhancers for testing (**Table 3**, **Table 4**). For the first three sets, we selected sequences where we had predictions in each of the orthologous loci for *D. melanogaster*, *T. castaneum,* and *A. mellifera* (*A. aegypti* was not considered for these sets), and where the *D.*

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

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

Table 2

Overlap of SCRMshaw predictions with FAIRE-seq and ATAC-seq peaks

melanogaster prediction mapped near a gene expressed in the wing imaginal discs (**Fig. 5**). These predictions were conducted using SCRMshaw's IMM scoring method only, with post-processing performed using the original method described in (21), and the Amel_4.5 version of the *A. mellifera* genome. For the second three sets, we chose sequences where we had predictions in orthologous loci for at least three of the four species and where the *D. melanogaster* sequence had previously been tested in a reporter gene assay and was known to be active in the relevant imaginal discs (**Fig. 5**). Imposing this latter criterion allowed us to leverage existing knowledge and reduce the necessary amount of *in vivo* testing for each set of predictions, enabling us to test a larger set of sequences overall. For this second set of three, we used all three SCRMshaw scoring methods with a revised post-processing algorithm (see Methods), and the Amel_Hav3.1 *A. mellifera* genome. For all six sets, predictions were chosen for testing based on high SCRMshaw scores and overlap with open-chromatin data (where available). We also considered the position of each prediction within the locus (i.e., first intron, downstream intergenic region, etc.), favoring sequences where position was maintained among the orthologs. Selected sequences were cloned into a cross-species compatible reporter vector (18, 41), and reporter gene activity was visualized either directly or by using the lineage tracing system G-TRACE (42). In the latter, enhancer activity is visualized through two reporters; the first reporter visualizes the direct enhancer activity while expression of the second reporter is induced and maintained in all cells that descend from a cell in which the enhancer is initially active, even if activity subsequently shuts off.

Ex

The *expanded (ex)* gene plays a crucial role in tissue growth control, including wings (43–45). There was a single prediction in the *D. melanogaster ex* locus, *Dm_ex_20p0*, which falls within an open chromatin region of the third intron (Supplemental Fig. S2), and which comprises an untested subsequence of a longer sequence that acts as an imaginal disc enhancer (**Fig. 5B**; (45)). This sequence drove reporter activity in the wing, leg, and antennal discs (**Fig. 6A**). The *T. castaneum* genome also had only one prediction for the *ex* locus, *Tc_ex_9p0*, which overlaps well with a FAIRE-seq peak (Supplemental Fig. S2) within the second intron. Although no imaginal disc activity was observed (**Fig. 6B**), *Tc_ex_9p0*-driven reporter gene expression was observed during late pupal stages in the legs (Supplemental Fig. S4G). The *A. mellifera* genome (v4.5) had one prediction, *Am_ex_20p3*, within the fourth intron of the *ex* locus (Supplemental Fig. S2; however, note that subsequent prediction using the updated Amel_Hav3.1 genome and revised SCRMshaw post-processing yielded additional predictions in this locus). *Am_ex_20p3* drove active but variable expression in both the pouch and notum regions of the wing imaginal disc (**Fig. 6C**). Although often significantly limited to a small number of cells, the pouch expression of *Am_ex_20p3* was similar in pattern to that seen with *Dm_ex_20p0* (cf. **Fig. 6A**). Weak and inconsistent expression in the pouch region of the haltere discs was also observed. In addition, *Am_ex_20p3* drove expression in the leg discs (**Fig. 6C**).

Klu

Klumpfuss (klu) encodes a zinc finger protein important for proper tissue specification and differentiation (46). The *D. melanogaster* genome had three predictions for the *klu* locus (Supplemental Fig. S2). We selected *Dm_klu_16p1*, which overlaps a region of open chromatin within the second intron present in most imaginal discs (most distinct in the T3 leg disc; Supplemental Fig. S2). *Dm_klu_16p1* drove active expression in the notum regions of the wing and haltere discs and in the leg discs (**Fig. 6D**). *Tc_klu_8p6* was the only prediction for the *T. castaneum klu* locus. It falls within the second intron, similar to *D. melanogaster's klu* prediction *Dm_klu_161p1* (Supplemental Fig. S2), but lacked enhancer activity (**Fig. 6E**). *A. mellifera* had only one prediction for the *klu* locus (*Am_klu_20p2*) (Supplemental Fig. S2). We included this prediction for validation even though its location in the third intron does not match with that in the other two species. *Am_klu_20p2* drove weak expression in the most proximal part of the leg

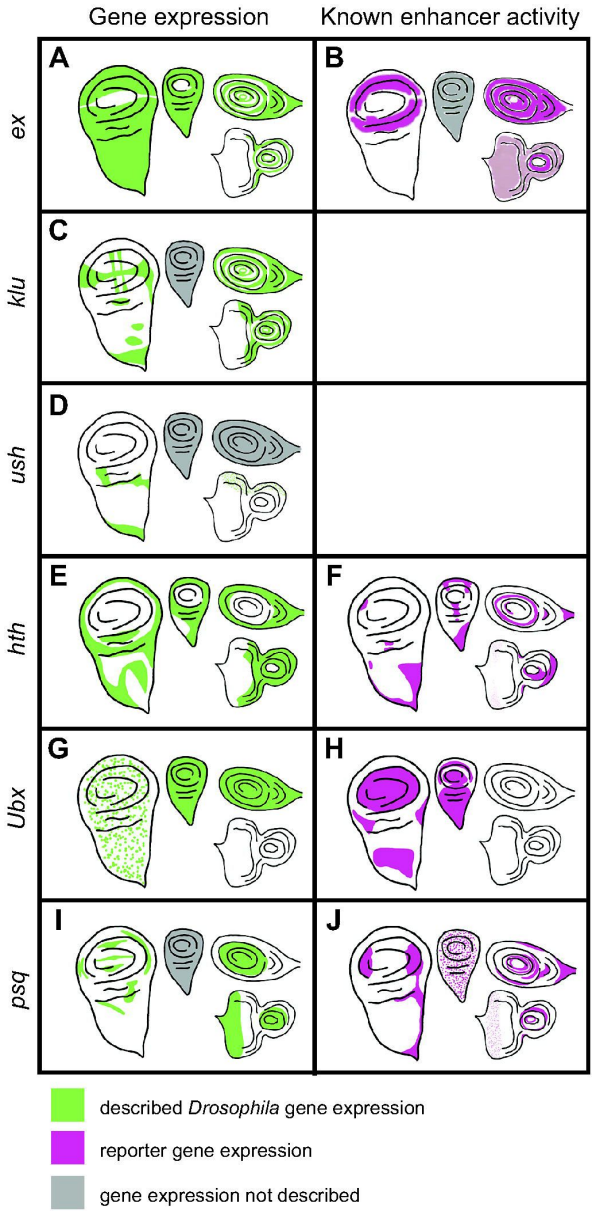


Figure 5

Previously described gene expression and enhancer activity for select *D. melanogaster* sequences predicted by SCRMshaw.

The lefthand column shows native *D. melanogaster* gene expression in imaginal discs (green), while the righthand column shows described enhancer activity (magenta). Gray shading indicates that expression has not been described. Moving clockwise from the left side of each panel are the wing, haltere, leg, and eye-antennal discs. The enhancers whose activities are described in the table are: (B) *ex_BCDE* (45), (F) *hth_GMR46D04* (50), (H) *Ubx_GMR39A02* (50), (J) *psq_GMR41E12* (50).

Table 3

Gene loci chosen for *in vivo* validation

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

Table 4

SCRMshaw predictions chosen for *in vivo* validation

	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

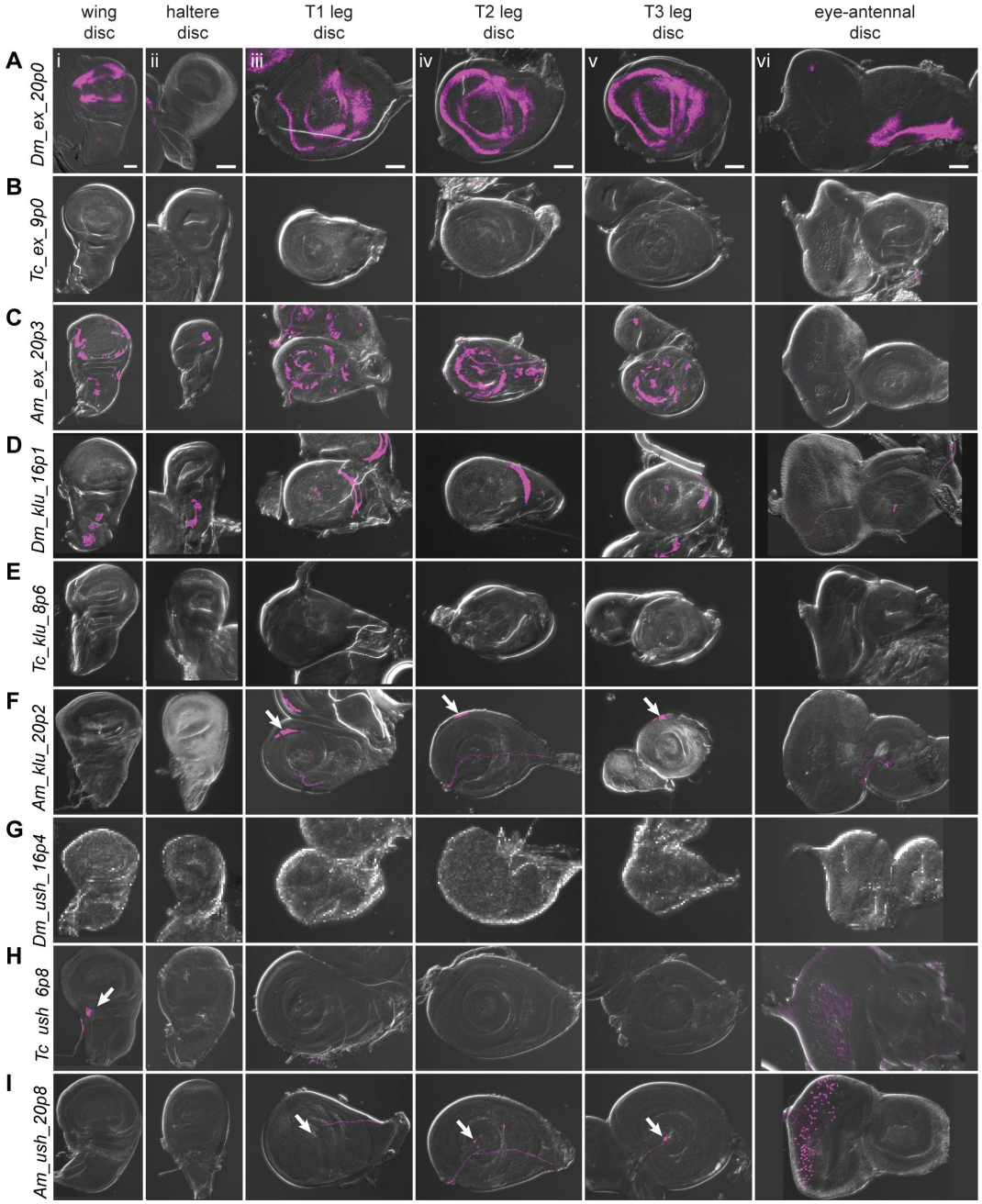


Figure 6

Reporter gene expression for tested *ex*, *klu*, and *ush* predicted enhancer sequences.

Each row shows expression for the indicated construct in (i) wing discs, (ii) haltere discs, (iii) T1 (prothoracic) leg discs, (iv) T2 (mesothoracic) leg discs, (v) T3 (metathoracic) leg discs, and (vi) eye-antennal discs (with eye portion to the left). Positive results were obtained by the enhancers associated with the *ex* locus of *D. melanogaster* (wing, legs, antenna, A) and *A. mellifera* (wing, haltere, legs, C); the *klu* locus of *D. melanogaster* (wing, haltere, legs, D) and *A. mellifera* (legs, arrows in F); the *ush* locus of *T. castaneum* (wing, arrow, H(i); eye, H(vi)) and *A. mellifera* (legs, arrows I(iii, iv, v); eye, I(vi)). Enhancer activities were visualized by UAS-tdTomato that was included in the reporter construct. Scale bar is 50 μm for each column.

discs (**Fig. 6F**, arrows), but no activity was observed in the other imaginal discs. (Note that additional predictions were subsequently produced with our revised post-processing algorithm and the updated Amel_Hav3.1 genome, for both the *T. castaneum* and *A. mellifera klu* loci).

Ush

Our final choice from our first set for *in vivo* validation was *u-shaped* (*ush*), which encodes a transcription factor with described imaginal disc expression (47–49). The *D. melanogaster* genome had two predictions for the *ush* locus (three when using the updated post-processing algorithm) (Supplemental Fig. S2). We tested *Dm_ush_16p4*, but did not observe any larval disc activity (**Fig. 6G**). *Tc_ush_6p8* and *Am_ush_20p8* were the only predictions for *T. castaneum* and *A. mellifera*, respectively, both located in the second intron (again, additional predictions are found using updated methods and genome versions). *Tc_ush_6p8* had expression in the wing disc as well as weak activity in the peripodial membrane of the eye-antennal disc (**Fig. 6H**, Supplemental Fig. S4K). *Am_ush_20p8* displayed active expression in the peripodial membrane surrounding the eye disc, as well as in a single cell, or small subset of cells, located at the center of each leg disc (**Fig. 6I**, arrows).

Hth

SCRMshaw predicted ten enhancers in the locus of the *D. melanogaster* Hox cofactor *homothorax* (*hth*). One of the two top-scoring predictions, *Dm_hth_30p1*, located in the fifth intron, overlapped the known *hth_GMR46D04* enhancer, which has activity in all of the larval imaginal discs (**Fig. 5F**, Supplemental Fig. S3)(50). *A. aegypti* had a single prediction in the orthologous *hth* locus, *Aa_hth_35p9*, located in the first intron (Supplemental Fig. S3). This sequence failed to display activity in imaginal discs (**Fig. 7A**).

T. castaneum had 23 *hth*-locus predictions. We selected a high-scoring (albeit not the highest-scoring) sequence, *Tc_hth_15p5*, due to the similarity of its position to that of the *D. melanogaster* enhancer—in the fifth intron—and overlapping FAIRE peak (Supplemental Fig. S3)(18). The *Tc_hth_15p5* reporter showed activity in the presumptive notum region of the wing disc and in the proximal region of the leg discs, resembling both the endogenous *D. melanogaster hth* expression pattern and that of the *GMR46D04* enhancer (**Fig. 7B**, cf. **Fig. 5F**). No enhancer activity was observed in the haltere or eye-antennal discs (**Fig. 7B**). In the pupal stage, *Tc_hth_15p5* exhibited active expression along the thorax, corresponding to adult *hth* expression in *D. melanogaster* (Supplemental Fig. S4I, arrows) (51). No *hth* predictions were obtained for *A. mellifera*.

Ubx

The classic homeotic gene *Ultrabithorax* (*Ubx*) regulates tissue identity in the thoracic and abdominal segments (52). Of six SCRMshaw predictions in the *D. melanogaster Ubx* locus, we noted that one, *Dm_Ubx_36p1*, overlaps a cluster of known enhancers in the third intron centered on *Ubx_GMR39A02* and *Ubx_abx6.8* (Supplemental Fig. S3). *Ubx_GMR39A02* mirrors the haltere activity of *Ubx*, but also displays ectopic activity in the pouch and notum regions of the wing disc (**Fig. 5H**)(50). Similarly, *Ubx_abx6.8* also drives both native and ectopic expression in the imaginal discs (53).

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

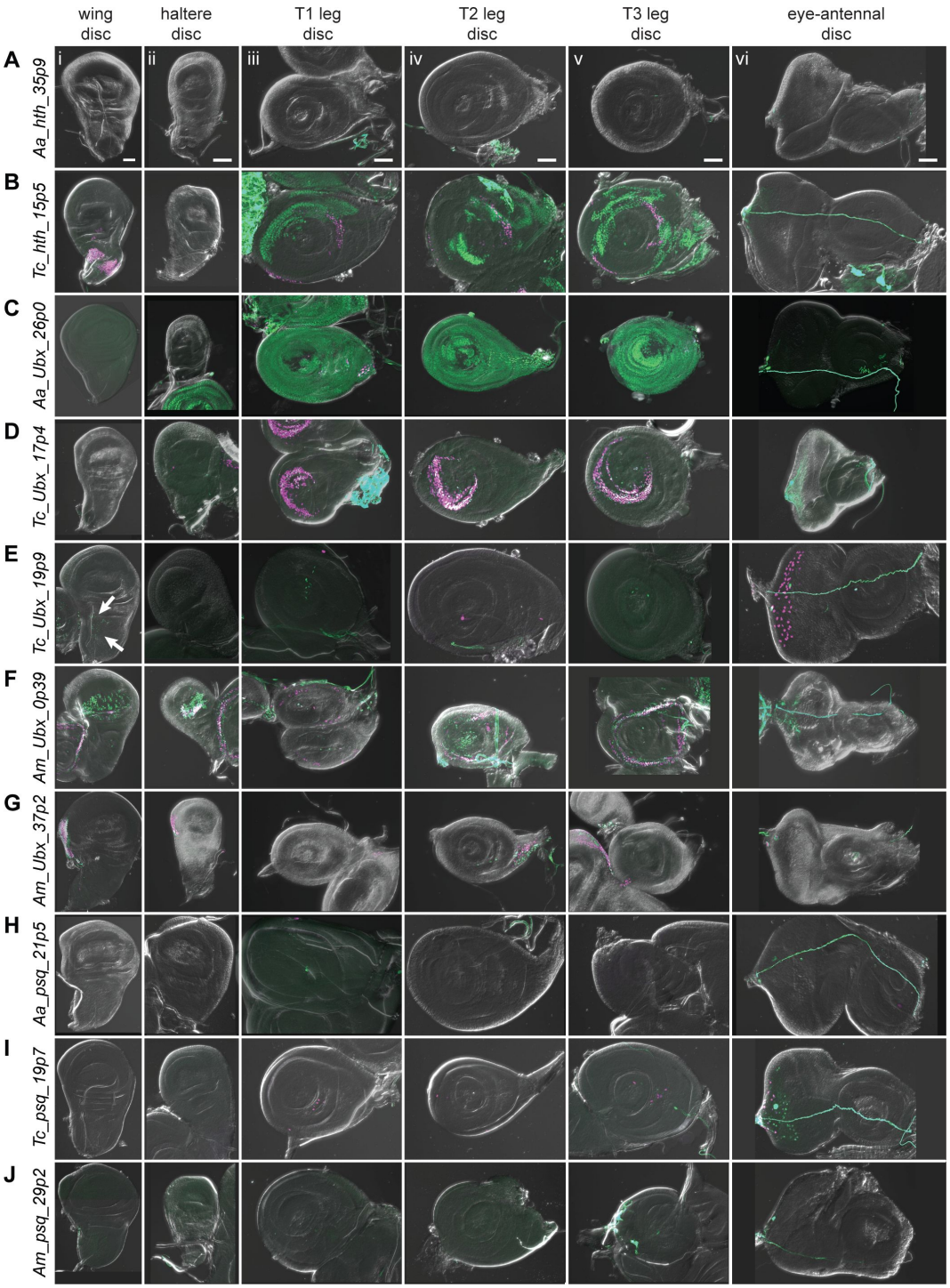


Figure 7

Reporter gene expression for tested *hth*, *Ubx*, and *psq* predicted enhancer sequences.

Each row shows expression for the indicated construct in (i) wing discs, (ii) haltere discs, (iii) T1 (prothoracic) leg discs, (iv) T2 (mesothoracic) leg discs, (v) T3 (metathoracic) leg discs, and (vi) eye-antennal discs (with eye portion to the left). Positive results were obtained by the enhancers associated with the *hth* locus of *T. castaneum* (the notum portion of the wing disc, legs, B); the *Ubx* locus of *A. aegypti* (legs, C), *T. castaneum* (legs, D) (myoblast cells in the wing disc, arrows, E9i); legs, eye, E(iii-vi)), and *A. mellifera* (wing, haltere, legs, F) (wing, haltere, G); the *psq* locus of *T. castaneum* (legs, eye, I). Enhancer activities were detected by the G-TRACE system; magenta represents direct enhancer activity detected by dsRed expression, while green indicates lineage-based GFP expression. Scale bar is 50 μm for each column.

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

The *A. mellifera* genome had eleven predictions at the *Ubx* locus (Supplemental Fig. S3). *Am_Ubx_37p2* was chosen based on a high SCRMshaw score, although it falls within the fourth, rather than the third, intron. *Am_Ubx_0p39*, on the other hand, was in the corresponding third intron location. *Am_Ubx_0p39* had activity in the pouch region of the wing and haltere discs, as well as in the proximal region of leg discs (**Fig. 7F**). *Am_Ubx_37p2* drove expression in specific portions of the wing and haltere discs (**Fig. 7G**). Although *Ubx* is not expressed in the *D. melanogaster* wing disc (i.e., the forewing of the fly), wing activity has been observed previously with tested *Ubx* enhancer fragments (e.g. 50). Moreover, *Ubx* is expressed in both the forewing and hindwing discs in honeybees, making the reporter expression we observed driven by the two predicted *A. mellifera* enhancers consistent with their potential native activities (54).

Psq

We chose *pipsqueak* (*psq*), a transcription factor involved in Polycomb group gene silencing (55), and its orthologous loci as the final targets for enhancer validation. There were four predicted *psq* enhancers in *D. melanogaster*, one of which, *Dm_psq_25p6,* overlaps known enhancer *psq_GMR41E12* (Supplemental Fig. S3). This enhancer is located within the second *psq* intron and drives expression in all imaginal discs (**Fig. 5J**)(50). The *A. aegypti* genome had four predictions for the *psq* locus (Supplemental Fig. S3); we chose *Aa_psq_21p5*, located in the second intron and with the highest local SCRMshaw score, for validation (Supplemental Fig. S3). However, *Aa_psq_21p5* did not have observable imaginal disc expression (**Fig. 7H**).

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

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

Embryonic activity

Although our SCRMshaw predictions were targeted toward imaginal disc activity, activity was also observed in embryos for many of the reporter lines (Supplemental Fig. S5). Analysis of this activity was complicated by the fact that our reporter lines, while not having any basal activity in imaginal discs (Supplemental Fig. S4, A-E), displayed reporter gene expression in several tissues including hemocytes, caudal visceral mesoderm, and the proventriculus, even in the absence of a putative enhancer sequence (Supplemental Fig. S5A-D). Similar expression is seen in control embryos of the G-TRACE line alone, i.e., even in the absence of a Gal4 driver (data not shown), suggesting that the observed basal activity might be coming from the UAS construct. In those lines that had reporter gene expression in other tissues (Supplemental Fig. S5I-S, X, Y), most of that expression was not clearly associated with the expected endogenous expression of the predicted target gene, although the complex embryonic expression patterns of these genes makes a definitive assessment difficult.

Moreover, for most of the species, we do not currently know the expression patterns of either the gene in its native species (as opposed to the expression of its *Drosophila* ortholog), or the expression patterns of other nearby potential target genes. Further analysis, including additional control experiments, use of different reporter vectors, and assessment of gene expression patterns in each of the relevant species will be necessary before drawing final conclusions as to embryonic enhancer activity.

An insect regulatory annotation resource

Taken together, the results from our simulations, our comparisons to open chromatin regions, and our *in vivo* validation experiments demonstrate that SCRMshaw is remarkably effective at predicting regulatory sequences across a wide range of insect species. To facilitate access to our SCRMshaw-based regulatory annotations, we created a database with the results from our 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, *http://redfly.ccr.buffalo.edu*), 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.

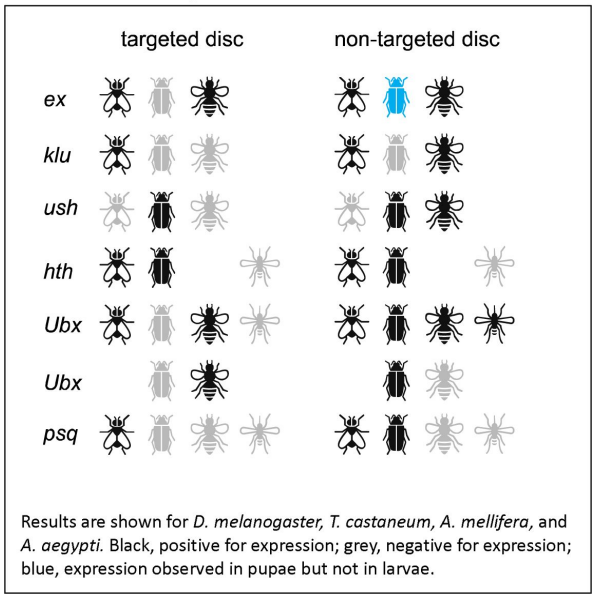
Discussion

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

Validation success rates

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

These numbers may in fact underestimate the success rate of our predictions, for several reasons. One, all of the testing was performed in transgenic *D. melanogaster*, despite the putative enhancers being from three additional species. Reduced efficiency of certain transcription factor or cofactor binding, reduced enhancer-promoter compatibility, or other species-specific differences may lead to elevated false negative results. Furthermore, the unique imaginal disc mode of adult epithelial development in *D. melanogaster* (56, 57) might have prevented some



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.

Table 5

Summary of *in vivo* validation results

enhancers of other species from working properly in *D. melanogaster* imaginal discs, likely producing additional false negative results. Evaluating enhancer activities in the native species will allow us to address the degree of false negatives produced by the cross-species setting. Two, predictions for *ex, klu,* and *ush* were made using a less-effective SCRMshaw post-processing algorithm and, for *A. mellifera,* a less-complete genome build. This may have led to suboptimal candidate enhancer selection. Three, it is possible that some of our predicted enhancers act as silencers rather than enhancers. Silencers, which attenuate rather than promote gene expression, are less well understood than enhancers, but in at least some instances have identical sequence characteristics (and in fact can act simultaneously as enhancers in some tissues and silencers in others) (58, 59). Our reporter gene assay was not designed to detect silencers, which would therefore appear as false-positive predictions. Indeed, an intriguing possibility is that some of our sequences that had reporter gene activity in non-targeted discs (e.g., leg discs) act as enhancers in those tissues but as silencers in the targeted wing discs. Alternatively, identified enhancers that drove expression in multiple imaginal discs may simply represent pleiotropic enhancers.

Enhancer pleiotropy is not uncommon (60, 61), and moreover, there is overlap in the enhancer content of our training sets for the various discs. There are many gene expression and developmental similarities among the discs, and likely significant regulatory overlap; chromatin profiling experiments have found that wing, leg, and eye discs share the majority of their regions of accessible chromatin (18, 39). Additional experiments will be necessary to distinguish between pleiotropic enhancers, dual-function silencers/enhancers, and other possibilities.

Redundant enhancers

Almost all of our training sets predicted multiple enhancers in certain loci at rates exceeding random expectation, consistent with the noted prevalence of redundant or “shadow” enhancers in numerous plant and animal species (reviewed by 34). The few training sets that did not may reflect types of tissues or regulatory processes for which having shadow enhancers is unusual, or may be indicative of poorly-performing training sets. Indeed, the sole training set that did not suggest the presence of shadow enhancers in either mosquito species, *adult_PNS,* also performed poorly by other metrics, such as failing to produce predictions in common over multiple species and scoring “poor” using our *pCRM_eval* training set evaluation tool (21).

Although the evolutionary origins of shadow enhancers are not well understood, several studies suggest that at least one important contribution of shadow enhancers is to provide phenotypic robustness during development, particularly during stress conditions (62–67). In some cases, a mechanistic basis for this robustness has been suggested by the finding that different groups of transcription factors act on individual members of a shadow enhancer set (68, 69). Further support for this idea can be found in the fact that shadow enhancers appear to have limited sequence similarity (68, 70). However, our successes with SCRMshaw have demonstrated that enhancers can have little overt sequence similarity but still rely on the same underlying subsequence (and presumably, transcription factor binding) model. As the potential shadow enhancers we identify are predicted using the same training set, they are likely to be functioning using similar, rather than independent, regulatory mechanisms. Thus, shadow enhancers appear to come in at least two flavors: those that use similar mechanisms, such as those identified here, and those that use different mechanisms, such as the *Krüppel* enhancers studied by (68).

Enhancer evolution at large divergence distances

SCRMshaw’s ability to find putatively “orthologous” enhancers—i.e., enhancers in orthologous loci predicted using the same regulatory model—not only substantiates the non-random nature of our prediction approach, but also opens the door to exciting studies of enhancer evolution over large divergence ranges. We previously illustrated the power of such an approach with an analysis of a small number of enhancers in just 2-3 highly diverged species (17). However, the greatly increased number of species for which we now have regulatory predictions should allow us to

follow the evolution of specific enhancers over their entire divergence range. Although the number of common loci tails off rapidly as the number of species considered increases, we suspect that this small (albeit statistically significant) number understates the true results. For one, only a limited number of species have been evaluated for common predictions so far, and these are not evenly distributed along the phylogenetic spectrum of sequenced species. Also, our analysis depends wholly on the presence of recognized *D. melanogaster* orthologous genes to define the common loci. This in turn is dependent on several factors, including the sensitivity and accuracy of our ortholog-calling pipeline, the reliability of the protein annotations we use in that pipeline, and on accurate target gene assignments. Each of these has known sources of error. For example, our ortholog-calling pipeline currently does not disambiguate multiple paralogs (see Methods), and ortholog detection has been shown to be sensitive to the method used for generating protein annotation, in particular when different annotation methods have been applied to different species in the comparison (71). Moreover, our target gene assignments are presently based solely on closest-gene relationships, a method which is known to mis-assign a fair number of enhancer-gene target pairings (e.g. 24, 25-27) and which does not take into account complexities in genome architecture such as nested genes, long non-coding RNAs, promoter competition, and the like (which are common in *D. melanogaster*, e.g. 72, 73). Developing effective and scalable solutions to these issues will be an important goal for future work.

As the number of species in our prediction database with fully mapped orthologs grows, it will become possible to ask increasingly sophisticated questions about the nature of enhancer evolution. For instance, we will be able to determine whether the numbers of putatively orthologous enhancer predictions follow phylogenetic relationships and degree of sequence divergence, and whether certain loci are only found in common for specific groups of species. Also in need of further investigation will be to determine whether sets of common enhancers are restricted to certain functional Gene Ontology categories, and how this might vary with phylogenetic grouping.

Ongoing regulatory annotation

While effective, SCRMshaw still has various limitations and aspects in need of improvement. Errors in genome assembly and insufficient repeat masking both appear to contribute to overly long predictions (multiple kilobases) that are unlikely to represent individual single enhancers (HA and MSH, unpublished observations). Poor assembly, while not having major effects on SCRMshaw overall, can significantly affect results at specific loci, as can inaccurate gene annotation (21). Also requiring further investigation is how to best combine and weight the scores from the three individual SCRMshaw scoring methods of IMM, hexMCD, and PAC-rc. Individual SCRMshaw predictions should therefore be treated as just that—predictions—and appropriate validation experiments are recommended for any sequences of interest. The regulatory annotations presented here represent initial “1.0” versions of the regulatory genome. Like all genome annotations, these will continue to be revised as updated models become available, including addition of new training sets, confidence scores based on validation experiments, and improved genome builds. Nevertheless, the enrichment scores from our *in silico* experiments and the true-positive rates from our *in vivo* reporter gene experiments suggest that overall, true positive rates for most training sets are between 50-85%, with most likely having success rates exceeding 70%. Note that even this lower bound of a 50% true positive rate means that one out of every two predictions is correct—a more than acceptable rate to encourage follow-up experiments for predicted enhancers in organisms of interest. The catalog of predicted enhancers we introduce here, spanning 33 insect species and growing, should thus be useful resource for both large-scale and small-scale studies of the insect regulatory genome.

Methods

Datasets

Sequence and annotation for *D. melanogaster* were obtained from FlyBase (74). For other species, wherever possible, genome sequence (FASTA) and annotation (GFF) files were downloaded from NCBI at *www.ncbi.nlm.nih.gov/datasets*. Otherwise, the genome sequences and annotations were obtained from the primary literature or directly from the data generators (see **Table 1** for references). Version numbers for genomes and annotations are provided in **Table 1**.

Scripts

All scripts used for the analyses described in this paper can be found in the GitHub repository Asma_etal_2024_eLife at *https://github.com/HalfonLab/Asma_etal_2024_eLife*.

SCRMshaw

A detailed SCRMshaw protocol can be found at Asma et al., 2024 (75).

Genome files were masked using Tandem Repeat Finder (22) using parameters 2 7 7 80 10 50 500 -m -h. Genome and annotation files were then assessed using our *preflight* script. Any sequence scaffolds not containing annotated genes were removed before passing the genome sequence to the main SCRMshaw program. SCRMshaw was run using the "HD" version, which scans the genome with a 500 bp window sliding in 10 bp increments, as described in Asma and Halfon (21). The *--thiwt* parameter was set to 5000, and *--lb* to [0,10,20,30,...,240] for each of 25 instances, respectively. All other parameters were kept at default values.

SCRMshaw makes use of three underlying statistical models, described briefly in (17) and in detail in (13, 14). *HexMCD* trains a fifth-order Markov chain on all 6-mers and their one-away mismatches in the training and background sequence sets, respectively, while *IMM* trains an interpolated Markov model that combines Markov chains of all orders from 0-5. For both of these methods, the score for a sequence in the log-likelihood ratio of the model trained on the training sequences divided by model trained on the background sequences. *PAC-rc* quantifies the overrepresentation of 6-mers in the training versus background sets, assuming a Poisson distribution of word counts. Our previous work has shown that each method is effective, and each tends to have sensitivity for a different group of enhancers. Roughly three-quarters of all predicted enhancers were identified by just a single method (see Supplemental Table S2), although there is a strong correlation between high-ranking predictions and prediction by more than one method. The best results are obtained from combining the output of all three methods.

SCRMshaw can be downloaded at *https://github.com/HalfonLab/SCRMshaw_HD*.

Postprocessing

The top 5000 hits were extracted using scripts *Generate_top_N_SCRM-hits.pl* and *concatenatenatingOffsetResults.sh* and passed to *postProcessingScrmshawPipeline.py* with *-num* = "5000" and *-topN* = "Median." Post-processing was performed essentially as previously described (15, 21), with the following modifications (Supplemental Fig. S6): before evaluating each 10 bp region, scores from each of the 25 individual SCRMshaw instances were assessed and any 500 bp window whose score was below the value of the 5000th ranked score was eliminated by having its score reset to zero (Supplemental Fig. S6B). The "elbow" point of the SCRMshaw score curve of the 5000 top scores from each instance was then determined (Fig. S6C), and any scores below the

elbow point were reset to zero (Supplemental Fig. S6D; gray boxes). This is a key modification to our previous SCRMshawHD protocol (21) and reduces the median prediction size by preventing concatenation of multiple adjacent low-scoring windows. Only after these two rounds of score evaluation were all windows grouped together (Fig. S6E, F) and subjected to peak calling on 10 bp intervals (Fig. S6G). Final "top predictions" were then any peaks with an amplitude above the selected amplitude threshold (elbow point of amplitude curve, represented by a red dot in Supplemental Fig. S6H), following the peak-calling step.

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

Orthology mapping

The final SCRMshaw output from the above steps was used as input to our orthology mapping pipeline. For each species, a FASTA-formatted file of all annotated proteins was downloaded from NCBI (*www.ncbi.nlm.nih.gov/datasets*). The annotated proteins from *D. melanogaster* plus each individual other species were used as input to *Orthologer* (31) to obtain the *Drosophila* ortholog for each protein (when existing). Our approach was designed to be minimally restrictive in that we did not enforce a one-to-one ortholog mapping; in cases of likely paralogs, we considered all of the paralogs as a potential result. Details on the orthology mapping protocol can be found in (75).

Evaluating the number of predictions per locus

For each training set, BEDTools "merge" was used to remove any overlapping predictions (76). The results were then permuted 1000 times using BEDTools "shuffle", with coding regions excluded. BEDTools "closest" was used to assign upstream and downstream flanking genes to each of the permuted predictions, with the following parameters: the '-io' flag was enabled to ignore any overlaps between predictions and genes, and '-D ref-id' and '-D ref-iu' were used to obtain the closest 5' and 3' genes (with respect to chromosome coordinates), respectively. A custom Python script, *checkSameLocus_ForSimulations.py* (see *https://github.com/HalfonLab/Asma_etal_2024_eLife*) was used to calculate the number of predictions per locus for both the real and permuted results. For this purpose, "locus" was defined as the entire region between the left and right flanking genes for each SCRMshaw prediction; that is, for each prediction we take the region between the 3' end of the upstream gene and the transcription start site of the downstream gene. If a prediction is within an intron, we define the locus as the entire span of the enclosing gene. If a prediction overlaps two genes, the locus is considered to be the entire span of the two genes. Average intergenic region sizes for the genomes were estimated using the distances between neighboring genes, discarding any nested or overlapping gene pairs, as provided by the SCRMshaw *preflight* script.

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

Evaluating the number of predictions in common across species

To determine the expected number of common predictions—i.e., predictions in orthologous loci—across species, we utilized SCRMshaw predictions for 16 species. Each set of predictions was sorted, merged, permuted, and mapped to new loci as described above for "Evaluating the number of predictions per locus." The permuted predictions were used as input for the script "*checkSameLocus_ForSimulations_crossSpecies.py*" (see *https://github.com/HalfonLab/Asma_etal_2024_eLife*). This script identifies the *Drosophila* orthologs of the nearest flanking genes for each species and calculates the number of common loci flanking the simulated SCRMshaw predictions for 5, 10, 11, 12, 13, 14, 15, and 16 species. For each species, the process was repeated for a total of

360 permutations. The mean and standard deviation of the permuted results were then used to calculate a *z*-score for each training set. We considered training sets with *z*-score ≥ 1.645 to be significant ($P<0.05$, not corrected for multiple testing).

Overlap between SCRMshaw predictions and open chromatin regions

Open chromatin data were obtained from the following sources:

D. melanogaster

Data for *D. melanogaster* were downloaded from GEO and consisted of ATAC-seq and FAIRE-seq data from accessions GSE101827, GSE38727 and GSE118240. These assays were performed using blastoderm embryos, eye-antennal discs, wing discs, haltere discs, leg discs, third instar central nervous system, and wing, leg, and haltere pharate appendages (38, 39, 77).

T. castaneum

FAIRE-seq data for three stages of embryogenesis, larval central nervous system, and larval second and third thoracic epidermal tissues were downloaded from GEO (GSE104495)(18). The FAIRE profiles were remapped to the version 5.2 of the *T. castaneum* genome (Tcas5.2) for this study. The remapped FAIRE profiles are available on iBeetle-Base (*https://ibeetle-base.uni-goettingen.de/*) (78).

A. gambiae

ATAC-seq data for adult midgut and salivary gland were downloaded from GEO (GSE152924) (79).

D. plexippus, J. coenia, H. himera, and V. cardui

ATAC-seq data for larval forewing and hindwing tissues at stage M5 were provided by Anyi Mazo-Vargas and Robert Reed (40).

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

To assess significance, the SCRMshaw predictions were permuted 500 times using BEDTools “shuffle” and the overlaps with open chromatin regions assessed as above. The mean and standard deviation of the permuted results were then used to calculate a *z*-score for each training set. We also calculated a “fold enrichment” score by dividing the observed number of overlapping regions by the expected number, to provide a sense of effect size in addition to statistical significance.

Reporter constructs and transgenic *Drosophila*

Sequences for reporter gene analysis, including attL1 and attL2 sites, were synthesized *de novo* and cloned into pUC57 Kan-r (GenScript, Piscataway NJ) as entry vectors suitable for Gateway cloning (80). Gateway LR recombination was then used to move the sequences into piggyPhiGUGd (*hth, Ubx,* and *psq* lines) and piggyPhiGUGd-TomatoI (*ex, klu,* and *ush* lines) (41) (These reporter vectors are available from the *Drosophila* Genomics Resource Center (*https://dgrc.bio.indiana.edu*)). Transgenic flies were generated by BestGene (Chino Hills, CA) using PhiC31 recombination and the attP2 third chromosome insertion site. piggyPhiGUGd lines were subsequently crossed to G-TRACE for visualizing enhancer activities (42).

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

Database implementation

The SCRMshaw results database is implemented as part of REDfly (RRID:SCR_006790) (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.

Data availability

All SCRMshaw output generated in this study has been deposited at Dryad at *https://doi.org/10.5061/dryad.3j9kd51t0* in both the standard SCRMshaw output format (modified BED) and in GFFv3 format. Data can also be obtained via the REDfly database at *http://redfly.ccr.buffalo.edu*, as described above.

Acknowledgements

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 2019-67013-29354 to M.S.H. and Y.T.).

References

1. NCBI. NCBI Datasets: Genome, Available from (2024) **NCBI. NCBI Datasets: Genome [Available from2024 [: https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=33208].**

2. Carroll SB, Grenier JK, Weatherbee SD (2005) **From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design**

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

4. Claringbould A, Zaugg JB (2021) **Enhancers in disease: molecular basis and emerging treatment strategies** *Trends Mol Med* **27**:1060–73

5. Rickels R, Shilatifard A (2018) **Enhancer Logic and Mechanics in Development and Disease** *Trends Cell Biol* **28**:608–30

6. Smith E, Shilatifard A (2014) **Enhancer biology and enhanceropathies** *Nature structural & molecular biology* **21**:210–9

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

8. Benton ML, Talipineni SC, Kostka D, Capra JA (2019) **Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function** *BMC Genomics* **20**

9. Halfon MS (2019) **Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases** *Trends Genet* **35**:93–103

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

11. IUCN. The IUCN list of threatened species, Available from (2022) **IUCN. The IUCN list of threatened species [Available from2022 [: https://www.iucnredlist.org.**

12. Royal Entomological Society. Understanding Insects: Facts and figures St. Albans, UK, Available from (2023) **Royal Entomological Society. Understanding Insects: Facts and figures St. Albans, UK [Available from2023 [: https://www.royensoc.co.uk/understanding-insects/facts-and-figures/.**

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

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

15. Kazemian M, Halfon MS (2019) **CRM Discovery Beyond Model Insects** *Methods Mol Biol* **1858**:117–39

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

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

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

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

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

21. Asma H, Halfon MS (2019) **Computational enhancer prediction: evaluation and improvements** *BMC bioinformatics* **20**

22. Benson G (1999) **Tandem repeats finder: a program to analyze DNA sequences** *Nucleic Acids Res* **27**:573–80

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

24. Sanyal A, Lajoie BR, Jain G, Dekker J (2012) **The long-range interaction landscape of gene promoters** *Nature* **489**:109–13

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

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

27. Qin T, Lee C, Li S, Cavalcante RG, Orchard P, Yao H, et al. (2022) **Comprehensive enhancer-target gene assignments improve gene set level interpretation of genome-wide regulatory data** *Genome Biol* **23**

28. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. (2017) **GeneHancer: genome-wide integration of enhancers and target genes in GeneCards** *Database: the journal of biological databases and curation*

29. Gschwind AR, Mualim KS, Karbalayghareh A, Sheth MU, Dey KK, Jagoda E, et al. (2023) **An encyclopedia of enhancer-gene regulatory interactions in the human genome** *bioRxiv*

30. Whalen S, Truty RM, Pollard KS (2016) **Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin** *Nat Genet* **48**:488–96

31. Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, et al. (2023) **OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity** *Nucleic Acids Res* **51**:D445–D51

32. Grosveld F, van Staalduinen J, Stadhouders R (2021) **Transcriptional Regulation by (Super)Enhancers: From Discovery to Mechanisms** *Annu Rev Genomics Hum Genet* **22**:127–46

33. Weinstein ML, Jaenke CM, Asma H, Spangler M, Kohnen KA, Konys CC, et al. (2023) **A novel role for trithorax in the gene regulatory network for a rapidly evolving fruit fly pigmentation trait** *PLoS Genet* **19**

34. Kvon EZ, Waymack R, Gad M, Wunderlich Z (2021) **Enhancer redundancy in development and disease** *Nat Rev Genet* **22**:324–36

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

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

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

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

39. McKay DJ, Lieb JD (2013) **A Common Set of DNA Regulatory Elements Shapes Drosophila Appendages** *Developmental Cell* **27**:306–18

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

41. Deem KD, Halfon MS, Tomoyasu Y (2024) **A new suite of reporter vectors and a novel landing site survey system to study cis-regulatory elements in diverse insect species** *Scientific reports* **14**

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

43. Boedigheimer M, Laughon A (1993) **Expanded: a gene involved in the control of cell proliferation in imaginal discs** *Development* **118**:1291–301

44. Wang LH, Baker NE (2015) **Salvador-Warts-Hippo pathway in a developmental checkpoint monitoring helix-loop-helix proteins** *Dev Cell* **32**:191–202

45. Wang LH, Baker NE (2018) **Spatial regulation of expanded transcription in the Drosophila wing imaginal disc** *PLoS One* **13**

46. Klein T, Campos-Ortega JA (1997) **klumpfuss, a Drosophila gene encoding a member of the EGR family of transcription factors, is involved in bristle and leg development** *Development* **124**:3123–34

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

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

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

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

51. Aldaz S, Morata G, Azpiazu N (2005) **Patterning function of homothorax/extradenticle in the thorax of Drosophila** *Development* **132**:439–46

52. Lewis EB (1978) **A gene complex controlling segmentation in Drosophila** *Nature* **276**:565–70

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

54. Prasad N, Tarikere S, Khanale D, Habib F, Shashidhara LS (2016) **A comparative genomic analysis of targets of Hox protein Ultrabithorax amongst distant insect species** *Scientific reports* **6**

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

56. Cohen SM., Bate M, Martinez Arias A (1993) **Imaginal disc development** *The development of Drosophila melanogaster*

57. Svacha P (1992) **What are and what are not imaginal discs: reevaluation of some basic concepts (Insecta Holometabola)** *Dev Biol* **154**:101–17

58. Halfon MS (2020) **Silencers, Enhancers, and the Multifunctional Regulatory Genome** *Trends Genet* **36**:149–51

59. Segert JA, Gisselbrecht SS, Bulyk ML (2021) **Transcriptional Silencers: Driving Gene Expression with the Brakes On** *Trends Genet* **37**:514–27

60. Laiker I, Frankel N (2022) **Pleiotropic Enhancers are Ubiquitous Regulatory Elements in the Human Genome** *Genome biology and evolution* **14**

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

62. Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL (2010) **Phenotypic robustness conferred by apparently redundant transcriptional enhancers** *Nature* **466**:490–3

63. Perry MW, Boettiger AN, Bothma JP, Levine M (2010) **Shadow enhancers foster robustness of Drosophila gastrulation** *Curr Biol* **20**:1562–7

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

65. Osterwalder M, Barozzi I, Tissieres V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al. (2018) **Enhancer redundancy provides phenotypic robustness in mammalian development** *Nature* **554**:239–43

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

67. Sagai T, Amano T, Maeno A, Kiyonari H, Seo H, Cho SW, et al. (2017) **SHH signaling directed by two oral epithelium-specific enhancers controls tooth and oral development** *Scientific reports* **7**

68. Waymack R, Fletcher A, Enciso G, Wunderlich Z (2020) **Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic** *eLife* **9**

69. Cannavo E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, et al. (2016) **Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks** *Curr Biol* **26**:38–51

70. Barth NKH, Li L, Taher L (2020) **Independent Transposon Exaptation Is a Widespread Mechanism of Redundant Enhancer Evolution in the Mammalian Genome** *Genome biology and evolution* **12**:1–17

71. Weisman CM, Murray AW, Eddy SR (2022) **Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes** *Curr Biol* **32**:2632–9

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

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

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

75. Asma H, Liu L, Halfon MS (2024) **SCRMshaw: supervised cis-regulatory module prediction for insect genomes** *protocols.io* https://doi.org/10.17504/protocols.io.e6nvw1129lmk/v2

76. Quinlan AR, Hall IM (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–2

77. Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, et al. (2018) **The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes** *Nat Genet* **50**:1011–20

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

79. Ruiz JL, Ranford-Cartwright LC, Gomez-Diaz E (2021) **The regulatory genome of the malaria vector Anopheles gambiae: integrating chromatin accessibility and gene expression** *NAR Genom Bioinform* **3**

80. Katzen F (2007) **Gateway((R)) recombinational cloning: a biological operating system** *Expert Opin Drug Discov* **2**:571–89

81. Asma H, Halfon MS (2021) **Annotating the Insect Regulatory Genome** *Insects* **12**

Editors

Reviewing Editor
Yamini Dalal
National Cancer Institute, Bethesda, United States of America

Senior Editor
Yamini Dalal
National Cancer Institute, Bethesda, United States of America

Reviewer #1 (Public review):

Summary:

The authors provide an genome annotation resource of 33 insects using a motif-blind prediction methods for tissue-specific cis-regulatory modules. This is a welcome addition that may facilitate further research in new laboratory systems, and the approach seem to be relatively accurate, although it should be combined with other sources of evidence to be practical.

Strengths:

The paper clearly presents the resource, including the testing of candidate enhancers identified from various insects in Drosophila. This cross-species analysis, and the inherent suggestion that training datasets generated in flies can predict a cis-regulatory activity in distant insects, is interesting. While I can not be sure this approach will prevail in the future, for example with approaches that leverage the prediction of TF binding motifs, the SCRMShaw tool is certainly useful and worth of consideration for the large community of genome scientists working on insects.

Weaknesses from the previous version were appropriately corrected in this revision, as the authors improved data availability including with genome annotation resources.

https://doi.org/10.7554/eLife.96738.2.sa2

Reviewer #3 (Public review):

Summary:

In this ambitious paper, the authors develop an unparalleled community resource of insect genome regulatory annotations spanning five insect orders. They employ their previously-

developed SCRMshaw method for computational cross-species enhancer prediction, drawing on available training datasets of validated enhancer sequence and expression from Drosophila melanogaster, which had been previously shown to perform well across select holometabolous insects (representing 160-345MY divergence). In this work they expand regulatory sequence annotation to 33 insect genomes spanning Holometabola and Hemiptera, which is even more distantly related to the fly model. They perform multiple downstream analyses of sets of predicted enhancers to assess the true-positive rate of predictions; the independent comparisons of real predictions with simulated predictions and with chromatin accessibility data, as well as the functional validation through reporter gene analysis strengthen their conclusions that their annotation pipeline achieves a high true-positive rate and can be used across long divergence times to computationally annotate regulatory genome regions, an ability that has been largely inaccessible for non-model insects and now is possible across the many newly-sequenced insect scaffold-level genomes.

Strengths:

This work fills a large gap in current methods and resources for predicting regulatory regions of the genome, a task that has long lagged behind that of coding region prediction and analysis.

Despite technical constraints in working outside of well-developed model insect systems, the authors creatively draw on existing resources to scaffold a pipeline and independently assess likelihood of prediction validity.

The established database will be a welcome community resource in its current state, and even more so as the authors continue to expand their annotations to more insect genomes as they indicate. Their available analysis pipeline itself will be useful to the community as well for research groups that may want to undertake their own regulatory genome annotation.

Weaknesses:

The work here is limited by the field-wide lack of an independently validated set of tissue specific enhancers that could be used to directly benchmark this pipeline. The prediction of true positive enhancer identification rates and in vivo reporter gene assays offer some insight into the rates of successful prediction, but the output of SCRMshaw regulatory annotation should be regarded as another prediction-generating tool.

https://doi.org/10.7554/eLife.96738.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Strengths:*
>
> *The paper clearly presents the resource, including the testing of candidate enhancers identified from various insects in Drosophila. This cross-species analysis, and the inherent suggestion that training datasets generated in flies can predict a cis-regulatory activity in distant insects, is interesting. While I can not be sure this approach will prevail in the future, for example with approaches that leverage the prediction of TF binding motifs, the SCRMShaw tool is certainly useful and worth consideration for the large community of genome scientists working on insects.*

We thank the reviewer for the positive comments, and would just like to point out that we agree: while we cannot of course know if other methods will overtake SCRMshaw for enhancer prediction—we assume they will, at some point (although motif-based approaches have not fared as well in the past)—for now, SCRMshaw provides strong performance and is a useful part of the current toolkit.

> *Weaknesses:*
>
> *While the authors made the effort to provide access to the SCRMShaw annotations via the RedFly database, the usefulness of this resource is somewhat limited at the moment. First, it is possible to generate tables of annotated elements with coordinates, but it would be more useful to allow downloads of the 33 genome annotations in GFF (or equivalent) format, with SCRMshaw predictions appearing as a new feature. Also, I should note that unlike most species some annotations seem to have issues in the current RedFly implementation. For example, Vcar and Jcoen turn empty.*

We have addressed these weaknesses in several ways:

(1) We have created GFF versions of the SCRMshaw predictions and provide them standalone and also merged into the available annotation GFFs for each of the 33 species

(2) We have made these GFF files, and also the original SCRMshaw output files, available for download in a Dryad repository linked to the publication (https://doi.org/10.5061/dryad.3j9kd51t0).

(3) We have added the inadvertently omitted species to the REDfly/SCRMshaw database.

We agree that the database functions are still somewhat limited, but note that database development is ongoing and we expect functionality to increase over time. In the meantime, the Dryad repository ensures that all results reported in this paper are directly available.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *… Upon identification of predicted enhancer regions, the authors perform post-processing step filtering and identify the most likely predicted enhancer candidates based on the proximity of an orthologous target gene. …*

We respectfully point out a small misunderstanding here on the part of the reviewer. We stress that putative target gene assignments and identities have no impact at all on our prediction of regulatory sequences, i.e., they are not "based on the proximity of an orthologous target gene." Predictions are solely based on sequence-dependent SCRMshaw scores, with no regard to the nature or identities of nearby annotated features. Putative target genes are mapped to Drosophila orthologs purely as a convenience to aid in interpreting and prioritizing the predicted regulatory elements. We have added language on page 8 (lines 189ff) to make this more clear in the text.

> *Weaknesses:*
>
> *This work provides predicted enhancer annotations across many insect species, with reporter gene analysis being conducted on selected regions to test the predictions. However, the code for the SCRMshaw analysis pipeline used in this work is not made available, making reproducibility of this work difficult. Additionally, while the authors claim the predicted enhancers are available within the REDfly database, the predicted*

> *enhancer coordinates are currently not downloadable as Supplementary Material or from a linked resource.*

We have placed all the code for this paper into a GitHub repository "Asma_etal_2024_eLife" (https://github.com/HalfonLab/Asma_etal_2024_eLife) to address this concern. As described in our response to Reviewer 1, above, all results are now available in multiple formats in a linked Dryad repository in addition to the REDfly/SCRMshaw database.

> *The authors do not validate or benchmark the application of SCRMshaw against other published methods, nor do they seek to apply SCRMshaw under a variety of conditions to confirm the robustness of the returned predicted enhancers across species. Since SCRMshaw relies on an established k-mer enrichment of the training loci, its performance is presumably highly sensitive to the selection of training regions as well as the statistical power of the given k-mer counts. The authors do not justify their selection of training regions by which they perform predictions.*

Our objective in this study was not to provide proof-of-principle for the SCRMshaw method, as we have established the efficacy of the approach at this point in several previous publications. Rather, the objective here was to make use of SCRMshaw to provide an annotation resource for insect regulatory genomics. Note that the training regions we used here are the same as those we have used in earlier work. Naturally, we performed various assessments to establish that the method was working here, but we make no claims in this work about SCRMshaw's relative efficiency compared to other methods. Some of our prior publications include assessments of the sort the reviewer references, which suggest that SCRMshaw is at least comparable to other enhancer discovery approaches. We note that benchmarking of such methods is in fact extremely complicated due to the fact that there are no established true positive/true negative data sets against which to benchmark (we have explored this in Asma et al. 2019 BMC Bioinformatics).

> *While there is an attempt made to report and validate the annotated predicted enhancers using previously published data and tools, the validation lacks the depth to conclude with confidence that the predicted set of regions across each species is of high quality. In vivo, reporter assays were conducted to anecdotally confirm the validity of a few selected regions experimentally, but even these results are difficult to interpret. There is no large-scale attempt to assess the conservation of enhancer function across all annotated species.*

We respectfully disagree that there is insufficient validation. We bring several different lines of evidence to bear suggesting that our results fall into the accuracy range—roughly 75%—established both here and in previous work. We are also clear about the fact that these are predictions only and need to be viewed as such (e.g. line 638). Although "large-scale" in vivo validation assays would certainly be both interesting and worthwhile, the necessary resources for such an assessment places it beyond our present capability.

> *Lastly, it is suggested that predicted regions are derived from the shared presence of sequence features such as transcription factor binding motifs, detected through k-mer enrichment via SCRMshaw. This assumption has not been examined, although there are public motif discovery tools that would be appropriate to discover whether SCRMshaw is assigning predicted regions based on previously understood motif grammar, or due to other sequence patterns captured by k-mer count distributions. Understanding the sequence-derived nature of what drives predictions is within the scope of this work and would boost confidence in the predicted enhancers, even if it is limited to a few training examples for the sake of clarity of interpretation.*

Again, we respectfully disagree that “this assumption has not been examined.” Although we did not undertake this analysis here, we have in the past, where we have shown that known TFBS motifs can be recovered from sets of SCRMshaw predictions (e.g., Kazemian et al. 2014 Genome Biology and Evolution). We return to this point when we address the Comments to Authors, below.

> ***Reviewer #3 (Public Review):***
>
> *Weaknesses:*
>
> *The rates of predicted true positive enhancer identification vary widely across the genomes included here based on the simulations and comparison to datasets of accessible chromatin in a manner that doesn't map neatly onto phylogenetic distance. At this point, it is unclear why these patterns may arise, although this may become more clear as regulatory annotation is undertaken for more genomes.*

We agree that we do not see clear patterns with respect to phylogenetic distance in our results. However, we note that this initial data set is still fairly small, and not carefully phylogenetically distributed. We are hoping that, as the reviewer suggests, some of these questions become more clear as we add more genomes to our analysis. Fortunately, the list of available genomes with chromosome-level assembly is growing rapidly, and as we move ahead we should have much greater ability to choose informative species.

> *Functional assessment of predicted enhancers was performed through reporter gene assays primarily in Drosophila melanogaster imaginal discs, a system amenable to transgenics. Unfortunately, this mode of canonical imaginal disc development is only representative of a subset of all holometabolous insects; therefore, it is difficult to interpret reporter gene expression in a fly imaginal disc as evidence of a true positive enhancer that would be active in its native species whose adult appendages develop differently through the larval stage (for example, Coleopteran and Lepidopteran legs). However, the reporter gene assays from other tissues do offer strong evidence of true positive enhancer detection, and constraints on transgenic experiments in other systems mean that this approach is the best available.*

Please see an extensive discussion of this point in our response to Reviewer 3, below.

> ***Recommendations for the authors:***
>
> ***Reviewer #2 (Recommendations For The Authors):***
>
> *Major Concerns:*
>
> *(1) While the GitHub source code for SCRMshaw is provided, the authors do not provide a repository of manuscriptspecific code and scripts for readers. This is a barrier to reproducibility and the code used to perform the analysis should be made available. Additionally, links to available scripts do not work, see Line 690. Post-processing scripts point to a general lab folder, but again, no specific analysis or code is sourced for the work in this specific manuscript (e.g. Line 637).*

As noted above, we have corrected this oversight and established a specific GitHub repository for this manuscript “Asma_etal_2024_eLife” (https://github.com/HalfonLab/Asma_etal_2024_eLife).

> *(2) On lines 479-488, there is a discussion about the annotations being provided on REDfly, though no link is provided.*

We have included a link in the text at this point (now line 515).

> *Additionally, for transparency, it would be valuable to provide in Supplementary Table 1 the genomic coordinates of the original training sets in addition to their identity.*

These coordinates have been added to Supplementary Table 1 as suggested.

> *Also, it is suggested to provide genomic coordinates of the predicted enhancers for each training set across all species, perhaps with a column denoting a linked ID of one genomic coordinate in a species to another species (i.e. if there is a linked region found from D. melanogaster to J. coenia, labeling this column in both coordinate sets as blastoderm.mapping1_region1). Providing these annotations directly in the work enhances the transparency of the results.*

We are unsure exactly what the reviewer means here by "a linked region." It is critical to understanding our approach to recognize that the genome sequences have diverged to the point where there is no alignment of non-coding regions possible. Thus there is no way to directly "link" coordinates of a predicted enhancer from one species to those of a predicted enhancer in another species. The coordinates for each prediction are available on a per-species basis either through the database or in the files now available in the linked Dryad repository; these can be filtered for results from a specific training set. The database will allow users to select all results for a given orthologous locus, from any subset of species. More complex searches will continue to become available as we improve functionality of the database, an ongoing project in collaboration with the REDfly team.

> *(3) Figure 2B: It is unclear what this figure shows. Are the No Fly Orthologs false positives, Orthology pipeline issues, or interesting biology?*

We have clarified this in the Figure 2 legend. "No Mapped Fly Orthologs" indicates that our orthology mapping pipeline did not identify clear D. melanogaster orthologs. For any given gene, this could reflect either a true lack of a respective ortholog, or failure of our procedure to accurately identify an existing ortholog.

> *(4) SCRMshaw appears to be a versatile tool, previously published in a variety of works. However, in this manuscript, there is little discussion of the sensitivity of SCRMshaw to different initial parameters, how the selection of training loci can impact outcomes, or how SCRMshaw k-mer discovery methods compare to other similar tools.*
>
> *- This paper would be strengthened by addressing this weakness. Some specific suggestions below:*
>
> *In order to strengthen confidence that SCRMshaw is a reliable predictor of enhancer regions in other species, it is suggested that you benchmark against other k-mer-derived methods to assign enhancers, such as GSK-SVM developed by the Beer Lab in 2016 (https://www.beerlab.org/gkmsvm/, https://www.biorxiv.org/content/10.1101/2023.10.06.561128v1).*

We have established the effectiveness of SCRMshaw as an enhancer discovery method in previous work, and the main goal of this study was to make use of the established method to annotate numerous insect genomes as a community resource. Our claim here is that SCRMshaw works well for this purpose; we do not attempt a strong claim about whether other approaches may work equally well or marginally better (although we do not believe this is the case, based on prior work). Benchmarking enhancer discovery is challenging, as we point out in Asma et al. 2019 (BMC Bioinformatics), and, while important, best left for a

dedicated comprehensive study. A major problem is that there are no independent objective "truth" sets for enhancers from the various species we interrogate here. Thus, while we could also run, e.g., GSK-SVM, what criteria would we use to establish which method had better accuracy for a given species? Note that the work from Beer's lab took advantage of the ability to match human-mouse orthologous (or syntenic) regions and available open-chromatin data to assess whether conserved enhancers were discovered, but this is not possible given the degree of divergence, limited synteny, and relative lack of additional data for the insect genomes we are annotating.

> *- In Table S1, we see that 7-146 regions are used as training sets, which is a huge variety. Does an increase in training set size provide a greater "rate of return" for predicted regions? Is the opposite true? Addressing this question would allow readers to understand if they wish to use SCRMshaw, a reasonable scope for their own training region selections.*
>
> *- Within a training set, does subsampling provide the same outcomes in terms of prediction rates? There is no exploration of how "brittle" the training sets are, and whether the generalized k-mer count distributions that are established in a training set are consistent across randomly selected subgroups. Performing this analysis would raise confidence in the method applied and the resulting annotations.*

These are interesting and important questions, but again we feel they are beyond the scope of this particular study, which is focused primarily on using SCRMshaw and not on optimizing various search parameters. That said, this is of course something we have investigated, although as with other aspects of enhancer discovery, the absence of a true gold standard enhancer set makes evaluation difficult. We have not found a clear correlation between training set size and performance beyond the very general finding that performance appears to be best when training set size is moderate, e.g. 20-40 initial enhancers. We suspect that larger training sets often contain too many members that don't fit the core regulatory model and thus add noise, whereas sets that are too small may not contain enough signal for best performance (although small sets can still be useful, especially if used in an iterative cycle; see Weinstein et al. 2023 PLoS Genetics). However, establishing this rigorously is highly challenging given the limitations with assessing true and false positive rates at scale.

> *(5) In Figure 2C, when plotting hexMCD, IMM, pacRC, and then the merged set, it is unclear whether the scorespecific bar allows coordinate redundancy, though this is implied. What might be more useful is a revision of this plot where the hexMCD/IMM/pac-RC-specific loci are plotted, with the merged set alongside as is currently reported. This would give the reader a clearer understanding of the variability between these scoring methods and why this variability occurs.*

We have added the breakdowns between IMM, hexMCD, and pacRC in Supplementary Table S2, and made more complete reference to this in the text (lines 682ff). Both the database and the data files in the Dryad repository allow exploration of the overlap between the different methods and contain both separate and merged (for overlap and redundancy) results.

> *Additionally, there is no information in the Methods section of these three SCRMshaw scores and what they represent, even colloquially. While SCRMshaw has been applied in several papers previously, it would help with scientific clarity to describe in a sentence or two what each score is meant to represent and why one is different from another.*

We had chosen to err on the side of brevity given prior publication of the SCRMshaw methodology, but we recognize now that we went too far in that direction. We have added

more complete descriptions of the methods in both the Results (lines 164-167) and the Methods (lines 667-681) sections.

> *(6) When describing results in Figure 2, an important question arises: "Is there an anti-correlation between the number of predicted regions and evolutionary distance?" This would be an expected result that could complement Figure 4's point that shared orthology across 16 species is rarer than across 10 species. Visualizing and adding this to Figure 2 or Figure 4 would be a powerful statement that would boost confidence in the returned predicted enhancers and/or orthologous regions.*

This is an important question and one in which we are very interested. Unfortunately, we do not have sufficient data at this time to address this proper statistical rigor. As we remarked above in response to Reviewer 3, “We agree that we do not see clear patterns with respect to phylogenetic distance in our results. However, we note that this initial data set is still fairly small, and not carefully phylogenetically distributed. We are hoping that, as the reviewer suggests, some of these questions become more clear as we add more genomes to our analysis. Fortunately, the list of available genomes with chromosome-level assembly is growing rapidly, and as we move ahead we should have much greater ability to choose informative species.”

> *(7) In Figure 3, the authors seek to convey that SCRMshaw predicts enhancer regions that are mapped nearby one another, across different loci widths, and that this occurrence of nearby predicted regions occurs more than a randomly selected control. This is presumably meant to validate that SCRMshaw is not providing predictions with low specificity, but rather to highlight the possibility that SCRMshaw is identifying groups of shadow enhancers. However, these plots are extremely difficult to decipher and do not strongly support the claims due to the low resolution and difficult interpretability of the boxplot interquartile distributions.*

Additionally, as the majority of predicted regions are around ~750bp, how does that address loci groups of <1000bp? This suggests that predicted regions are overlapping, and therefore cannot be meaningfully interpreted as shadow enhancers. This plot should either be moved to the supplements or reworked to more effectively convey the point that "SCRMshaw is detecting predicted regions that are proximal to one another and that this proximity is not due to chance".

> *- A suggestion to rework this plot is to change this instead to a bar plot, where the y-axis instead represents "number of predictions with at least 2 predicted regions proximal to one another" divided by "total number of predictions", separating bar color by simulated/observed values. The x-axis grouping can remain the same. Because this plot is a broad generalization of the statement you're trying to make above, knowing whether a few loci have 2 versus 4 proximal predicted enhancers doesn't enhance your point.*

We agree with the reviewer that these are not the clearest plots, and thank them for the suggestions regarding revision. We tried many variations on visualizing these complex data, including those suggested by the reviewer, and have concluded that despite their weaknesses, these plots are still the best visualization. The main problem is that the observed data cluster heavily around zero, so that the box plots are very squat and mainly only the outlier large values are observed. The key point, however, is that the expected values almost never give values much greater than one, so that the observed outlier points are the only points seen in the upper ranges of the y-axis. This is true across the three species, across the bins of locus sizes, and across training sets (averaged into the box plots). The reviewer is correct as well about the bins where locus size is < 1000. However, inspection of the data shows that this is not a large concern, as very few data points lie in this range and we never see multiple

predicted enhancers there. Thus we believe while not the prettiest of graphs, Figure 3 does effectively support the claims made in the text. In keeping with our view that it is preferable to have data in the main paper whenever possible, we choose to keep the figure in place rather than move it to the Supplement.

> *- Label the species for the reader's understanding of each subplot on the plot.*

We apologize for this oversight and have now labeled each plot with its relevant species.

> *(8) SCRMshaw operates on k-mer count distributions compared to a genomic background across different species, allowing it to assign predicted regions without prior knowledge of an organism's cis-regulatory sequences. This is powerful and boosts the versatility of the method. However, understanding the cis-regulatory origins of the kinds of kmers that are driving the detection of orthologous regions across species is crucial and absolutely within the scope of the paper, particularly for the justification of the provided annotations. Is SCRMshaw making use of enriched motifs within the training region set to assign regions in other species? One would presume so, but it is necessary to show this. There are many motif discovery tools that are readily available and require little up-front knowledge and little to no use of a CLI, such as MEMESuite (https://meme-suite.org/meme/tools/meme). It is highly recommended that, even for a few training pairs that are well understood (e.g. mesoderm.mapping1, dorsal_ectoderm.mapping1), assess the motif enrichment within the original sequence set, then see whether motif enrichments are reflected in the predicted enhancers. As evolutionary distance increases between D. melanogaster and the species of interest, is the assignment of enriched motifs more sparse? Is there a loss of a key motif? These are the kinds of questions that will allow readers to understand how these annotations are assigned as well as boost confidence in their usage.*

This is a very important point and a subject of significant interest to us. We have demonstrated in earlier work (e.g., Kazemian et al. 2014 Genome Biol. Evol.) that SCRMshaw-predicted enhancers do contain expected TFBS motifs, across multiple species—and that even an overall arrangement of sites is sometimes conserved. Thus we have previously answered, in part, the reviewer's question.

What we also learned from our previous work is that filtering out relevant motifs from the noise inherent in motif-finding is both arduous and challenging. As the reviewer is no doubt aware, while using motif discovery tools is simple, interpreting the output is much less so. In response to the reviewer's comments, we revisited this issue with data from a small sample of training sets. We can discover motifs; we can see that the motif profiles are different between different training sets; and we can observe the presence of expected motifs based on the activity profile of the enhancers (e.g., Single-minded binding sites in our mesectoderm/midline training and result data). However, to do this cleanly and with appropriate statistical rigor is beyond what we feel would be practical for this paper. We hope to return to this important question in the future when we have a larger and phylogenetically more evenly-distributed set of species, and the time and resources to address it appropriately.

> *(9) Figures 5-7 need to have better descriptions.*

We have added to the figure 6 and 7 legends in response to this comment; please note as well that there is substantial detail provided in the text. If there are specific aspects of the figures that are not clear or which lack sufficient description, we are happy to make additional changes.

> *(1) In Figure 1A, it is implied that "k-mer count distributions" are actually only "5-mer count distributions". However, in the published documentation of SCRMshaw, it is suggested that k-mers between 1-6 bp are involved in establishing sequence distributions. Please add a justification for the selection of these criteria. It would be helpful to understand the implications of using up to a 3-mer versus a 12-mer when assessing k-mer counts using SCRMshaw.*

We have clarified in the Figure 1 legend that this is just an example, and the k-mers of different sizes are used in the IMM method; we have also increased the description of the basic method in the Methods section. To be clear, the hexMCD sub-method is 6-mer based (5th-order Markov chain), as is pacRC, while the IMM method considers Markov chains of orders 0-5.

> *(2) Control the y-axis to remove white space from Figure 2D.*

We have amended the figure as suggested.

> *Additionally, expand in the manuscript on expected results from SCRMshaw. Given training regions of 750 bp, is the expectation that you return predicted enhancers of the same length? This is not explicitly stated, only a description of outliers.*

The scoring is not dependent on the length of the training sequences, and there is no direct expectation of predicted enhancer length. Scores are calculated on 10-bp intervals, and a peak-calling algorithm is used to determine the endpoints of each prediction based on where the scores drop below a cutoff value. Thus there is no explicit minimum prediction length beyond the smallest possible length of 10-bp. That said, the initial scoring takes place over a 500-bp sequence window (for reasons of computational efficiency), which does influence scores away from the smaller end of the possible range. We correct for this in part by reducing scores below a certain threshold to zero, to prevent multiple low-scoring regions from combining to give a low but positive score over a long interval. Indeed, we found that in the original version of SCRMshawHD (Asma et al. 2019), multiple low-scoring but above-threshold intervals would get concatenated together in broad peaks, leading to an unrealistically large average prediction length. In the version used here, described in Supplementary Figure S6, low-scoring windows are now first reset to zero and a new threshold is calculated before overlapping scores are summed. This helps to prevent the broad peak problem, and we find that it results in a median prediction length ~750 bp, more in line with expected enhancer sizes.

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *Line 161: Given that the SCRMshaw HD method is the basis for the pipeline, the methodology deserves at least an "in brief" recapitulation in this manuscript.*

As we remark in our response to Reviewer 2, above, “We had chosen to err on the side of brevity given prior publication of the SCRMshaw methodology, but we recognize now that we went too far in that direction. We have added more complete descriptions of the methods in both the Results (lines 164-167) and the Methods (lines 667-681) sections.”

> *Line 219: Throughout the reporting of the results, there appeared to be a bit of inconsistency/potential typos regarding whether threshold or exact P values were reported. In lines 219, 222, 265, 696, and 811, the reported values seem to clearly be thresholds (< a standard cutoff), while in lines 291,293, 297,300, values appear to be exact but are reported as thresholds (<).*

This is not an error but rather reflects two different types of analysis. The predictions per locus (originally lines 219, 222 etc) are evaluated using an empirical P-value based on 1000 permutations. As such, they are thresholded at 1/1000. The overlap with open chromatin regions, on the other hand, are based on a z-score with the P-values taken from a standard conversion of z-scores to P-values.

> *Page 13/Table 2: At face value, it seems surprising that the overlap between Dmel SCRMshaw predictions with open chromatin is so much smaller than the overlap between predictions and open chromatin in other species, both in raw % (Tcas, D plexippus, H. himera) and fold enrichment (Tcas), given that the training sets for SCRMshaw are all derived from Dmel data. The discussion here does not touch on this aspect of the results, and the interpretation of this approach, in general, would be strengthened if the authors could comment on potential reasons why this pattern may be arising here, or at least acknowledge that this is an open question.*

There are many variables at play here, as the data are from different species, from different tissues, and from different methods. Thus we think it is difficult to read too much into the precise results from these comparisons—the main take-home is really just that there is a significant amount of overlap. In acknowledgment of this, we have slightly modified the text in this section so that it now notes (line 302ff): “These comparisons are imperfect, as the tissues used to obtain the chromatin data do not precisely correspond to the training sequences used for SCRMshaw, and the data were obtained using a variety of methods.”

> *Line 318-329: The inferences from the reporter gene assay deserve a more nuanced treatment than they are given here. The important nuance that was not addressed by the discussion here is that the imaginal disc mode of development in Drosophila is not broadly representative of the development of larval/adult epithelial tissues across Holometabola; thus, inference of a true positive validation becomes complicated in cases where predicted enhancers from a species were tested and shown to drive expression in a fly imaginal disc that the native species have no direct disc counterpart to. For example, in line 388 a Tcas enhancer is reported to drive expression in the eye-antennal disc, and in lines 404 and 423 additional Tcas enhancers were reported to drive expression in the leg discs; however, Tribolium larvae do not possess antennal discs or leg discs set aside during embryogenesis in the sense that flies do - instead the homologous epithelial tissues form larval antennae and larval legs external to the body wall that are actively used at this life stage and are starkly different in morphology than an internally invaginated epithelial disc, that will directly give rise to adult tissues in subsequent molts. Is the interpretation of an expression pattern driven in a fly disc as a true positive really as straightforward as it was presented here, when in the native species the expression pattern driven by the enhancer in question would be in the context of an extremely different tissue morphology? That said, I understand and am deeply sympathetic to the constraints on the authors in performing transgenic experiments outside of the model fly; but these divergent modes of development across Holometabola deserve a mention and nuance in the interpretation here.*

This is indeed a very important point, and we greatly appreciate Reviewer 3 pointing out this caveat when interpreting the outcomes of our cross-species reporter assay. Reviewer 3 is correct that the imaginal disc mode of adult tissue (i.e. imaginal) development found in Diptera does not represent the imaginal development across Holometabola.

In fact, imaginal development is quite diverse among Holometabola. For instance, larval leg and antennal cells appear to directly develop into the adult legs and antennae in Coleoptera (i.e. primordial imaginal cells function as larval appendage cells), while some cells within the larval legs and antennae are set aside during larval development specifically for adult

appendages in Lepidopteran species (i.e. imaginal cells exist within the larval appendages but do not contribute to the formation of larval appendages). In contrast, an almost entire set of cells that develop into adult epithelia are set aside as imaginal discs during embryogenesis in Diptera. Furthermore, the imaginal disc mode of development appears to have evolved independently in

Hymenoptera. Therefore, determining how imaginal primordial tissues correspond to each other among Holometabola has been a challenging task and a topic of high interest within the evo-devo and entomology communities.

Nevertheless, despite these differences in mode of imaginal development, decades of evo-devo studies suggest that the gene regulatory networks (GRNs) operating in imaginal primordial tissues appear to be fairly well conserved among holometabolan species (for example, see Tomoyasu et al. 2009 regarding wing development and Angelini et al. 2012 regarding leg development between flies and beetles). These outcomes imply that a significant portion of the transcriptional landscape might be conserved across different modes of imaginal development. Therefore, an enhancer functioning in the Tribolium larval leg tissue (which also functions as adult leg primordium) could be active even in the leg imaginal disc of Drosophila, if the trans factors essential for the activation of the enhancer are conserved between the two imaginal tissues.

That being said, we fully expect there to be both false negative and false positive results in our cross-species reporter assay. We are optimistic about the biological relevance of the positive outcomes of our crossspecies reporter assay, especially when the enhancer activity recapitulates the expression of the corresponding gene in Drosophila (for example, Am_ex Fig6B and Tc_hth Fig7B). Nonetheless, the biological relevance of these enhancer activities needs to be further verified in the native species through reporter assays, enhancer knock-outs, or similar experiments.

In recognition of the Reviewer's important point, we added the following caveat in our Discussion (lines 549553): "Furthermore, the unique imaginal disc mode of adult epithelial development in D. melanogaster might have prevented some enhancers of other species from working properly in D. melanogaster imaginal discs, likely producing additional false negative results. Evaluating enhancer activities in the native species will allow us to address the degree of false negatives produced by the cross-species setting." We moreover mention this caveat in the Results section when we first introduce the reporter assays (line 342).

> *Line 580: This is the first time that the weakness of the closest-gene pairing approach is mentioned. This deserves mention earlier in the manuscript, as unfortunately, this is one of the major bottlenecks to this and any other approaches to investigating enhancer function. Could the authors address this earlier, perhaps pages 7-8, and provide citations for current understanding in the field of how often closest-gene pairing approaches correctly match enhancers to target genes?*

We have added text as suggested on p.7-8 acknowledging the shortcomings of the closest-gene approach. We also clarify at the end of that section (lines 173-181) that target gene assignments, while useful for interpretation, have no bearing on the enhancer predictions themselves (which are generated prior to the target gene assignment steps).

https://doi.org/10.7554/eLife.96738.2.sa0