insects

MDPI

Review

Annotating the Insect Regulatory Genome

Hasiba Asma [1] and Marc S. Halfon [1,2,3,4,5,*]

[1] Program in Genetics, Genomics, and Bioinformatics, University at Buffalo-State University of New York, Buffalo, NY 14203, USA; hasibaas@buffalo.edu
[2] Department of Biochemistry, University at Buffalo-State University of New York, Buffalo, NY 14203, USA
[3] Department of Biomedical Informatics, University at Buffalo-State University of New York, Buffalo, NY 14203, USA
[4] Department of Biological Sciences, University at Buffalo-State University of New York, Buffalo, NY 14203, USA
[5] NY State Center of Excellence in Bioinformatics & Life Sciences, Buffalo, NY 14203, USA
* Correspondence: mshalfon@buffalo.edu; Tel.: +1-716-829-3126

Simple Summary: Insects comprise the largest and most diverse class of animals on earth, and have major impacts on human health and agriculture. The effort to better understand insect biology has led to the sequencing of hundreds of insect genomes. However, the usefulness of having a genome sequence is limited in the absence of a comprehensive annotation—a description of the function of each part of the sequence. Functional parts of the genome include not only genes, but also regulatory sequences that mediate gene expression. We discuss here methods used to identify regulatory sequences within the genome, with the emphasis on a pair of tools we have developed, REDfly and SCRMshaw, that can be used in tandem to carry out this task in an efficient and economical manner.

Abstract: An ever-growing number of insect genomes is being sequenced across the evolutionary spectrum. Comprehensive annotation of not only genes but also regulatory regions is critical for reaping the full benefits of this sequencing. Driven by developments in sequencing technologies and in both empirical and computational discovery strategies, the past few decades have witnessed dramatic progress in our ability to identify *cis*-regulatory modules (CRMs), sequences such as enhancers that play a major role in regulating transcription. Nevertheless, providing a timely and comprehensive regulatory annotation of newly sequenced insect genomes is an ongoing challenge. We review here the methods being used to identify CRMs in both model and non-model insect species, and focus on two tools that we have developed, REDfly and SCRMshaw. These resources can be paired together in a powerful combination to facilitate insect regulatory annotation over a broad range of species, with an accuracy equal to or better than that of other state-of-the-art methods.

Keywords: arthropod; genomics; enhancer; cis-regulation; genome annotation

Citation: Asma, H.; Halfon, M.S. Annotating the Insect Regulatory Genome. *Insects* **2021**, *12*, 591. https://doi.org/10.3390/insects12070591

Academic Editors: Monica Poelchau and Surya Saha

Received: 28 May 2021
Accepted: 25 June 2021
Published: 29 June 2021

1. Introduction

Significant resources have been invested in sequencing insect genomes over the last decade, with over 768 species being fully or partially sequenced [1]. For instance, the i5k project was initiated to organize the sequencing and analysis of as many as 5000 arthropod genomes [2]. As valuable as these genome sequences are, however, sequencing a genome by itself is merely a beginning—the first step in a transformative process that builds on new information to generate fresh insights. Genomes are of limited value without comprehensive annotation: in addition to the DNA sequence itself, it is necessary to attach biological information to the genome, including not only the locations and identities of genes but also of non-coding regulatory elements. While gene annotation typically follows fairly quickly after assembly, regulatory annotation can be a long time in coming, if at all, and even then typically consists of putative locations only, without descriptions of function.

This lack of regulatory annotation is unfortunate, as regulatory sequences are the building blocks of transcriptional regulatory networks and essential for mediating both development and homeostasis [3,4]. Genes within a regulatory network need to be transcribed at the right time and in the right cells. This regulation is the result of the interaction of transcription factors (TFs) with specific *cis*-regulatory modules (CRMs, including but not limited to "enhancers") that are frequently organized in a modular fashion to regulate specific spatiotemporal aspects of the expression of nearby genes. CRMs are typically a few hundred base pairs in length and are located upstream, downstream, or even within intronic regions of their target regulated gene, often at a considerable distance from the promoter.

One reason why regulatory annotation so often lags behind genome sequencing is that, historically, finding regulatory elements in the genome has been challenging even in well-studied model organisms because of their distant positions from target genes, the absence of a clear universal biochemical CRM marker, and the cell-type specificity of CRM activity [5–8]. For non-model organisms, where only limited functional genomic data tend to be available, regulatory annotation is even more difficult. In an effort to ameliorate this situation, we have developed two tools that facilitate CRM identification, in particular with respect to insect regulatory genomics. REDfly, the the Regulatory Element Database for *Drosophila* and other insects, is a comprehensive knowledge base of published insect CRMs. REDfly contains more than 25,000 experimentally validated *Drosophila melanogaster* CRMs associated with over 1700 genes, along with their sequences and the expression patterns for which they are responsible, accompanied by a growing number of CRMs identified in other insects [9]. SCRMshaw is a computational approach to Supervised *cis*-Regulatory Module discovery that can locate CRMs responsible for directing specific patterns of gene expression in a rapid fashion, with minimal required input [10–12]. Used jointly, these two tools have enabled us to identify CRMs across a range of insect species spanning over 345 million years of evolutionary divergence. In this paper, we briefly review current approaches to CRM discovery, and then show how REDfly and SCRMshaw together constitute a powerful platform for insect CRM discovery and regulatory genome annotation.

2. Empirical Approaches to CRM Discovery

2.1. Reporter Gene Assays

Traditionally, CRMs have been identified through empirical testing of genomic sequence fragments in reporter gene assays. This is a time-consuming and resource-intensive approach, as a CRM can sometimes reside hundreds of thousands of base pairs away from the gene that it regulates. Moreover, in vivo reporter gene assays are not well suited for genome-wide annotation, as they require the generation of a great many transgenic lines to be able to survey a sufficient length of the genome, many of which will provide negative results. More recently, with the availability of fully sequenced genomes and next-generation sequencing methods, high-throughput methods have been developed to functionally assay putative CRMs on a genomic scale (e.g., [13–15]). STARR-seq, which elegantly converts CRMs into their own reporters by cloning them downstream of a core promoter and sequencing the output, is one increasingly popular method [14,16–20]. Although reporter-based methods have long been viewed as a gold standard, due to the fact that they provide a direct functional readout of regulatory activity, there is growing recognition that these methods can lead to both false-positive and false-negative results [7,8,21]. However, the overall accuracy of reporter gene assays is believed to be high, and these remain the most definitive assays for regulatory function.

2.2. ChIP-Based Assays

CRM activity is strictly dependent on the binding of transcription factors, which makes it possible to use genome-wide methods that determine in vivo transcription factor binding sites for the prediction of active CRMs. One common method for genome-wide CRM discovery has been chromatin immunoprecipitation, followed by deep sequencing (ChIP–seq) [22,23]. ChIP-seq not only can be used to identify binding sites for specific TFs,

but it can also be used to identify the in vivo binding sites of transcriptional coactivators present at large numbers of CRMs, such as the acetyltransferase p300. Coactivators do not bind to DNA directly but are recruited by TFs, and carry out various biochemical activities, ultimately leading to activation or repression. The advantage of this approach is that one does not need to know sets of relevant transcription factors a priori; the shortcoming is that focusing on a widely deployed coactivator does not allow for the preferential discovery of CRMs active in a specific tissue. Genome-wide chromatin profiling can also be obtained through ChIP-seq. For that purpose, antibodies against specific chromatin post-translational modifications are used. For example, histone H3, with its lysine at position 27 acetylated (H3K27Ac), shows a high correlation with active CRMs, whereas histone H3, with its K27 trimethylated (H3 K27me3), is indicative of inactive regions [24,25]. However, no single marker appears to universally distinguish CRMs.

Increased chromatin accessibility is also critical to facilitate precise gene regulation. Active CRMs reside in regions of open (nucleosome-depleted) chromatin, which can be identified on a genome-wide scale through a variety of methods. In particular, the formaldehyde-assisted identification of regulatory elements (FAIRE) [26–28], the assay for transposase-accessible chromatin using sequencing (ATAC-seq) [29], and DNase-seq [30] are all used to identify open chromatin regions across the genome. Improvements in single-cell technologies have led to the ability to measure chromatin accessibility using single-cell ATAC-seq (scATAC-seq), a potential game-changer in the ability to undertake cell-type-specific CRM discovery without needing to obtain large numbers of purified cells of uniform type [31–37].

2.3. Limitations

None of the abovementioned empirical methods are without limitations. They remain costly (compared to in silico approaches); can be difficult to validate, depending on the availability of biological resources such as cell lines, antibodies, and tissue samples, or the existence of relevant technologies, such as transgenesis; and carry false-positive and false-negative rates that can be surprisingly high—false-positive rates range as high as 40% for some ChIP-based methods [38–40] and from 10–20% for some ATAC-seq studies [31,32]. Moreover, CRMs may be functional only in certain cell types or under specific conditions, and thus can only be identified when assays include those cells or conditions. The features used by the abovementioned empirical approaches must, therefore, be assessed in multiple tissues over many developmental stages and/or under varying environmental conditions, in order to achieve comprehensive CRM discovery.

3. Computational Approaches

As part of the contemporary arsenal of methods for CRM discovery, computational approaches have proven to be an important complement to experimental ones. Computational CRM discovery has several potential benefits, including a low cost, rapid results, and no requirement for access to expensive and/or limiting biological resources and assays. This is of particular benefit when working with non-model organisms, for which there may be a genome sequence but frequently not extensive other resources. However, many current computational approaches still rely on experimental data for either training or as input (see below), which often negates these advantages.

3.1. Supervised Machine Learning

Recently, genome-wide computational CRM prediction has gained prominence with supervised machine-learning (ML) algorithms that are trained using one or more features from known CRMs. Features can consist of the DNA sequence itself, but frequently also include experimentally derived epigenetic information, such as histone post-translational modifications, transcription factor binding, and chromatin accessibility. In supervised ML, a classification algorithm is trained to distinguish between labeled positive and negative training examples (e.g., CRMs vs. non-CRMs) based on the features of these examples. The

trained classifier is then used to predict the labels for uncharacterized input (e.g., unseen genomic regions). Among the ML approaches that have been used for CRM discovery are support vector machines (SVMs) [41,42], artificial neural networks (ANNs) [43], decision trees (DTs) [44], random forests (RFs) [45], probabilistic models (PGMs) [46,47], and, more recently, deep learning (e.g., [48–52]).

The commonly used SVM seeks to find a hyperplane in an N-dimensional space (where N equals the number of input features) that distinctly separates the data points. SVMs work very effectively when there is a clear margin of distinction between classes, but suffer when the data set is big, noisy, and/or overlapping, and they are prone to over-fitting.

Random forest classifiers, on the other hand, are an ensemble method that has become a popular machine-learning technique due to its ability to run efficiently on large datasets without over-fitting, and to deal with unbalanced and missing data [45]. Usefully, random forests can also indicate the relative importance of individual input features.

Deep learning algorithms are particularly suitable for dealing with large, high-throughput data sets. They utilize a hierarchical assembly of ANNs—nodes connected in a web acting like neurons in the brain, where connections between nodes are strengthened during training if they lead to successful outcomes—to carry out the process of machine learning. Due to their inherent non-linearity and high-level representation of features, algorithms using deep learning often outperform other ML-based methods in predicting CRMs.

3.2. Chromatin and Epigenetic Features

Examples of supervised learning tools based on chromatin features include chromatin signature identification by an artificial neural network (CSI-ANN) [43], ChromagenSVM [41], random forest-based enhancer identification using chromatin states (RFECS) [45], and deep learning for identifying *cis*-regulatory elements (DECRES) [49]. CSI-ANN uses an artificial neural network model to predict CRMs based on histone methylation and acetylation signatures, whereas ChromagenSVM employs a genetic algorithm to choose an optimum combination of histone epigenetic marks for use in an SVM-based classifier. Taking advantage of the strong feature-selection capabilities of random forests, Rajagopal et al. [45] evaluated the importance of different histone modifications using RFECS and an extended panel of histone-ChIP data, to conclude that a combination of three histone modifications, H3K4me1, H3K4me2, and H3K4me3, gave the strongest performance. However, almost as good a performance was achieved when substituting the more commonly assessed H3K27ac for H3K4me2. When all three of these aforementioned chromatin-based tools were applied to the same histone modification datasets in $CD4^+$ T-cells, RFECS achieved the highest validation rate (70% vs. 57% for ChromaGenSVM and 51% for CSI-ANN) and lowest misclassification rate (7% vs. 27% and 35% as compared to ChromagenSVM and CSI-ANN, respectively). While in relative terms RFECS, therefore, gives the strongest performance, the actual validation rates should be viewed with caution, as the study used an extremely generous criterion for true positive predictions, defined as falling with $\pm$2.5 kb of a set of themselves non-dispositive CRM markers, including DNaseI hypersensitive sites and the binding of several TFs and coactivators [49]. The deep-learning approach, DECRES, uses a comprehensive feature set integrating histone modification, TF binding, DNase-seq, FAIRE-seq, and ChIA-PET data from the ENCODE project [53], along with the transcriptionally active enhancers and promoters cataloged by the FANTOM project [54], to predict CRMs. When evaluated against two empirically determined CRM datasets from K562 cells, DECRES displayed high sensitivity (predicting 65% and 98% of validated CRMs from the two sets, respectively) but also a high false-positive rate (predicting 53% and 92% of non-CRMs) [49].

3.3. Sequence Features

Although substantial amounts of cell-type-specific epigenetic data exist for human and mouse such data are often much more limited for other organisms, and reliance on extensive experimental data nullifies many of the advantages of having a computational

approach in the first place. Therefore, a number of methods have explored using sequence features alone as a basis for CRM prediction. Common features used in sequence-based approaches are DNA subsequence composition (e.g., kmers) or TF binding site motifs, although additional features such as G + C and A + T frequencies or CpG island length have been used as well [42]. Kmer-based methods have often been coupled with SVM classifiers, e.g., "kmer-support vector machine" (kmer-SVM) [55]. Kmer-svm can accurately predict CRMs (defined by the binding of the p300 co-activator) from genomic sequence alone and can also discriminate CRMs from non-functional elements with high accuracy in a cross-validation framework. This approach was subsequently improved by "gapped kmer SVM" (gkm-SVM), which takes kmers with gaps into consideration [56].

Deep-learning approaches have been increasing in popularity as the underlying methods become more mature and as advances in computing power make them more feasible. DECRES (introduced above) can be trained and applied using only sequence features, although its performance is then significantly reduced, compared to training with a full range of experimental chromatin-level data [49]. BiRen [52] leverages the deep learning power of convolutional neural networks (CNNs) and bidirectional recurrent neural networks (BRNNs) to predict CRMs using DNA sequences alone as input, although chromatin and histone modification data are used during training. Although this approach exhibited high performance in a cross-validation setting, validation based on overlap with genomic features suggestive of CRMs was relatively weak: association with DNaseI hypersensitive sites, histone H3K27ac, and high-occupancy target for TF ChIP (HOT) regions was only 55%, 40% and 23%, respectively. Nevertheless, BiRen outperformed competing methods also based solely on sequence-feature input [42,55] in head-to-head comparisons, showing the promise of contemporary deep-learning approaches for CRM discovery.

Unsupervised machine-learning methods have suggested the presence of different CRM classes, such as "strong" and "weak" enhancers, based on histone modifications and other ChIP-derived data [46,47]. Although reporter-gene-based validation has called these specific designations into question [13], the sequences themselves are independently discoverable using supervised methods based solely on sequence features. iEnhancer-EL [57] uses a combination of sequence features to build a two-layer ensemble model formed from 16 individual SVM classifiers. In testing, it was shown to perform well at both CRM detection and at stratifying the identified CRMs by their ChromHMM-defined [46] "weak" and "strong" designations, with a measured accuracy of ~75% for the CRM discovery task. This was modestly improved upon (77% accuracy on the same data set) by Nguyen et al. [58], whose iEnhancer-ECNN replaces the ensemble SVM model with an ensemble CNN model instead, using a combination of one-hot encoding and kmers as the sequence input. Another recently proposed method, iEnhancer-5Step, makes use of a word-embedding approach borrowed from natural language programming along with an SVM classifier [59]. This model showed further improvements—an accuracy of 79%—on the same data set analyzed by the other "iEnhancer" methods. However, it is important to note that all of these methods have only been tested using a pre-classified set, with equal numbers of positive and negative sequences (as defined by ChromHMM). This raises significant questions as to what their performance would look like in a "real-world" test using a complete genome, where negative sequences significantly outnumber positive ones, and if validated against a set of CRMs defined using orthogonal criteria.

4. Cross-Species or Non-Model Insect CRM Discovery

CRM discovery in non-traditional insect models presents a particular challenge. Researchers studying these insects often lack the backing of large well-established communities, able to dedicate "big science" resources to gathering the necessary extensive genomic data, and many methods, in particular transgenesis, are not well-established outside of the primary research organisms. Recent advances in technologies that appear to be broadly applicable, such as CRISPR-Cas9 genome manipulation, as well as rapidly declining costs for accessing sequencing-based approaches such as ATAC-seq, are helping to level the

playing field [29,60]. Single-cell methods, especially scATAC-seq, show great promise as they should help to get around the problem of needing to purify large numbers of individual cell types in order to obtain chromatin profiles [31,32].

Computational methods that can be rapidly and inexpensively applied are, in principle, an attractive solution for obtaining a quick regulatory annotation of newly sequenced insect genomes. However, as demonstrated in the preceding discussion, most effective computational methods still require input data, either in the form of training data using known CRMs, or chromatin-level genomic assays. Paradoxically, therefore, computational CRM discovery appears to be dependent on just the sorts of empirical studies that it is intended to bypass.

A potential way past this dilemma is to leverage data from one species to train a computational algorithm to search the genome of a second species [48,61–63]. Such an approach has two basic requirements. First, there must be sufficient similarity in sequence-level CRM properties among species for a cross-species approach to be feasible, as the only input data for the second species may be its genome sequence. This requirement raises a biological question: even when there is a lack of obvious sequence similarity, do functionally related CRMs contain shared features? A growing body of literature suggests that this is the case, at least for CRMs regulating genes involved in core developmental processes. For instance, the similarity of the co-occurrence of sequence patterns has been used to make use of known *Drosophila melanogaster* CRMs to identify "orthologous" CRMs in the distantly related drosophilid and sepsid fly species [61]. The Capra lab has used two machine-learning frameworks (SVMs and CNNs) to distinguish CRMs from the genomic background, based on DNA sequence patterns, and models trained to predict CRMs from one species also accurately identify CRMs in the same cellular context in other species—from humans to opossums [48]. While these sets of species maintain a reasonable amount of alignable sequence, cross-species CRM prediction has also been demonstrated across more highly sequence-diverged species pairs. Minnoye et al. [63] designed a multi-class neural network-based method, DeepMEL, that when trained on human melanoma ATAC-seq data successfully predicted enhancers for two related but distinct cell types across six different species (human, dog, horse, pig, mouse, zebrafish); the latter pairing begins to approach the level of divergence observed in family-level comparisons among the holometabola [64]. Transcription factor binding site clustering, based on *Drosophila melanogaster* CRMs, has been used to discover CRMs in other holometabolous insects [65–68], and the SCRMshaw algorithm (described more fully below) has used *Drosophila* data to successfully identify CRMs in species as distantly diverged as the Hemiptera e.g., [62,69,70] (H.A. and M.S.H., unpublished data), based on statistical similarities in subsequence (kmer) counts among the CRMs. Detailed analysis of such distant but related CRMs has revealed the presence of common sets of transcription factor binding sites, and even potential enhancer grammars (conserved arrangements of binding sites), shared features that blend into the noise of whole-genome analyses but that become possible to detect once CRM locations have been defined [62]. So far, most success in cross-species CRM discovery where genomes are not alignable has been seen with CRMs that function in well-conserved developmental pathways, and it remains to be determined how widespread these deep sequence-level homologies are with respect to less fundamental regulatory networks.

The second requirement for cross-species CRM discovery is that there must be an extensive enough body of data of a suitable type in the training species. Sets of known CRMs are ideal for this purpose, as this is purely sequence-based data that can be readily applied to other sequenced genomes. Here insects have a significant advantage, due to the extensive set of known CRMs available for *Drosophila melanogaster*, many of which are functionally validated. Moreover, the majority of these have been carefully curated into the insect-specific REDfly database, a comprehensive source of regulatory data unique among the metazoa [9]. In the remaining sections of the paper, we focus on how REDfly and SCRMshaw can be used together for CRM discovery across multiple insect species as a powerful platform to facilitate the field of insect regulatory genomics.

5. REDfly and SCRMshaw: Powerful Tools for Insect Regulatory Genomics

5.1. REDfly

REDfly is a one-stop curated knowledgebase for insect *cis*-regulatory data [9]. Historically focused on experimentally verified *Drosophila melanogaster* CRMs, REDfly has grown over its 15 years of existence to include *Drosophila* transcription factor binding sites, CRMs identified from epigenetic profiles and computational prediction, and, most recently, CRMs from an increasing number of other insects, including important disease vectors such as mosquitoes. (Currently, three insect species in addition to *D. melanogaster* have been incorporated—*Anopheles gambiae, Aedes aegypti*, and *Tribolium castaneum*—with more on the way.) To date, REDfly has over 25,500 *D. melanogaster* CRMs (60% from in vivo reporter genes, most of the remainder from cell-culture assays) associated with 12–15% of protein-coding genes, and ~2700 defined transcription factor binding sites (TFBSs). These data are based on more than 1200 curated publications. The core REDfly *Drosophila* CRM annotations are provided to FlyBase, making *Drosophila* the only model organism whose genome annotation provides comprehensive coverage of validated CRMs, enabling direct integration with other *Drosophila* genomic and genetic data.

One strength of REDfly is the extensive detail it provides about the CRMs it curates (Figure 1). Regulatory activity is described using the *Drosophila* anatomy and development ontologies [71], which allows for retrieval of tissue-specific and stage-specific CRM datasets at multiple degrees of granularity. Terms from the Gene Ontology [72,73] are incorporated to annotate regulatory elements that respond to specific physiological or environmental cues (e.g., wound-healing, hypoxia). CRMs can be filtered by size, genomic location, position relative to target genes (e.g., upstream, downstream, intronic), and sex-specificity. Overlapping regions between multiple CRMs are automatically calculated to suggest potential minimal CRM sequences and their regulatory activity.

As the most detailed existing platform offering regulatory element annotation for any animal, REDfly serves as an important platform for supporting both empirical and computational research. REDfly has contributed to numerous studies in multiple areas relating to non-*Drosophila*, as well as *Drosophila* systems including studies of basic CRM biology (e.g., [74–83]) and interpretation of genomic data such as TF binding studies (e.g., [84–86]), studies of insulators [87,88], of chromosome domains and "states" (e.g., [89–92]), of 3D-chromatin conformation [93,94], and of ncRNA and eRNA expression [95,96]. REDfly data have facilitated the validation of ATAC-seq approaches (e.g., [97,98]), have been used to establish TF-CRM associations for the study of gene regulatory networks [74,99–104], and have enabled studies of CRM evolution and TFBS turnover (e.g., [104–111]).

REDfly has also played a dramatic role in developing methods for computational CRM discovery. Its extensive collection of experimentally verified CRMs provides a ready source of validation data for assessing CRM predictions and for comparing among methods [21,112–116]. Perhaps more importantly, REDfly's advanced search and filtering features make it an unmatched source for compiling training data for machine-learning approaches [10,11,61,117]. As we review in the following section, we have used REDfly to develop nearly 50 individual training sets, spanning numerous *Drosophila* tissues and time points, for use in conjunction with our SCRMshaw algorithm. This has enabled considerable new CRM discovery in *Drosophila* but, more excitingly, also allows for cross-species identification of CRMs in a wide range of insect species.

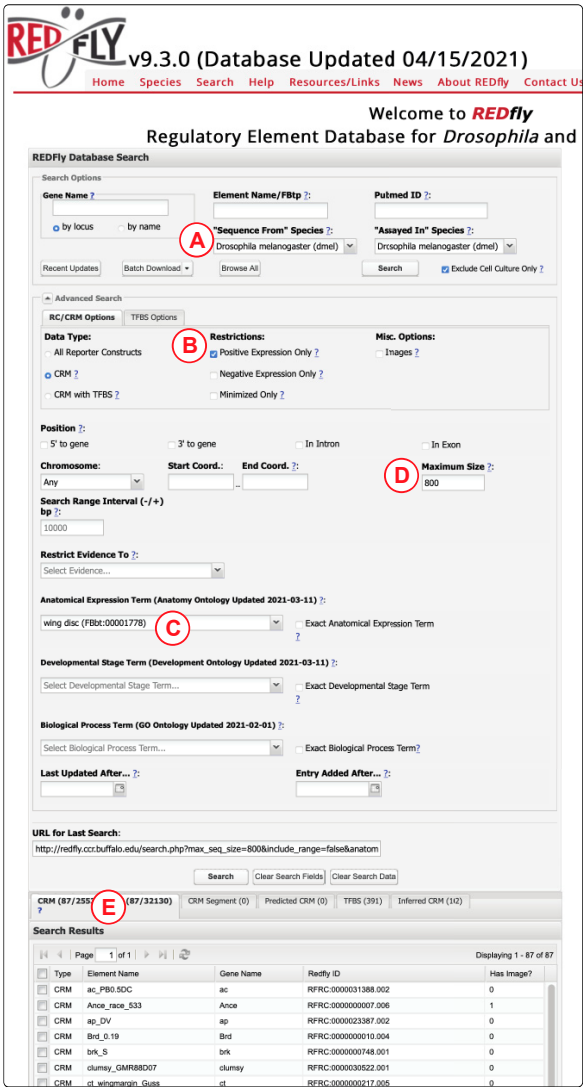

Figure 1. The REDfly search interface. Researchers can make use of REDfly's comprehensive search capabilities to assemble sets of CRMs with specific properties. In this example, for instance, over 25,000 "CRM" records will be searched for (**A**) sequences belonging to the *Drosophila melanogaster* genome that (**B**) positively regulate gene expression in the (**C**) wing imaginal disc, and that are (**D**) no greater than 800 bp in length. Results are listed in the Results Table (**E**). Clicking on an individual result opens a Detailed Results window with extensive further information, or multiple records can be selected using the checkboxes for download in a variety of formats.

5.2. SCRMshaw

The CRM training data made available by REDfly enabled us, in collaboration with Saurabh Sinha's laboratory at the University of Illinois, to develop SCRMshaw (for supervised *cis*-regulatory module discovery), a highly effective method for computational CRM discovery [10–12]. Other than the possession of a moderately sized training set of known CRMs—we have had success with fewer than ten CRMs, although 20–30 is preferred—the only input SCRMshaw requires is an annotated genome. As no other genomic data are required (e.g., no binding site data, chromatin conformation or state data, or histone modification data), SCRMshaw is ideal for use with newly sequenced and less-studied insect species.

SCRMshaw (Figure 2) uses a training set composed of known CRMs, defined by a common functional characterization (e.g., "nervous system", "midgut") 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-CRM "background" sequences in a machine-learning framework. The kmers likely serve as proxies for the unknown TFBSs, but TFBSs 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 CRMs. SCRMshaw has proven to be remarkably effective: when SCRMshaw predictions are tested empirically using reporter gene assays, success rates have averaged ~80%, with some training sets yielding over 90% true positives [10,11,62,69,118].

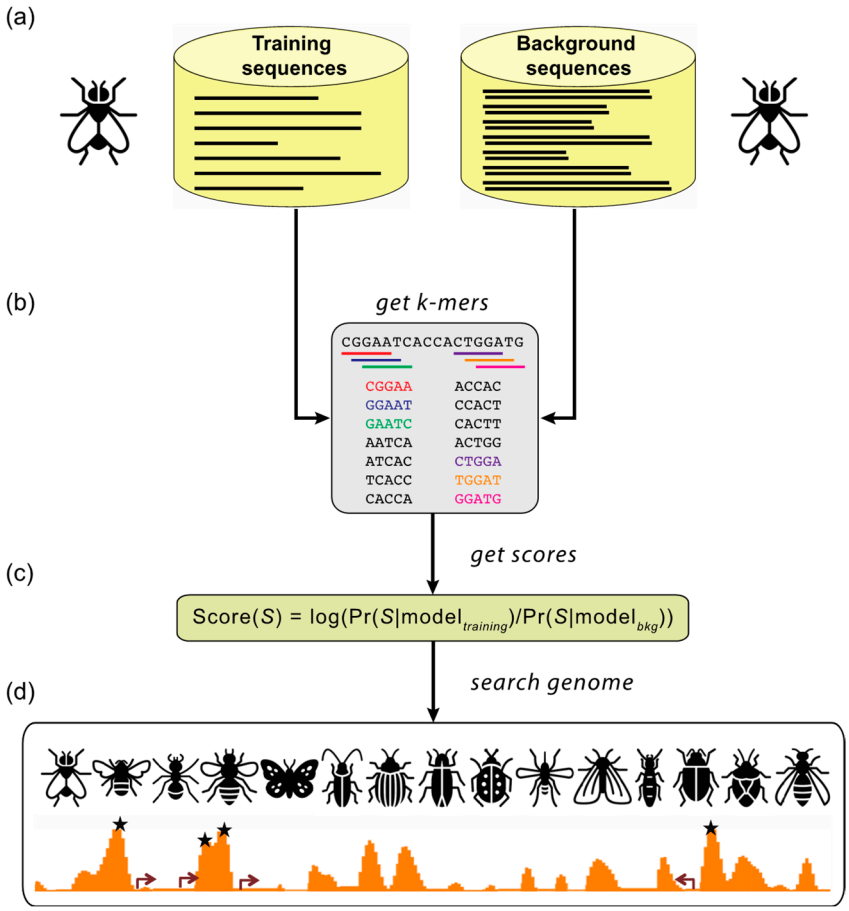

Figure 2. Supervised motif-blind CRM discovery (SCRMshaw). (**a**) SCRMshaw uses a training set of known *Drosophila melanogaster* CRMs ("training sequences"), drawn from REDfly, that are defined by common functional characterization, and a 10-fold larger background set of similarly sized non-CRM sequences ("background sequences"). (**b**) The short DNA subsequence (*kmer*) count distributions of these sequences are then used to train a statistical model. The trained model (**c**) is used to score overlapping windows in the "target genome"; to date, we have successfully used multiple different species from the Holometabola and several Hemiptera. (**d**) High-scoring regions are predicted to be functional regulatory sequences (asterisks). Figure adapted from [69]. Insect images downloaded from TheNounProject.com (accessed on 28 April 2021) set "Bugs" by Georgiana Ionescu, under the CC-BY license.

5.3. Cross-Species Prediction

As noted above, SCRMshaw's true power is its ability to predict CRMs across species. Kazemian et al. [62] used SCRMshaw with *Drosophila* CRM training data to discover CRMs across the entire ~345 Mya range of the holometabolous insects. By using the same methods and training sets as previously used for within-species CRM discovery in *D. melanogaster* [10], and instead searching the genomes of *Anopheles gambiae, Aedes aegypti, Tribolium castaneum, Apis mellifera, and Nasonia vitripennis*, SCRMshaw successfully

predicted CRMs in a cross-species fashion with an approximately 75% prediction success rate, based on reporter gene assays in xenotransgenic flies [62,69,118]. Direct testing of a predicted *Tribolium* CRM in transgenic *Tribolium* confirmed that SCRMshaw can find bona fide CRMs cross-species [70]. Preliminary data using hemipteran genomes suggest at least some ability to predict CRMs in this even more diverged insect order as well (H.A. and M.S.H., unpublished data).

5.4. Training Set Improvement

As with any machine-learning method, a key factor affecting SCRMshaw's importance is the quality of the training data. Interestingly, our testing has shown that even training sets made up of randomly grouped but validated CRMs outperform groups of random non-coding, non-regulatory sequences when used for SCRMshaw training [10,21]. This is most likely due to the increased presence of binding sequences for common transcription factor families (e.g., E-boxes, homeodomain binding sites) in the true CRM sequences, and may account for the positive predictive performance obtained even with our least effective training sets. However, increasing the cohesiveness of the training sets, so that they fully represent groups of CRMs with common activity profiles, should improve predictive performance.

We have constructed over 48 training sets spanning a broad range of gene expression patterns across tissues and stages, using the most current available CRM data in the REDfly database. For the most part, these new sets are generated automatically by filtering REDfly CRMs based on anatomical classifications derived directly from the *Drosophila* anatomy ontology. Because the anatomy terms are not all temporally specific—while some terms distinguish between embryonic, larval, and adult stages, others do not—some of the training sets are likely too broadly constituted, leading to reduced performance. Taking advantage of updated temporal staging data in REDfly [9], as well as performing manual refinement to create more specific groupings of known CRM-driven expression patterns, should help to improve training set quality. We recently developed a training set evaluation pipeline, *pCRMeval*, that enables unbiased assessments of SCRMshaw performance on a training set by training set basis, to aid in this process [21].

What can be done when a cohesive enough set of training CRMs of sufficient number is not available? In such cases, an iterative search-and-validation strategy has proven very effective (Figure 3). For example, in collaboration with Thomas Williams' laboratory at the University of Dayton, we used SCRMshaw on a small training set of just seven CRMs that drive gene expression in the *Drosophila* adult abdomen, in a first round of SCRMshaw, to identify CRMs potentially involved in abdominal pigmentation. Empirical testing of 18 of these predictions revealed 10 new CRMs regulating the desired specific expression pattern (55%); an additional three appeared to be bona fide CRMs but with a different expression profile (for a total of 13/18 or 72% validating as CRMs). The ten CRMs driving the "correct" expression pattern were then combined with the original seven members of the training set to generate a new, 2.5-fold expanded training set for a second round of SCRMshaw prediction. Notably, the top prediction results from this second round did not include seven of the eight sequences previously found to be false positives by empirical testing—including true CRMs with non-targeted expression profiles—but still contained all ten of the previous true positives. Empirical validation for the second round of predictions confirms this improved performance: out of 21 CRMs tested, 20 were functional CRMs (95%) of which 17 (81%) had the expected pattern of expression (T. Williams, personnal communication). These results demonstrate that iterative approaches can serve to augment weak training sets to improve true-positive: false-positive ratios and underscore the importance of having a well-constructed training set. Although so far we have used this approach exclusively for *Drosophila* CRM prediction, iterative searching should also be useful for increasing the success of our cross-species predictions. Validated cross-species sequences can be added to our training sets in the same way that we have added validated *Drosophila* CRMs, with similarly improved results. Indeed, we expect that we may see even more

dramatic improvement in some cases, as by adding in additional sequences from the species being searched, we will be increasing the number of same-species sequences, moving the SCRMshaw search closer to a same-species search.

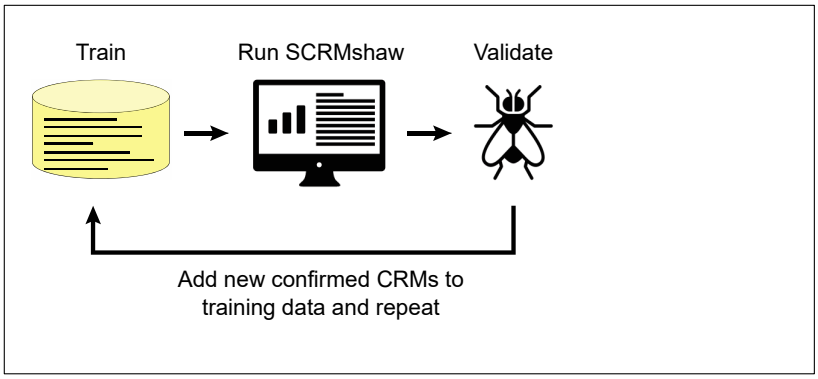

Figure 3. Iterative search and validation. Sequences validating as true CRMs from an initial SCRMshaw search can be added back to the training set, and the new, enlarged training set used for a subsequent round of predictions. This strategy can effectively compensate for an initially weak training set. Images downloaded from TheNounProject.com (accessed on 28 April 2021), "Fly" by Georgiana Ionescu, and "analytics" by Wilson Joseph, under the CC-BY license.

5.5. Limitations

While many insect species are being sequenced, the genome assemblies and annotations are of varying quality, ranging from extremely well-assembled genomes (e.g., *Aedes aegypti* with a scaffold N50 size of 409,777,670 bp) to very poorly assembled genomes (Yellow Sally stonefly with a scaffold N50 of only 457 bp) [119,120]. How do these factors affect CRM prediction using SCRMshaw? Since, by default, SCRMshaw searches the genome in 500 bp windows, we reasoned that assembly is not likely to be a major limiting factor for any but the most poorly assembled genomes. By simulating different degrees of genome assembly for the *Drosophila* genome, and comparing the results to those obtained with the fully assembled genome, we determined that this is indeed the case: SCRMshaw maintains its strong predictive power when applied to increasingly less well-assembled genomes, with only a minor drop-off in sensitivity—less than 15% on average—and a negligible increase in the false-positive rate [21]. Assembly quality does not, therefore, appear to present a significant barrier to successful CRM prediction using SCRMshaw. Moreover, with long-read sequencing technologies becoming more prevalent, less fractured assemblies are now frequently available [121].

The second factor to consider is genome annotation, which can lag considerably behind sequencing and assembly: only 40% of all i5k-sequenced species currently have an accompanying predicted gene set [120]. Because SCRMshaw looks at kmer-level patterns, we typically exclude coding sequences from the analysis out of concern that the inherent constraints on coding sequences (e.g., codon biases and the limited number of valid codon triplets) will affect the SCRMshaw scoring. To test how important this is, we have compared the results of running SCRMshaw on the *Drosophila* genome with and without exons masked (H.A. and B. Yuen, unpublished data). These tests show that there is only minimal overlap in the top predictions between the masked and unmasked versions, confirming the importance of excluding coding sequences from the analysis. Therefore, having a good draft gene annotation is an important requirement for ensuring optimal SCRMshaw predictions. Fortunately, common genome annotation pipelines such as Maker2 [122] and Braker2 [123] are generally effective at detecting protein-coding regions, enabling adequate initial annotations to be generated.

6. Integrating Computational and Experimental Approaches

With empirical methods such as ATAC-seq becoming increasingly more accessible and affordable, it is fair to ask the question, what is the future of computational CRM prediction for insect genomes? We believe that the combination of REDfly and SCRMshaw retains several powerful advantages. For one, the approach remains both rapid and cost-effective. With no required access to biological samples, negligible expense, and only a few days' time—much of which is hands-off—a reasonable first-pass annotation of the regulatory genome is possible for any recently sequenced insects within the holometabola, and probably their nearest neighboring families. This may prove especially useful for acquiring a broad sampling of CRM data for evolutionary studies; because SCRMshaw uses the same training data to search each species, the chances of discovering homologous (or functionally similar) CRMs for orthologous genes are high. In the long run, however, the greatest benefit is likely to be seen from combining computational prediction with other forms of genomic CRM discovery, where the in silico results can help to sharpen and refine empirical data. For instance, computational CRM discovery with subsequent alignment across groups of moderately closely related species can pinpoint important sequence motifs, and provide insights into possible enhancer grammar [62]. When used in conjunction with open-chromatin assays, SCRMshaw can help distinguish CRMs from non-CRM open regions, and, for whole-animal assays, is able to help home in on the most relevant tissue-specific CRMs. When used with single-cell methods, the fact that SCRMshaw predictions are based on tissue-specific training sets will help to assign identities to individual cell types to better interpret the single-cell results, and may help to develop improved methods for distinguishing sets of CRMs that are most closely related (i.e., integrate similar transcription factor inputs) from those that use different input strategies to achieve the same regulatory output (e.g., [124,125]). Chromatin profiling and next-generation sequencing assays each contain various biases (e.g., [126–129]). Those that are known can be corrected for to a certain extent, whereas other biases may not yet be well understood. Combining such assays with SCRMshaw, a wholly orthogonal method, should help in weeding out false-positive results from both types of assays, leading to more accurate CRM prediction overall.

7. Conclusions

The rapid growth in sequenced insect genomes requires the development of equally rapid and economical means for annotating the regulatory components of these genomes. Although many methods for CRM identification rely on extensive empirical data, a subset of computational approaches, including SCRMshaw, function effectively using sequence features only as input for both training and discovery. We discussed here how SCRMshaw, using the wealth of *Drosophila* regulatory data curated by REDfly, can be used cross-species to produce reasonable first-draft annotations of regulatory sequences throughout the holometabola and likely beyond. We focused on SCRMshaw because that is where data demonstrating good cross-species performance currently exist; however, it is likely that other sequence-based CRM discovery methods will similarly be capable of cross-species discovery. Our experience suggests that the most critical factor is the quality and cohesiveness of the training data. As the CRM data in REDfly become more finely annotated for developmental timing and cellular identities, it should be possible to generate improved training data and subsequently, an ever more accurate prediction of CRM locations in newly sequenced genomes. Moreover, as data for CRMs accumulate in more phylogenetically basal species, the ability to push prediction success into further diverged insect orders may become feasible.

Computational pipelines are able to provide rapid first-pass gene annotations of newly sequenced genomes, but require experimental follow-up in order to refine gene models for a final, accurate, finished annotation [130]. In the same way, REDfly and SCRMshaw can be paired as a powerful combination to generate initial, albeit imperfect, regulatory annotations for insect genomes, to be further refined by subsequent empirical CRM identification.

8. URLs

REDfly is freely accessible to the public at http://redfly.ccr.buffalo.edu.

SCRMshaw software and associated useful utility programs can be downloaded from the Halfon lab GitHub site at https://github.com/HalfonLab. Protocols for using SCRMshaw can be found in references [12] and [21].

Author Contributions: Writing—original draft, H.A.; writing—review and editing, M.S.H.; funding acquisition, M.S.H. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by U.S. Department of Agriculture grant 2018-08230, National Institutes of Health grant R01 GM114067 and National Science Foundation grant DBI-1758252.

Acknowledgments: We thank Thomas Williams for sharing data in advance of publication, Brandon Yuen for help with SCRMshaw testing, and Isabella Schember for comments on the manuscript. REDfly and SCRMshaw both make use of the resources of the University at Buffalo Center for Computational Research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ANN	Artificial neural network
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
auROC	Area under the receiver operating characteristic
BRNN	Bidirectional recurrent neural networks
ChIA-PET	Chromatin interaction analysis with paired-end tag
ChIP	Chromatin immuno precipitation
CNN	Convolutional neural network
CRM	*cis*-regulatory module
CSI-ANN	Chromatin signature identification by artificial neural network
DECRES	Deep learning for identifying *cis*-regulatory elements
ECNN	Ensemble of CNN
ENCODE	Encyclopedia of DNA elements
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FANTOM	Functional annotation of the mammalian genome
Gkm-SVM	Gapped kmer support vector machine
HMM	Hidden Markov model
Kmer-SVM	kmer-support vector machine
ML	Machine learning
REDfly	Regulatory element database for *Drosophila*
RF	Random forest
RFECS	Random forest-based enhancer identification using chromatin states
ROC	Receiver operating characteristic
scATAC-seq	Single cell assay for transposase-accessible chromatin using sequencing
SCRMshaw	Supervised *cis*-regulatory module discovery
STARR-seq	Self-transcribing active regulatory region sequencing
SVM	Support vector machine
TF	Transcription factor
TFBS	Transcription factor binding site

References

1. NCBI Genome Information by Organism. Available online: https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/ (accessed on 25 May 2021).
2. i5K Consortium. The i5K Initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.* **2013**, *104*, 595–600. [CrossRef]
3. Davidson, E.H.; Erwin, D.H. Gene regulatory networks and the evolution of animal body plans. *Science* **2006**, *311*, 796–800. [CrossRef] [PubMed]
4. Carroll, S.B.; Grenier, J.K.; Weatherbee, S.D. *From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design*; Blackwell Science: Malden, MA, USA, 2001.

5. Pennacchio, L.A.; Bickmore, W.; Dean, A.; Nobrega, M.A.; Bejerano, G. Enhancers: Five essential questions. *Nat. Rev. Genet.* **2013**, *14*, 288–295. [CrossRef] [PubMed]
6. Halfon, M.S. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends Genet.* **2019**, *35*, 93–103. [CrossRef] [PubMed]
7. Catarino, R.R.; Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* **2018**, *32*, 202–223. [CrossRef] [PubMed]
8. Gasperini, M.; Tome, J.M.; Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **2020**, *21*, 292–310. [CrossRef]
9. Rivera, J.; Keranen, S.V.E.; Gallo, S.M.; Halfon, M.S. REDfly: The transcriptional regulatory element database for *Drosophila*. *Nucleic Acids Res.* **2019**, *47*, D828–D834. [CrossRef]
10. Kantorovitz, M.R.; Kazemian, M.; Kinston, S.; Miranda-Saavedra, D.; Zhu, Q.; Robinson, G.E.; Göttgens, B.; Halfon, M.S.; Sinha, S. Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Dev. Cell* **2009**, *17*, 568–579. [CrossRef]
11. Kazemian, M.; Zhu, Q.; Halfon, M.S.; Sinha, S. Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res.* **2011**, *39*, 9463–9472. [CrossRef]
12. Kazemian, M.; Halfon, M.S. CRM Discovery Beyond Model Insects. In *Insect Genomics: Methods and Protocols*; Brown, S.J., Pfrender, M.E., Eds.; Springer: New York, NY, USA, 2019; pp. 117–139.
13. Kwasnieski, J.C.; Mogno, I.; Myers, C.A.; Corbo, J.C.; Cohen, B.A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 19498–19503. [CrossRef]
14. Arnold, C.D.; Gerlach, D.; Stelzer, C.; Boryn, L.M.; Rath, M.; Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **2013**, *339*, 1074–1077. [CrossRef]
15. Murtha, M.; Tokcaer-Keskin, Z.; Tang, Z.; Strino, F.; Chen, X.; Wang, Y.; Xi, X.; Basilico, C.; Brown, S.; Bonneau, R.; et al. FIREWACh: High-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat. Methods* **2014**, *11*, 559–565. [CrossRef]
16. Kim, Y.S.; Johnson, G.D.; Seo, J.; Barrera, A.; Cowart, T.N.; Majoros, W.H.; Ochoa, A.; Allen, A.S.; Reddy, T.E. Correcting signal biases and detecting regulatory elements in STARR-seq data. *Genome Res.* **2021**, *31*, 877–889. [CrossRef]
17. Lee, D.; Shi, M.; Moran, J.; Wall, M.; Zhang, J.; Liu, J.; Fitzgerald, D.; Kyono, Y.; Ma, L.; White, K.P.; et al. STARRPeaker: Uniform processing and accurate identification of STARR-seq active regions. *Genome Biol.* **2020**, *21*, 298. [CrossRef]
18. Peng, T.; Zhai, Y.; Atlasi, Y.; Ter Huurne, M.; Marks, H.; Stunnenberg, H.G.; Megchelenbrink, W. STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol.* **2020**, *21*, 243. [CrossRef] [PubMed]
19. Benoit, M. Shooting for the STARRs: A Modified STARR-seq Assay for Rapid Identification and Evaluation of Plant Regulatory Sequences in Tobacco Leaves. *Plant Cell* **2020**, *32*, 2057–2058. [CrossRef]
20. Zhang, J.; Lee, D.; Dhiman, V.; Jiang, P.; Xu, J.; McGillivray, P.; Yang, H.; Liu, J.; Meyerson, W.; Clarke, D.; et al. An integrative ENCODE resource for cancer genomics. *Nat. Commun.* **2020**, *11*, 3696. [CrossRef] [PubMed]
21. Asma, H.; Halfon, M.S. Computational enhancer prediction: Evaluation and improvements. *BMC Bioinform.* **2019**, *20*, 174. [CrossRef]
22. Ghavi-Helm, Y.; Furlong, E.E. Analyzing transcription factor occupancy during embryo development using ChIP-seq. *Methods Mol. Biol.* **2012**, *786*, 229–245. [CrossRef]
23. Park, P.J. ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **2009**, *10*, 669–680. [CrossRef] [PubMed]
24. Bannister, A.J.; Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **2011**, *21*, 381–395. [CrossRef] [PubMed]
25. Bell, O.; Tiwari, V.K.; Thoma, N.H.; Schubeler, D. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* **2011**, *12*, 554–564. [CrossRef] [PubMed]
26. Giresi, P.G.; Kim, J.; McDaniell, R.M.; Iyer, V.R.; Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **2007**, *17*, 877–885. [CrossRef] [PubMed]
27. McKay, D.J. Using Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) to Identify Functional Regulatory DNA in Insect Genomes. *Methods Mol. Biol.* **2019**, *1858*, 89–97. [CrossRef]
28. McKay, D.J.; Lieb, J.D. A common set of DNA regulatory elements shapes Drosophila appendages. *Dev. Cell* **2013**, *27*, 306–318. [CrossRef]
29. Buenrostro, J.D.; Wu, B.; Chang, H.Y.; Greenleaf, W.J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **2015**, *109*, 21.29.21–21.29.29. [CrossRef]
30. Boyle, A.P.; Davis, S.; Shulha, H.P.; Meltzer, P.; Margulies, E.H.; Weng, Z.; Furey, T.S.; Crawford, G.E. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **2008**, *132*, 311–322. [CrossRef]
31. Cusanovich, D.A.; Daza, R.; Adey, A.; Pliner, H.A.; Christiansen, L.; Gunderson, K.L.; Steemers, F.J.; Trapnell, C.; Shendure, J. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **2015**, *348*, 910–914. [CrossRef]
32. Bravo Gonzalez-Blas, C.; Quan, X.J.; Duran-Romana, R.; Taskiran, I.I.; Koldere, D.; Davie, K.; Christiaens, V.; Makhzami, S.; Hulselmans, G.; de Waegeneer, M.; et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Mol. Syst. Biol.* **2020**, *16*, e9438. [CrossRef]
33. Buenrostro, J.D.; Wu, B.; Litzenburger, U.M.; Ruff, D.; Gonzales, M.L.; Snyder, M.P.; Chang, H.Y.; Greenleaf, W.J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **2015**, *523*, 486–490. [CrossRef]

34. Chen, X.; Miragaia, R.J.; Natarajan, K.N.; Teichmann, S.A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **2018**, *9*, 5345. [CrossRef] [PubMed]
35. Mezger, A.; Klemm, S.; Mann, I.; Brower, K.; Mir, A.; Bostick, M.; Farmer, A.; Fordyce, P.; Linnarsson, S.; Greenleaf, W. High-throughput chromatin accessibility profiling at single-cell resolution. *Nat. Commun.* **2018**, *9*, 3647. [CrossRef] [PubMed]
36. Baek, S.; Lee, I. Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1429–1439. [CrossRef]
37. Fiers, M.; Minnoye, L.; Aibar, S.; Bravo Gonzalez-Blas, C.; Kalender Atak, Z.; Aerts, S. Mapping gene regulatory networks from single-cell omics data. *Brief Funct. Genom.* **2018**, *17*, 246–254. [CrossRef]
38. Blow, M.J.; McCulley, D.J.; Li, Z.; Zhang, T.; Akiyama, J.A.; Holt, A.; Plajzer-Frick, I.; Shoukry, M.; Wright, C.; Chen, F.; et al. ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **2010**, *42*, 806–810. [CrossRef] [PubMed]
39. May, D.; Blow, M.J.; Kaplan, T.; McCulley, D.J.; Jensen, B.C.; Akiyama, J.A.; Holt, A.; Plajzer-Frick, I.; Shoukry, M.; Wright, C.; et al. Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* **2011**, *44*, 89–93. [CrossRef] [PubMed]
40. Visel, A.; Blow, M.J.; Li, Z.; Zhang, T.; Akiyama, J.A.; Holt, A.; Plajzer-Frick, I.; Shoukry, M.; Wright, C.; Chen, F.; et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **2009**, *457*, 854–858. [CrossRef] [PubMed]
41. Fernandez, M.; Miranda-Saavedra, D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res.* **2012**, *40*, e77. [CrossRef] [PubMed]
42. Kleftogiannis, D.; Kalnis, P.; Bajic, V.B. DEEP: A general computational framework for predicting enhancers. *Nucleic Acids Res.* **2015**, *43*, e6. [CrossRef]
43. Firpi, H.A.; Ucar, D.; Tan, K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* **2010**, *26*, 1579–1586. [CrossRef]
44. Lu, Y.; Qu, W.; Shan, G.; Zhang, C. DELTA: A Distal Enhancer Locating Tool Based on AdaBoost Algorithm and Shape Features of Chromatin Modifications. *PLoS ONE* **2015**, *10*, e0130622. [CrossRef]
45. Rajagopal, N.; Xie, W.; Li, Y.; Wagner, U.; Wang, W.; Stamatoyannopoulos, J.; Ernst, J.; Kellis, M.; Ren, B. RFECS: A random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.* **2013**, *9*, e1002968. [CrossRef]
46. Ernst, J.; Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods* **2012**, *9*, 215–216. [CrossRef] [PubMed]
47. Hoffman, M.M.; Buske, O.J.; Wang, J.; Weng, Z.; Bilmes, J.A.; Noble, W.S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **2012**, *9*, 473–476. [CrossRef] [PubMed]
48. Chen, L.; Fish, A.E.; Capra, J.A. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput. Biol.* **2018**, *14*, e1006484. [CrossRef]
49. Li, Y.; Shi, W.; Wasserman, W.W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinform.* **2018**, *19*, 202. [CrossRef] [PubMed]
50. Liu, F.; Li, H.; Ren, C.; Bo, X.; Shu, W. PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. *Sci. Rep.* **2016**, *6*, 28517. [CrossRef]
51. Min, X.; Zeng, W.; Chen, S.; Chen, N.; Chen, T.; Jiang, R. Predicting enhancers with deep convolutional neural networks. *BMC Bioinform.* **2017**, *18*, 478. [CrossRef]
52. Yang, B.; Liu, F.; Ren, C.; Ouyang, Z.; Xie, Z.; Bo, X.; Shu, W. BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* **2017**, *33*, 1930–1936. [CrossRef]
53. Encode Project Consortium; Moore, J.E.; Purcaro, M.J.; Pratt, H.E.; Epstein, C.B.; Shoresh, N.; Adrian, J.; Kawli, T.; Davis, C.A.; Dobin, A.; et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **2020**, *583*, 699–710. [CrossRef]
54. Abugessaisa, I.; Ramilowski, J.A.; Lizio, M.; Severin, J.; Hasegawa, A.; Harshbarger, J.; Kondo, A.; Noguchi, S.; Yip, C.W.; Ooi, J.L.C.; et al. FANTOM enters 20th year: Expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res.* **2021**, *49*, D892–D898. [CrossRef]
55. Lee, D.; Karchin, R.; Beer, M.A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **2011**, *21*, 2167–2180. [CrossRef]
56. Ghandi, M.; Lee, D.; Mohammad-Noori, M.; Beer, M.A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **2014**, *10*, e1003711. [CrossRef]
57. Liu, B.; Li, K.; Huang, D.S.; Chou, K.C. iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* **2018**, *34*, 3835–3842. [CrossRef] [PubMed]
58. Nguyen, Q.H.; Nguyen-Vo, T.H.; Le, N.Q.K.; Do, T.T.T.; Rahardja, S.; Nguyen, B.P. iEnhancer-ECNN: Identifying enhancers and their strength using ensembles of convolutional neural networks. *BMC Genom.* **2019**, *20*, 951. [CrossRef]
59. Le, N.Q.K.; Yapp, E.K.Y.; Ho, Q.T.; Nagasundaram, N.; Ou, Y.Y.; Yeh, H.Y. iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* **2019**, *571*, 53–61. [CrossRef] [PubMed]
60. Shukla, A.; Huangfu, D. Decoding the noncoding genome via large-scale CRISPR screens. *Curr. Opin. Genet. Dev.* **2018**, *52*, 70–76. [CrossRef] [PubMed]
61. Arunachalam, M.; Jayasurya, K.; Tomancak, P.; Ohler, U. An alignment-free method to identify candidate orthologous enhancers in multiple *Drosophila* genomes. *Bioinformatics* **2010**, *26*, 2109–2115. [CrossRef]

62. Kazemian, M.; Suryamohan, K.; Chen, J.-Y.; Zhang, Y.; Samee, M.A.H.; Halfon, M.S.; Sinha, S. Evidence for Deep Regulatory Similarities in Early Developmental Programs across Highly Diverged Insects. *Genome Biol. Evol.* **2014**, *6*, 2301–2320. [CrossRef]
63. Minnoye, L.; Taskiran, I.I.; Mauduit, D.; Fazio, M.; Van Aerschot, L.; Hulselmans, G.; Christiaens, V.; Makhzami, S.; Seltenhammer, M.; Karras, P.; et al. Cross-species analysis of enhancer logic using deep learning. *Genome Res.* **2020**, *30*, 1815–1834. [CrossRef]
64. Zdobnov, E.M.; Bork, P. Quantification of insect genome divergence. *Trends Genet.* **2007**, *23*, 16–20. [CrossRef] [PubMed]
65. Cande, J.; Goltsev, Y.; Levine, M.S. Conservation of enhancer location in divergent insects. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 14414–14419. [CrossRef] [PubMed]
66. Erives, A.; Levine, M. Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3851–3856. [CrossRef] [PubMed]
67. Zinzen, R.P.; Cande, J.; Ronshaugen, M.; Papatsenko, D.; Levine, M. Evolution of the ventral midline in insect embryos. *Dev. Cell* **2006**, *11*, 895–902. [CrossRef] [PubMed]
68. Goltsev, Y.; Fuse, N.; Frasch, M.; Zinzen, R.P.; Lanzaro, G.; Levine, M. Evolution of the dorsal-ventral patterning network in the mosquito, *Anopheles gambiae*. *Development* **2007**, *134*, 2415–2424. [CrossRef] [PubMed]
69. Suryamohan, K.; Halfon, M.S. Overview Article: Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdiscip. Rev. Dev. Biol.* **2015**, *4*, 59–84. [CrossRef]
70. Lai, Y.T.; Deem, K.D.; Borras-Castells, F.; Sambrani, N.; Rudolf, H.; Suryamohan, K.; El-Sherif, E.; Halfon, M.S.; McKay, D.J.; Tomoyasu, Y. Enhancer identification and activity evaluation in the red flour beetle, *Tribolium castaneum*. *Development* **2018**, *145*. [CrossRef]
71. Costa, M.; Reeve, S.; Grumbling, G.; Osumi-Sutherland, D. The *Drosophila* anatomy ontology. *J. Biomed. Semant.* **2013**, *4*, 32. [CrossRef]
72. Gene Ontology Consortium. Gene Ontology Consortium: Going forward. *Nucleic Acids Res.* **2015**, *43*, D1049–D1056. [CrossRef]
73. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
74. De Renzis, S.; Elemento, O.; Tavazoie, S.; Wieschaus, E.F. Unmasking activation of the zygotic genome using chromosomal deletions in the *Drosophila* embryo. *PLoS Biol.* **2007**, *5*, e117. [CrossRef]
75. Li, L.; Zhu, Q.; He, X.; Sinha, S.; Halfon, M.S. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. *Genome Biol.* **2007**, *8*, R101. [CrossRef] [PubMed]
76. Papatsenko, D.; Goltsev, Y.; Levine, M. Organization of developmental enhancers in the *Drosophila* embryo. *Nucleic Acids Res.* **2009**, *37*, 5665–5677. [CrossRef] [PubMed]
77. Zinzen, R.P.; Girardot, C.; Gagneur, J.; Braun, M.; Furlong, E.E. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **2009**, *462*, 65–70. [CrossRef] [PubMed]
78. Erceg, J.; Pakozdi, T.; Marco-Ferreres, R.; Ghavi-Helm, Y.; Girardot, C.; Bracken, A.P.; Furlong, E.E. Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements. *Genes Dev.* **2017**, *31*, 590–602. [CrossRef] [PubMed]
79. Blick, A.J.; Mayer-Hirshfeld, I.; Malibiran, B.R.; Cooper, M.A.; Martino, P.A.; Johnson, J.E.; Bateman, J.R. The Capacity to Act in Trans Varies among *Drosophila* Enhancers. *Genetics* **2016**, *203*, 203–218. [CrossRef]
80. Vincent, B.J.; Staller, M.V.; Lopez-Rivera, F.; Bragdon, M.D.J.; Pym, E.C.G.; Biette, K.M.; Wunderlich, Z.; Harden, T.T.; Estrada, J.; DePace, A.H. Hunchback is counter-repressed to regulate even-skipped stripe 2 expression in *Drosophila* embryos. *PLoS Genet.* **2018**, *14*, e1007644. [CrossRef]
81. Samee, M.A.H.; Lydiard-Martin, T.; Biette, K.M.; Vincent, B.J.; Bragdon, M.D.; Eckenrode, K.B.; Wunderlich, Z.; Estrada, J.; Sinha, S.; DePace, A.H. Quantitative Measurement and Thermodynamic Modeling of Fused Enhancers Support a Two-Tiered Mechanism for Interpreting Regulatory DNA. *Cell Rep.* **2017**, *21*, 236–245. [CrossRef]
82. Gisselbrecht, S.S.; Palagi, A.; Kurland, J.V.; Rogers, J.M.; Ozadam, H.; Zhan, Y.; Dekker, J.; Bulyk, M.L. Transcriptional Silencers in *Drosophila* Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Mol. Cell* **2020**, *77*, 324–337.e8. [CrossRef]
83. Soluri, I.V.; Zumerling, L.M.; Payan Parra, O.A.; Clark, E.G.; Blythe, S.A. Zygotic pioneer factor activity of Odd-paired/Zic is necessary for late function of the *Drosophila* segmentation network. *eLife* **2020**, *9*. [CrossRef]
84. Li, X.Y.; MacArthur, S.; Bourgon, R.; Nix, D.; Pollard, D.A.; Iyer, V.N.; Hechmer, A.; Simirenko, L.; Stapleton, M.; Luengo Hendriks, C.L.; et al. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* **2008**, *6*, e27. [CrossRef]
85. Li, X.Y.; Thomas, S.; Sabo, P.J.; Eisen, M.B.; Stamatoyannopoulos, J.A.; Biggin, M.D. The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol.* **2011**, *12*, R34. [CrossRef]
86. MacArthur, S.; Li, X.Y.; Li, J.; Brown, J.B.; Chu, H.C.; Zeng, L.; Grondona, B.P.; Hechmer, A.; Simirenko, L.; Keranen, S.V.; et al. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **2009**, *10*, R80. [CrossRef] [PubMed]
87. Negre, N.; Brown, C.D.; Shah, P.K.; Kheradpour, P.; Morrison, C.A.; Henikoff, J.G.; Feng, X.; Ahmad, K.; Russell, S.; White, R.A.; et al. A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet.* **2010**, *6*, e1000814. [CrossRef] [PubMed]
88. Moshkovich, N.; Nisha, P.; Boyle, P.J.; Thompson, B.A.; Dale, R.K.; Lei, E.P. RNAi-independent role for Argonaute2 in CTCF/CP190 chromatin insulator function. *Genes Dev.* **2011**, *25*, 1686–1701. [CrossRef] [PubMed]

89. Khoroshko, V.A.; Levitsky, V.G.; Zykova, T.Y.; Antonenko, O.V.; Belyaeva, E.S.; Zhimulev, I.F. Chromatin Heterogeneity and Distribution of Regulatory Elements in the Late-Replicating Intercalary Heterochromatin Domains of *Drosophila melanogaster* Chromosomes. *PLoS ONE* **2016**, *11*, e0157147. [CrossRef]
90. Zhou, J.; Troyanskaya, O.G. Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states. *Nat. Commun.* **2016**, *7*, 10528. [CrossRef]
91. Mateo, L.J.; Murphy, S.E.; Hafner, A.; Cinquini, I.S.; Walker, C.A.; Boettiger, A.N. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **2019**, *568*, 49–54. [CrossRef]
92. Bozek, M.; Cortini, R.; Storti, A.E.; Unnerstall, U.; Gaul, U.; Gompel, N. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the *Drosophila* blastoderm. *Genome Res.* **2019**, *29*, 771–783. [CrossRef]
93. Ghavi-Helm, Y.; Klein, F.A.; Pakozdi, T.; Ciglar, L.; Noordermeer, D.; Huber, W.; Furlong, E.E. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **2014**, *512*, 96–100. [CrossRef]
94. Li, L.; Wunderlich, Z. An Enhancer's Length and Composition Are Shaped by Its Regulatory Task. *Front. Genet.* **2017**, *8*, 63. [CrossRef]
95. Schor, I.E.; Bussotti, G.; Males, M.; Forneris, M.; Viales, R.R.; Enright, A.J.; Furlong, E.E.M. Non-coding RNA Expression, Function, and Variation during *Drosophila* Embryogenesis. *Curr. Biol.* **2018**, *28*, 3547–3561.e9. [CrossRef]
96. Mikhaylichenko, O.; Bondarenko, V.; Harnett, D.; Schor, I.E.; Males, M.; Viales, R.R.; Furlong, E.E.M. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev.* **2018**, *32*, 42–57. [CrossRef] [PubMed]
97. Haines, J.E.; Eisen, M.B. Patterns of chromatin accessibility along the anterior-posterior axis in the early *Drosophila* embryo. *PLoS Genet.* **2018**, *14*, e1007367. [CrossRef] [PubMed]
98. Cusanovich, D.A.; Reddington, J.P.; Garfield, D.A.; Daza, R.M.; Aghamirzaie, D.; Marco-Ferreres, R.; Pliner, H.A.; Christiansen, L.; Qiu, X.; Steemers, F.J.; et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **2018**, *555*, 538–542. [CrossRef]
99. Costello, J.C.; Dalkilic, M.M.; Beason, S.M.; Gehlhausen, J.R.; Patwardhan, R.; Middha, S.; Eads, B.D.; Andrews, J.R. Gene networks in *Drosophila melanogaster*: Integrating experimental data to predict gene function. *Genome Biol.* **2009**, *10*, R97. [CrossRef]
100. Kazemian, M.; Blatti, C.; Richards, A.; McCutchan, M.; Wakabayashi-Ito, N.; Hammonds, A.S.; Celniker, S.E.; Kumar, S.; Wolfe, S.A.; Brodsky, M.H.; et al. Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biol.* **2010**, *8*. [CrossRef] [PubMed]
101. Marbach, D.; Roy, S.; Ay, F.; Meyer, P.E.; Candeias, R.; Kahveci, T.; Bristow, C.A.; Kellis, M. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* **2012**, *22*, 1334–1349. [CrossRef]
102. Pesch, R.; Zimmer, R. Cross-species Conservation of context-specific networks. *BMC Syst. Biol.* **2016**, *10*, 76. [CrossRef]
103. Reda, C.; Wilczynski, B. Automated inference of gene regulatory networks using explicit regulatory modules. *J. Theor. Biol.* **2020**, *486*, 110091. [CrossRef]
104. Yang, B.; Wittkopp, P.J. Structure of the Transcriptional Regulatory Network Correlates with Regulatory Divergence in *Drosophila*. *Mol. Biol. Evol.* **2017**, *34*, 1352–1362. [CrossRef]
105. Drosophila 12 Genomes Consortium; Clark, A.G.; Eisen, M.B.; Smith, D.R.; Bergman, C.M.; Oliver, B.; Markow, T.A.; Kaufman, T.C.; Kellis, M.; Gelbart, W.; et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **2007**, *450*, 203–218. [CrossRef] [PubMed]
106. Hare, E.E.; Peterson, B.K.; Iyer, V.N.; Meier, R.; Eisen, M.B. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet.* **2008**, *4*, e1000106. [CrossRef] [PubMed]
107. He, B.Z.; Holloway, A.K.; Maerkl, S.J.; Kreitman, M. Does positive selection drive transcription factor binding site turnover? A test with *Drosophila* cis-regulatory modules. *PLoS Genet.* **2011**, *7*, e1002053. [CrossRef] [PubMed]
108. Holloway, A.K.; Begun, D.J.; Siepel, A.; Pollard, K.S. Accelerated sequence divergence of conserved genomic elements in *Drosophila melanogaster*. *Genome Res.* **2008**, *18*, 1592–1601. [CrossRef] [PubMed]
109. Jiang, P.; Ludwig, M.Z.; Kreitman, M.; Reinitz, J. Natural variation of the expression pattern of the segmentation gene even-skipped in melanogaster. *Dev. Biol.* **2015**, *405*, 173–181. [CrossRef]
110. Khoueiry, P.; Girardot, C.; Ciglar, L.; Peng, P.C.; Gustafson, E.H.; Sinha, S.; Furlong, E.E. Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *eLife* **2017**, *6*. [CrossRef]
111. Macdonald, S.J.; Long, A.D. Fine scale structural variants distinguish the genomes of *Drosophila melanogaster* and *D. pseudoobscura*. *Genome Biol.* **2006**, *7*, R67. [CrossRef]
112. Aerts, S.; van Helden, J.; Sand, O.; Hassan, B.A. Fine-tuning enhancer models to predict transcriptional targets across multiple genomes. *PLoS ONE* **2007**, *2*, e1115. [CrossRef]
113. Brody, T.; Rasband, W.; Baler, K.; Kuzin, A.; Kundu, M.; Odenwald, W.F. cis-Decoder discovers constellations of conserved DNA sequences shared among tissue-specific enhancers. *Genome Biol.* **2007**, *8*, R75. [CrossRef]
114. Guo, H.; Huo, H.; Yu, Q. SMCis: An Effective Algorithm for Discovery of Cis-Regulatory Modules. *PLoS ONE* **2016**, *11*, e0162968. [CrossRef] [PubMed]
115. Ivan, A.; Halfon, M.S.; Sinha, S. Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs. *Genome Biol.* **2008**, *9*, R22. [CrossRef] [PubMed]

116. Su, J.; Teichmann, S.A.; Down, T.A. Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol.* **2010**, *6*, e1001020. [CrossRef] [PubMed]
117. Arbel, H.; Basu, S.; Fisher, W.W.; Hammonds, A.S.; Wan, K.H.; Park, S.; Weiszmann, R.; Booth, B.W.; Keranen, S.V.; Henriquez, C.; et al. Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 900–908. [CrossRef] [PubMed]
118. Schember, I.; Halfon, M.S. Identification of new Anopheles gambiae transcriptional enhancers using a cross-species prediction approach. *Insect Mol. Biol.* **2021**. [CrossRef] [PubMed]
119. NCBI Assembly. Available online: https://www.ncbi.nlm.nih.gov/assembly/GCA_001676475.1 (accessed on 26 May 2021).
120. i5K Sequenced Arthropod Genomes. Available online: http://i5k.github.io/arthropod_genomes_at_ncbi (accessed on 26 May 2021).
121. Hotaling, S.; Sproul, J.S.; Heckenhauer, J.; Powell, A.; Larracuente, A.M.; Pauls, S.U.; Kelley, J.L.; Frandsen, P.B. Long-reads are revolutionizing 20 years of insect genome sequencing. *Genome Biol. Evol.* **2021**. [CrossRef]
122. Holt, C.; Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **2011**, *12*, 491. [CrossRef] [PubMed]
123. Bruna, T.; Hoff, K.J.; Lomsadze, A.; Stanke, M.; Borodovsky, M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **2021**, *3*, lqaa108. [CrossRef] [PubMed]
124. Wunderlich, Z.; Bragdon, M.D.; Vincent, B.J.; White, J.A.; Estrada, J.; DePace, A.H. Kruppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. *Cell Rep.* **2015**, *12*, 1740–1747. [CrossRef]
125. Cannavo, E.; Khoueiry, P.; Garfield, D.A.; Geeleher, P.; Zichner, T.; Gustafson, E.H.; Ciglar, L.; Korbel, J.O.; Furlong, E.E. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol.* **2016**, *26*, 38–51. [CrossRef] [PubMed]
126. Gontarz, P.; Fu, S.; Xing, X.; Liu, S.; Miao, B.; Bazylianska, V.; Sharma, A.; Madden, P.; Cates, K.; Yoo, A.; et al. Comparison of differential accessibility analysis strategies for ATAC-seq data. *Sci. Rep.* **2020**, *10*, 10150. [CrossRef]
127. Martins, A.L.; Walavalkar, N.M.; Anderson, W.D.; Zang, C.; Guertin, M.J. Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res.* **2018**, *46*, e9. [CrossRef] [PubMed]
128. Orchard, P.; Kyono, Y.; Hensley, J.; Kitzman, J.O.; Parker, S.C.J. Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. *Cell Syst.* **2020**, *10*, 298–306.e4. [CrossRef] [PubMed]
129. Wang, J.R.; Quach, B.; Furey, T.S. Correcting nucleotide-specific biases in high-throughput sequencing data. *BMC Bioinform.* **2017**, *18*, 357. [CrossRef]
130. Yandell, M.; Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **2012**, *13*, 329–342. [CrossRef] [PubMed]