insects

MDPI

Review

REDfly: An Integrated Knowledgebase for Insect Regulatory Genomics

Soile V. E. Keränen [1], Angel Villahoz-Baleta [2,3,†], Andrew E. Bruno [2,3] and Marc S. Halfon [3,4,5,6,7,*]

[1] Independent Research, Berkeley, CA 94705, USA; soileredfly@gmail.com
[2] Center for Computational Research, State University of New York at Buffalo, Buffalo, NY 14203, USA; angel.villahoz-baleta@usda.gov (A.V.-B.); aebruno2@buffalo.edu (A.E.B.)
[3] New York State Center of Excellence in Bioinformatics and Life Sciences, State University of New York at Buffalo, Buffalo, NY 14203, USA
[4] Department of Biochemistry, State University of New York at Buffalo, Buffalo, NY 14203, USA
[5] Department of Biomedical Informatics, State University of New York at Buffalo, Buffalo, NY 14203, USA
[6] Department of Biological Sciences, State University of New York at Buffalo, Buffalo, NY 14203, USA
[7] Department of Molecular and Cellular Biology and Program in Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA
* Correspondence: mshalfon@buffalo.edu; Tel.: +1-716-829-3126
† Present Address: United States Department of Agriculture, Coastal Plain Experiment Station, 2316 Rainwater Rd., Tifton, GA 31793, USA.

Simple Summary: Understanding how genes are regulated is a vital area of current biological research and a crucial adjunct to ongoing efforts to sequence entire genomes. Knowing the DNA sequences responsible for gene regulation—transcriptional *cis*-regulatory modules (CRMs, e.g., "enhancers") and transcription factor binding sites (TFBSs)—is important for many areas of research including interpretation and validation of data developed by large-scale genomics projects, providing training data for machine-learning CRM-discovery methods, genome annotation, modeling gene-regulatory networks, studying the evolution of gene regulation, and numerous aspects of the basic biology of transcriptional regulation. Knowledge of insect CRMs is also an important step in developing biotechnology methods for control of insect disease vectors and for eliminating pathogen transmission. The REDfly (Regulatory Element Database for Fly) database integrates all of the available insect *cis*-regulatory information from multiple sources to provide a comprehensive collection of known regulatory elements. In this paper, we describe REDfly's basic contents and data model, emphasizing recently added features, and provide illustrated walk-throughs of some common search scenarios.

Abstract: We provide here an updated description of the REDfly (Regulatory Element Database for Fly) database of transcriptional regulatory elements, a unique resource that provides regulatory annotation for the genome of *Drosophila* and other insects. The genomic sequences regulating insect gene expression—transcriptional *cis*-regulatory modules (CRMs, e.g., "enhancers") and transcription factor binding sites (TFBSs)—are not currently curated by any other major database resources. However, knowledge of such sequences is important, as CRMs play critical roles with respect to disease as well as normal development, phenotypic variation, and evolution. Characterized CRMs also provide useful tools for both basic and applied research, including developing methods for insect control. REDfly, which is the most detailed existing platform for metazoan regulatory-element annotation, includes over 40,000 experimentally verified CRMs and TFBSs along with their DNA sequences, their associated genes, and the expression patterns they direct. Here, we briefly describe REDfly's contents and data model, with an emphasis on the new features implemented since 2020. We then provide an illustrated walk-through of several common REDfly search use cases.

Keywords: insects; *Drosophila*; regulatory genomics; gene regulation; *cis*-regulatory module; enhancer; genome annotation

check for updates

Citation: Keränen, S.V.E.; Villahoz-Baleta, A.; Bruno, A.E.; Halfon, M.S. REDfly: An Integrated Knowledgebase for Insect Regulatory Genomics. *Insects* **2022**, *13*, 618. https://doi.org/10.3390/insects13070618

Academic Editors: Sourav Roy and Rebecca Spokony

Received: 16 June 2022
Accepted: 6 July 2022
Published: 11 July 2022

1. Introduction

The turn-of-the-century advent of fully sequenced metazoan genomes brought with it the first genome annotations, which were largely confined to positions of confirmed and predicted genes, and typically housed in community-specific model-organism databases, e.g., [1–4]. Remarkably, over two decades later, the major databases (see [5]) are still mostly lacking annotation of non-coding regulatory sequences. These sequences include distal "*cis*-regulatory modules" (CRMs), a generic term encompassing such regulatory elements as enhancers, which mediate positive gene regulation; silencers, involved in negative regulation; and a growing number of additional elements that are not easily classified including PREs, super-enhancers, insulators, tethering elements, and others [6–11].

Obtaining a comprehensive annotation of regulatory sequences is important not only for its intrinsic value in illuminating the structure and function of the genome, but also for its practical value in facilitating bioinformatics analyses of CRMs and their interactions with other genomic features. To this end, the REDfly database (Regulatory Element Database for fly) [12–15] plays a critical role for regulatory bioinformatics and genomics, particularly with respect to insects. REDfly is a highly curated knowledgebase dedicated to annotating and integrating the growing body of information on insect transcriptional regulatory sequences curated from the published literature, with an emphasis on empirically validated CRMs. Although originally focused solely on *Drosophila melanogaster*, REDfly now includes data from a growing number of additional insect species.

2. Utility of REDfly

Prior to the development of REDfly, large-scale analyses of regulatory sequences were challenging to conduct, as the bulk of the existing regulatory data was distributed among hundreds of individual publications. Consequently, what few analyses were completed were performed on small and frequently biased sets of CRMs, such as a limited subset of early developmental pair-rule stripe enhancers in *Drosophila*, e.g., [16–18]. By curating these data and making them findable, accessible, interoperable, and usable (FAIR) [19], REDfly made it possible to bring statistical, computational, and comparative genomics methods to bear on their study. REDfly enabled the first-ever large-scale, relatively unbiased analysis of CRMs, which immediately revealed novel insights into CRM-sequence composition, differences among tissue-specific groups of CRMs, and an early indication of the presence of enhancer RNAs (eRNAs) as a prevalent CRM characteristic [20]. REDfly, by continuing to compile the data from hundreds and eventually thousands of individual experiments scattered throughout four decades of literature, subsequently proved instrumental in facilitating studies in a wide variety of research areas, including:

Biology of CRMs. REDfly has been used to investigate the organization of TFBSs within CRMs [21] and how combinatorial binding influences CRM activity [22]. Soluri et al. [23] investigated how pioneer TFs control chromatin accessibility, and Blick et al. [24] examined the ability of CRMs to act in *trans*. REDfly data helped to illustrate how CRMs can have multiple functions [25], such as dual use as both enhancers and Polycomb response elements [26], or as both enhancers and silencers [27].

Interpretation of genomic data. REDfly has been critical for interpreting data from large-scale genomics projects including TF binding studies, e.g., [28,29] and studies of insulators [30,31]. A study challenging our understanding of which epigenetic marks characterize regulatory sequences depended on REDfly data [32]. REDfly has been used to study chromosome domains and chromatin "states", e.g., [33–36], to explore 3D-chromatin conformation [37,38], to study ncRNA and eRNA expression [39,40], and to validate scATAC-seq approaches, e.g., [41,42].

Computational CRM discovery. REDfly has played a dramatic role in methods for computational CRM discovery, both as a source of training data and as a method for validating predictions, e.g., [43–51]. Su et al. [52] used REDfly data to assess CRM-discovery approaches, which would have been impossible without REDfly. Computational CRM-discovery methods using REDfly for training data also can identify CRMs in diverse insect species [53] and, as such, provide a powerful tool for annotating insect regulatory genomes [54].

CRM evolution. REDfly has enabled studies of CRM evolution and TFBS turnover, e.g., [55–61]. Wang et al. [62] used REDfly data to investigate the selective pressure on DNA shape at TF binding sites, and Peng et al. [63] explored the relationship between chromatin accessibility and TF binding to predict evolutionary changes in enhancer activity.

As can be seen from these examples, REDfly is an important source of raw data for analysis, hypothesis generation, assessment, validation, and empirical research. In the remainder of the paper, we describe the REDfly data model and provide a guide to some common REDfly uses.

3. REDfly Data Model

REDfly curates two types of data: CRMs and transcription factor binding sites (TFBSs), with CRMs being the main focus. Historically, CRM annotations have been drawn from reporter gene assays in transgenic animals or cultured cells, but an increasingly diverse set of assays are starting to be included. In particular, several years ago REDfly started to capture CRMs identified through various "X-seq" assays, such as ATAC-seq, FAIRE-seq, DNase-seq, ChIP-seq, etc., as well as from purely computational predictions, in recognition of the fact that many regulatory sequences are presently being defined by these methods. There is considerable debate in the regulatory genomics field as to just how CRMs should be defined, with several studies indicating that the different methods of CRM identification have led to widely non-overlapping results and raising questions as to which, if any, of these methods most accurately identify CRMs [64–66]. As a result, REDfly separates its CRM data into four distinct subclasses: *reporter constructs (RC)*, *CRM_segments*, *predicted CRMs (pCRM)*, and *inferred CRMs (iCRM)*. *RCs* and *CRM_segments* are drawn from activity-based assays (Figure 1, left), primarily gene-expression data from either reporter genes or from native genes following mutation or deletion of regulatory sequences. *RCs* mainly represent reporter-gene results, assayed either in transgenic animals or in cell-culture assays; the two types of assays can be independently searched. The *CRM_segment* class contains sequences that are demonstrated to be necessary for gene regulation but, unlike in a reporter gene assay, are not necessarily sufficient. Such sequences can be obtained from the analysis of small chromosomal deletions or site-directed mutagenesis but are increasingly being found in the literature as a result of CRISPR/Cas9-mediated targeted sequence deletions. *pCRMs*, on the other hand, reflect CRMs identified by assays that do not require demonstration of activity, for example, the presence of histone modifications, or computational predictions (Figure 1, right). *iCRMs* are not curated from the literature, but represent putative regulatory elements based on the analysis of other REDfly data (see below).

3.1. "Reporter Constructs" and Their Attributes

RCs in REDfly have three primary associated attributes: *expression*, *CRM*, and *minimization* (Figure 2). *Expression* can be either "positive" or "negative" and denotes whether or not the sequence drives gene expression in the reporter-gene assay. Positive expression is described using the *Drosophila* Anatomy Ontology [67] (or, for non-*Drosophila* species, an appropriate species-specific anatomy ontology). Every expression pattern is linked with the stage(s) of development at which this expression is observed, using community-standard developmental staging ontologies, and with an indicator for whether that expression is consistent with or ectopic to that of the assigned target gene. Sexually dimorphic expression is also captured. Terms from the Gene Ontology [68] allow for annotation of regulatory elements that respond to specific signals or environmental cues (e.g., wound healing, hypoxia, circadian cycling).

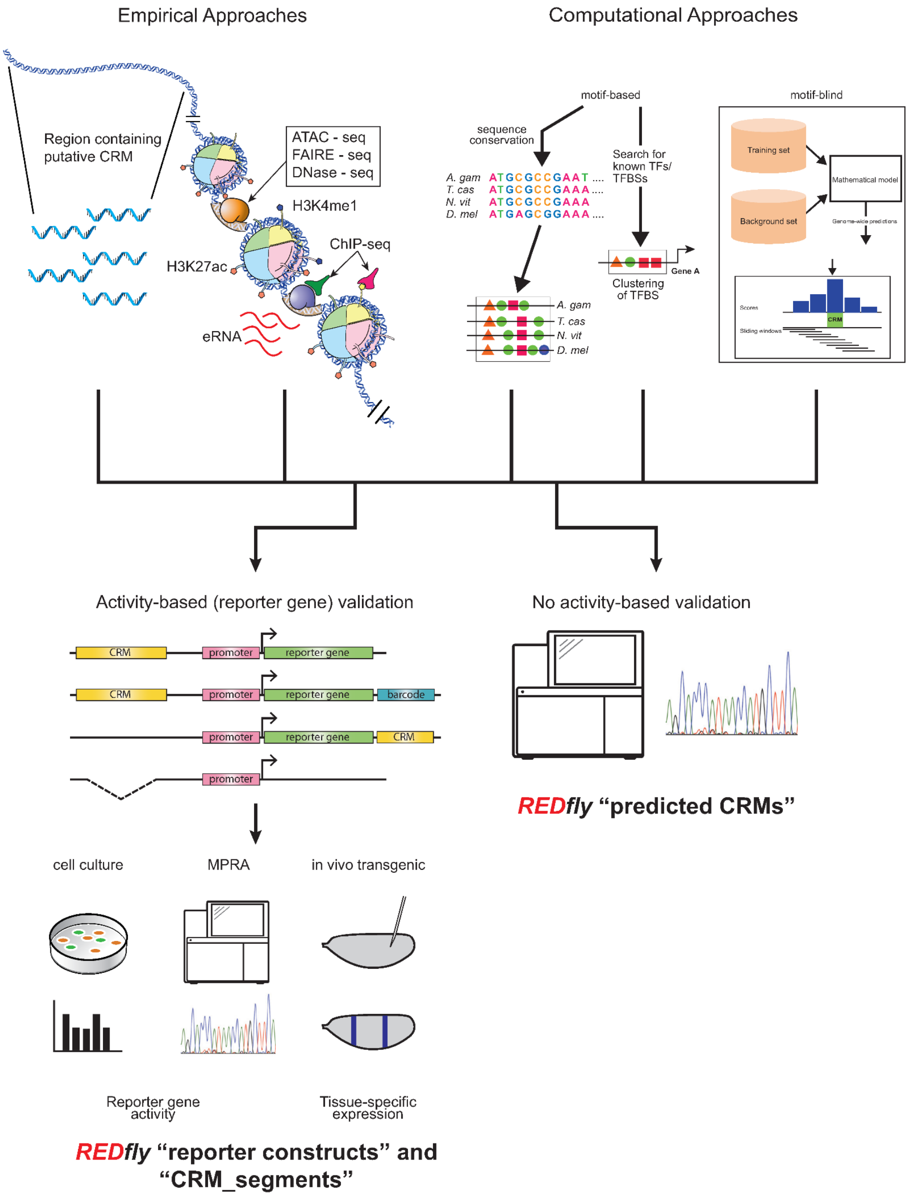


Figure 1. Activity-based and non-activity-based methods for defining regulatory sequences. Top: A wide variety of methods exist for identifying regulatory sequences based on both empirical (**left**) and computational (**right**) approaches. Empirical approaches include unbiased testing of non-coding DNA regions as well as selection of sequences based on chromatin accessibility, histone post-translational modification, transcription factor binding, production of enhancer RNAs, and others. Computational approaches may include assessment of sequence conservation, presence of transcription factor binding motifs, or various machine-learning methods. Bottom: Results from these regulatory element discovery methods can be obtained with (**left**) or without (**right**) the use of activity-based criteria. Activity-based criteria typically involve some sort of reporter-gene assay, which might be performed in cultured cells, using next-generation sequencing in a "massively parallel reporter assay" (MPRA), or in transgenic animals; recently, testing via genomic deletion via CRISPR/Cas9 has also been gaining popularity. REDfly classifies regulatory sequences derived from these methods as *reporter constructs (RCs)* and *CRM_segments*, while any methods that identify regulatory sequences without recourse to activity-based criteria are referred to as *predicted CRMs (pCRMs)*. These somewhat historical definitions should not be construed to imply that one or the other type of data is more accurate or "correct".

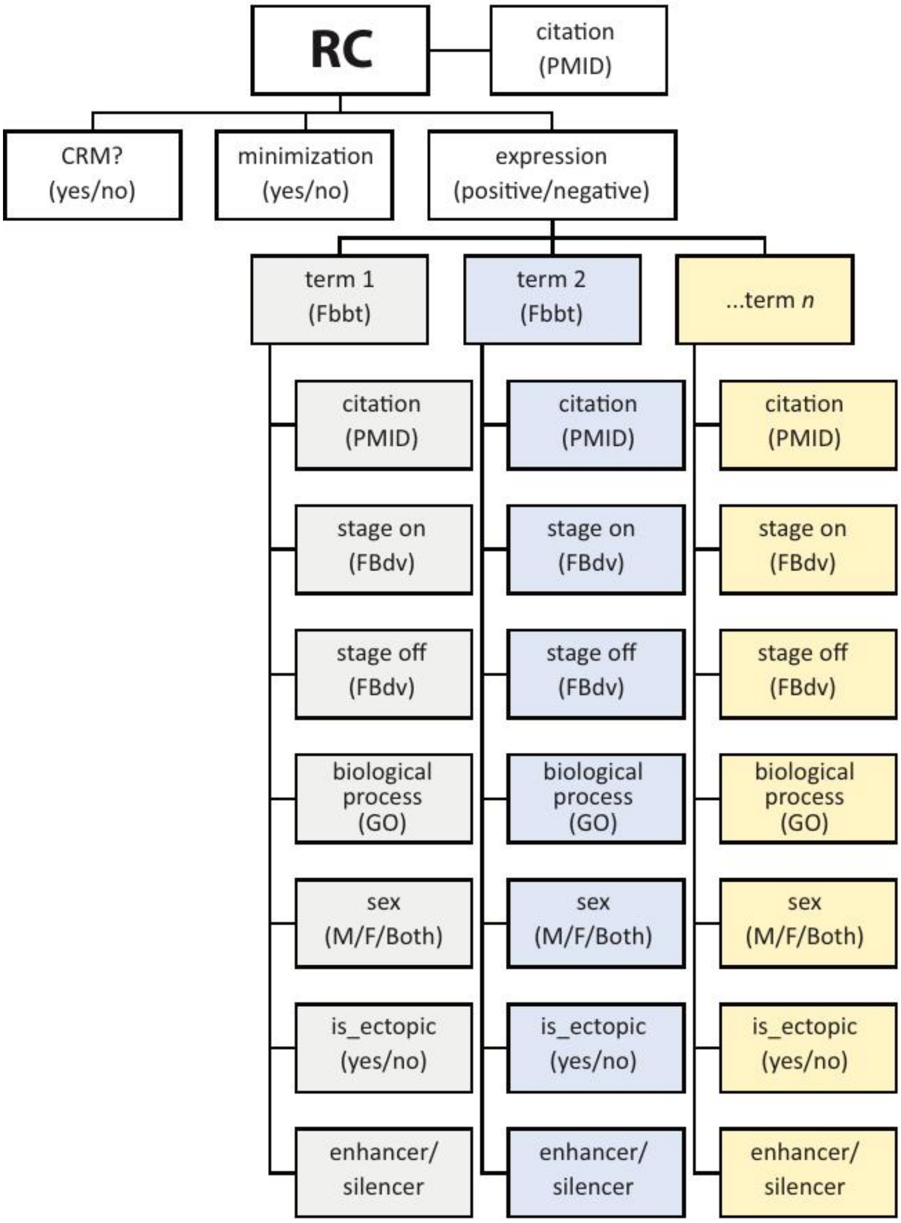


Figure 2. The basic REDfly *Reporter Construct* data model. Depicted is a partial illustration of the data model for REDfly *Reporter Construct (RC)* records. Each RC has the three basic attributes of *CRM*, *minimization*, and *expression*. If *expression* is positive, the RC is annotated with each anatomical location where expression is observed, to the most granular degree available, based on a species-relevant anatomy ontology. Each annotated expression pattern is then associated with additional data including citation; the stages at which expression is observed; a biological process (where relevant) drawn from the Gene Ontology; the sex in which the expression is observed; whether or not the expression is ectopic with respect to the known pattern of the associated gene; and whether the RC is acting as an enhancer or a silencer in the current tissue. Not included in the schematic are additional RC-associated data such as species, gene names, relevant figure panels, evidence terms, and others.

Recent data suggest that many regulatory sequences can act as enhancers in one cell type, while simultaneously acting as transcriptional silencers in another [27]. An "enhancer/silencer" tag associated with each described expression pattern allows for this dual functionality to be accurately represented (Figure 2). For example, [69] describe a sequence, *Tm2_intronI1B(b)*, that acts as a silencer by suppressing activity of the *Tm2_intronI1b(a)* enhancer in embryonic and larval muscle, but which also acts as a weak enhancer to promote expression in adult leg muscle. In the REDfly "anatomical expression" tab, this *RC* would, therefore, have the annotation "embryonic/larval somatic muscle" and "silencer" in one row, and "skeletal muscle of leg" and "enhancer" in another.

The *CRM* attribute indicates whether an *RC* is the shortest of a set of nested sequences with identical activity, i.e., what is commonly referred to as a "minimal enhancer" (Figure 3B,D,F). When an *RC* is part of a set of nested sequences, we say the set of *RCs* has undergone *minimization* (Figure 3C–I). This attribute is included as an aid to researchers to help decide whether to undertake experimental analysis of a region, as a minimized region might provide less new information than one for which only a single construct has been tested. Note that none of these attributes are fixed for a given RC record, as the attribute values might change over time as new information becomes available through follow-up studies and new literature.

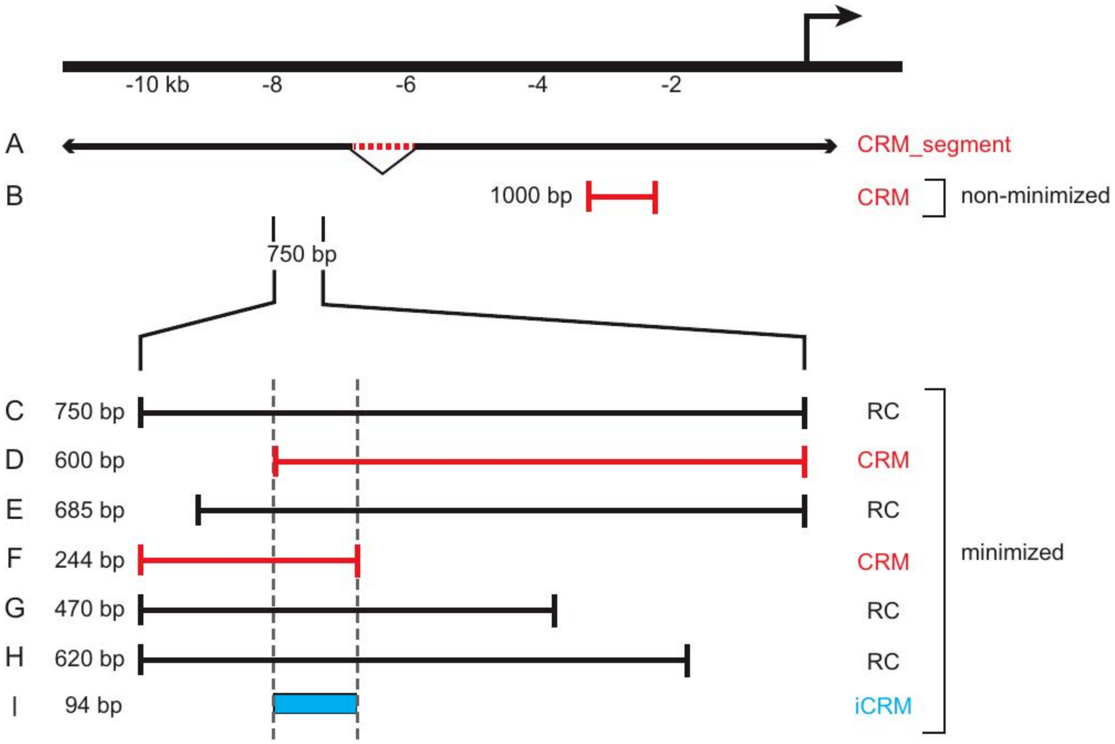


Figure 3. REDfly data subclasses and *Reporter Construct* attributes. The figure illustrates a hypothetical locus for which different sequences have been tested in vivo. Sequence (**A**) represents an approximately 1 kb genomic deletion (red dotted line) that reduces expression from the nearby promoter 6 kb downstream (bent arrow). As such, it is considered a *CRM_segment*. Sequence (**B**) is a 1 kb sequence fragment located roughly 2 kb upstream of the transcription start. Since it is an isolated sequence, it is considered to be a *CRM* that has not been subject to minimization. If this construct showed reporter gene activity, it would be designated as "expression positive"; otherwise, it would be labeled "expression negative." Constructs (**C**–**H**) are part of an overlapping and partially nested series of sequences spanning 750 bp of DNA 7.25 kb upstream of the transcription start. In this example, each drives the identical pattern of reporter gene expression. Since each of these sequences overlaps at least one other, we consider this region and the six sequences to have undergone minimization. Sequences (**D**,**F**) are each the shortest of a respective set of fully nested sequences and are, therefore, considered to be *CRMs* (marked in red). The remaining sequences are designated as *RCs* (black). A 94 bp sequence, (**I**), marks the minimal region of overlap among all of the sequences and is, thus, registered in REDfly as an *inferred CRM* (*iCRM*, blue).

3.2. "Inferred CRMs"

iCRMs represent sequences that have not been explicitly identified as CRMs from any assay, but that are inferred to have regulatory activity based on analysis of other sequences in REDfly. For instance, overlapping sequences may have the same regulatory activity when assayed in vivo, and a logical—although unproven—supposition in such a case is that the overlapping region contains the "true" minimal CRM (Figure 3I). These overlaps can arise from *RCs* that were assayed in different publications and, therefore, are only discovered through integrated curation in REDfly.

4. Species in REDfly

Although REDfly has historically been focused on *Drosophila melanogaster*, the clear value of comparative genomics and of working with non-traditional model organisms, a vast increase in the number of sequenced insect genomes, and the small but growing availability of both predicted and validated insect regulatory sequences has led us to expand REDfly by implementing the ability to curate regulatory sequences for additional insect species. Information on which species are represented can be found on the "Species" page; current species include the mosquitoes *Anopheles gambiae* and *Aedes aegypti* and the beetle *Tribolium castaneum*. Additional species will be added as data accumulate. Since insect CRMs are often tested using transgenic *Drosophila*, REDfly divides the sequence and gene-feature data and the expression pattern and cell-line data into separate components. Each REDfly record has both a "sequence from" and an "assayed in" component. Sequence and gene-feature data are linked to the former, and anatomy and staging data to the latter. While it is preferable to describe species-specific reporter-gene-expression patterns using the proper species-specific anatomy ontology, many species lack an ontology as rich in terms as that for *Drosophila*. Therefore, terms from the *Drosophila* ontology can also be used to annotate expression in other species.

In order to facilitate research using these newly added genomes, we have implemented interfaces for *BLAT* [70] and in silico *PCR* [71] for each species included in REDfly. These can be accessed through the "Species" page.

5. Contents of REDfly

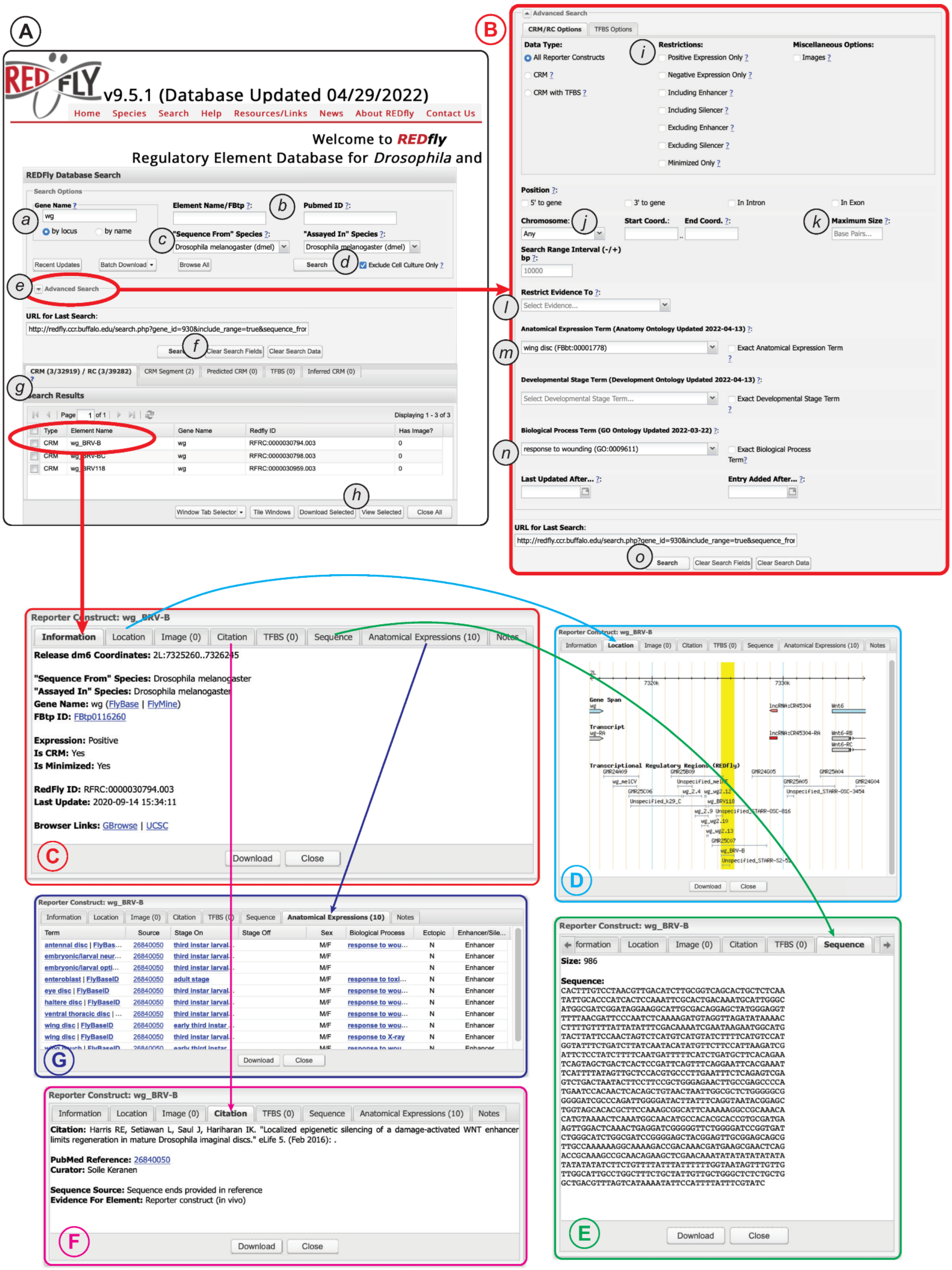
REDfly has continued to expand its contents at a rapid rate (Table 1). Since the end of 2019, the number of curated publications has increased by 30%, leading to an increase in the total number of Reporter Construct records by over 25% and in the number of pCRMs by almost 60%. Not reflected in these numbers, however, is an ambitious endeavor to update all RC records with the full set of RC expression attributes (developmental staging, sexually dimorphic activity, ectopic activity, and enhancer/silencer activity), which did not become a full part of the REDfly data model until the release of REDfly v6 in 2020. Since that time, over one-third of the RC records have been updated to contain this information.

Table 1. REDfly contents as of 1 July 2022.

Reporter Constructs (RCs)	43,819
From in vivo *reporter genes*	*21,690*
Associated with staging and other attributes	*17,961*
CRM_segments	16
Predicted CRMs (pCRMs)	14,318
Inferred CRMs (iCRMs)	7760
Transcription Factor Binding Sites (TFBS)	2717
Publications curated	1366

6. Using REDfly

The extensive data and metadata REDfly provides for each of its records allows for detailed customized searching of the database contents. Typical entry points for a REDfly search are via a gene name (Figure 4A(a)) or a literature reference (via PubMed ID, Figure 4A(b)). By default, searching for a gene name will execute a "by locus" search in which any elements annotated as being associated with that gene, as well as any elements found within a user-customizable range of 10 kb upstream or downstream of the gene (regardless of assigned target gene), will be returned. Moreover, by default, elements identified solely by assays performed in cultured cells are omitted from the results (Figure 4A(d)); unchecking the check-box causes these results to be included.



v9.5.1 (Database Updated 04/29/2022)

Welcome to ***RED*fly**
Regulatory Element Database for *Drosophila* and

REDFly Database Search

Search Options
Gene Name ?: wg
● by locus ○ by name
Element Name/FBtp ?:
Pubmed ID ?:
"Sequence From" Species ?: Drosophila melanogaster (dmel)
"Assayed In" Species ?: Drosophila melanogaster (dmel)
☑ Exclude Cell Culture Only ?

Advanced Search

URL for Last Search: http://redfly.ccr.buffalo.edu/search.php?gene_id=930&include_range=true&sequence_fror

CRM (3/32919) / RC (3/39282) ? | CRM Segment (2) | Predicted CRM (0) | TFBS (0) | Inferred CRM (0)

Search Results

Page 1 of 1 — Displaying 1 - 3 of 3

	Type	Element Name	Gene Name	Redfly ID	Has Image?
□	CRM	wg_BRV-B	wg	RFRC:0000030794.003	0
□	CRM	wg_BRV-BC	wg	RFRC:0000030798.003	0
□	CRM	wg_BRV118	wg	RFRC:0000030959.003	0

Advanced Search

Data Type:
● All Reporter Constructs
○ CRM ?
○ CRM with TFBS ?

Restrictions:
□ Positive Expression Only ?
□ Negative Expression Only ?
□ Including Enhancer ?
□ Including Silencer ?
□ Excluding Enhancer ?
□ Excluding Silencer ?
□ Minimized Only ?

Miscellaneous Options:
□ Images ?

Position ?: □ 5' to gene □ 3' to gene □ In Intron □ In Exon
Chromosome: Any
Start Coord.: .. **End Coord. ?**:
Maximum Size ?: Base Pairs...
Search Range Interval (-/+) bp ?: 10000
Restrict Evidence To ?: Select Evidence...
Anatomical Expression Term (Anatomy Ontology Updated 2022-04-13) ?: wing disc (FBbt:00001778) □ Exact Anatomical Expression Term ?
Developmental Stage Term (Development Ontology Updated 2022-04-13) ?: Select Developmental Stage Term... □ Exact Developmental Stage Term ?
Biological Process Term (GO Ontology Updated 2022-03-22) ?: response to wounding (GO:0009611) □ Exact Biological Process Term?
Last Updated After... ?:
Entry Added After... ?:

URL for Last Search: http://redfly.ccr.buffalo.edu/search.php?gene_id=930&include_range=true&sequence_fror

Reporter Construct: wg_BRV-B

Release dm6 Coordinates: 2L:7325260..7326245

"Sequence From" Species: Drosophila melanogaster
"Assayed In" Species: Drosophila melanogaster
Gene Name: wg (FlyBase | FlyMine)
FBtp ID: FBtp0116260

Expression: Positive
Is CRM: Yes
Is Minimized: Yes

RedFly ID: RFRC:0000030794.003
Last Update: 2020-09-14 15:34:11

Browser Links: GBrowse | UCSC

Reporter Construct: wg_BRV-B

Term	Source	Stage On	Stage Off	Sex	Biological Process	Ectopic	Enhancer/Sile...
antennal disc | FlyBas...	26840050	third instar larval...		M/F	response to wou...	N	Enhancer
embryonic/larval neur...	26840050	third instar larval...		M/F		N	Enhancer
embryonic/larval opti...	26840050	third instar larval...		M/F		N	Enhancer
enteroblast | FlyBaseID	26840050	adult stage		M/F	response to toxi...	N	Enhancer
eye disc | FlyBaseID	26840050	third instar larval...		M/F	response to wou...	N	Enhancer
haltere disc | FlyBaseID	26840050	third instar larval...		M/F	response to wou...	N	Enhancer
ventral thoracic disc | ...	26840050	third instar larval...		M/F	response to wou...	N	Enhancer
wing disc | FlyBaseID	26840050	early third instar ...		M/F	response to wou...	N	Enhancer
wing disc | FlyBaseID	26840050	third instar larval...		M/F	response to X-ray	N	Enhancer
wing pouch | FlyBaseID	26840050	early third instar ...		M/F	response to wou...	N	Enhancer

Reporter Construct: wg_BRV-B

Citation: Harris RE, Setiawan L, Saul J, Hariharan IK. "Localized epigenetic silencing of a damage-activated WNT enhancer limits regeneration in mature Drosophila imaginal discs." eLife 5. (Feb 2016): .

PubMed Reference: 26840050
Curator: Soile Keranen

Sequence Source: Sequence ends provided in reference
Evidence For Element: Reporter construct (in vivo)

Reporter Construct: wg_BRV-B

Reporter Construct: wg_BRV-B

Size: 986

Sequence:
```
CACTTTGTCCTAACGTTGACATCTTGCGGTCAGCACTGCTCTCAA
TATTGCACCCATCACTCCAAATTCGCACTGACAAATGCATTGGGC
ATGGCGATCGGATAGGAAGGCATTGCGACAGGAGCTATGGGAGGT
TTTTAACGATTCCCAATCTCAAAAGATGTAGGTTAGATATAAAAC
CTTTTGTTTTATTATATTTCGACAAAATCGAATAAGAATGGCATG
TACTTATTCCAACTAGTCTCATGTCATGTATCTTTTCATGTCCAT
GGTATTTCTGATCTTATCAATACATATGTTCTTCCATTAAGATCG
ATTCTCCTATCTTTTCAATGATTTTTCATCTGATGCTTCACAGAA
TCAGTAGCTGACTCACTCCGATTCAGTTTCAGGAATTCACGAAAT
TCATTTTATAGTTGCTCCACGTGCCCTTGAATTTCTCAGAGTCGA
GTCTGACTAATACTTCCTTCCGCTGGGAGAACTTGCCGAGCCCCA
TGAATCCACAACTCACAGCTGTAACTAATTGGCGCTCTGGGGGCG
GGGGATCGCCCAGATTGGGGATACTTATTTCAGGTAATACGGAGC
TGGTAGCACACGCTTCCAAAGCGGCATTCAAAAAGGCCGCAAACA
CATGTAAAACTCAAATGGCAACATGCCACACGCACCGTGCGATGA
AGTTGGACTCAAACTGAGGATCGGGGGTTCTGGGGATCCGGTGAT
CTGGGCATCTGGCGATCCGGGGAGCTACGGAGTTGCGGAGCAGCG
TTGCCAAAAAAGGCAAAAGACCGACAAACGATGAAGCGAACTCAG
ACCGCAAAGCCGCAACAGAAGCTCGAACAAATATATATATATATA
TATATATATCTTCTGTTTTATTTATTTTTTGGTAATAGTTTGTTG
TTGGCATTGCCTGGCTTTCTGCTATTGTTGCTGGGCTCTCTGCTG
GCTGACGTTTAGTCATAAAATATTCCATTTTATTTCGTATC
```

Figure 4. The REDfly search interface. See text for details. (**A**) The basic search panel. (**B**) The Advanced Search panel. (**C**) The Detailed Results "Information" pane. (**D**) The Detailed Results "Location" pane. (**E**) The Detailed Results "Sequence" pane. (**F**) The Detailed Results "Citation" pane. (**G**) The Detailed Results "Anatomical Expression" pane.

Clicking on the “Advanced Search” arrow (Figure 4A(e)) allows access to a large variety of additional options (Figure 4B), including the ability to restrict searches to specific *RC* attributes, genomic locations, anatomical regions, developmental stages, or biological processes. More detailed and complex search capabilities are under development.

Regardless of whether “basic” or “advanced” search is used, a summary of the results will appear in the “Search Results” pane directly below the main search window (Figure 4A(g)). Results for each REDfly data class—*RCs/CRMs*, *CRM_segments*, *pCRMs*, *TFBSs*, and *iCRMs*—are displayed in individual tabs to make it easier to view results by type. Checkboxes allow selection of records for download (Figure 4A(h)) in any of a number of convenient formats. Alternatively, clicking on an individual result will open a multipaned “Detailed View” window containing full information for the selected record (Figure 4C). Basic location and attribute data are displayed in the “Information” tab, along with links to relevant model-organism databases and genome browsers (Figure 4C). The “Location” tab provides a snapshot of the element in its genomic milieu (Figure 4D). The “Sequence” tab displays the genomic sequence and its size (Figure 4E), while the “Citation” tab (Figure 4F) provides a citation and link to the publication describing the current element, plus a description of the evidence used by REDfly curators to annotate sequence and expression information. The “Anatomical Expression” tab (Figure 4G) lists each cell type or tissue where the regulatory element is active, along with a specific citation for that activity data (since activity data may be drawn from multiple references), developmental staging for the observed activity, and the other attributes discussed above, e.g., sexually dimorphic activity, ectopic expression, and enhancer or silencer activity.

Since CRM activity can be complex and not easily summarized using the anatomical and staging terms available in the relevant ontologies, we also supply a “Notes” tab containing details and clarifications.

7. Use Cases

Below, we illustrate several common scenarios for using REDfly, with step-by-step instructions.

Note: The “Clear Search Fields” button (Figure 4A(f)) can be used to reset the search interface to the default settings and empty all search fields.

Case 1: I Want to Find All *D. melanogaster* Regulatory Features within a Specified Locus

One common use of REDfly is to explore what is known about the regulation of a particular gene. This is easily done:

(a) Make sure that both the “sequence from” and “assayed in” fields (Figure 4A(c)) are set to *Drosophila melanogaster*;
(b) Select the locus of interest using the “Gene Name” field (Figure 4A(a));
(c) Make sure the radio button is set to “by locus”;
(d) Click on “search” (Figure 4A(d)).

This will retrieve all features directly associated with the specified gene as well as any other features within 10 kb upstream or downstream of that gene. To alter the size of the region to be searched do the following:

(e) Open the Advanced Search box (Figure 4A(e));
(f) Change the value in the “Search Range Interval (−/+)” field from 10,000 to the desired number of basepairs;
(g) Click on “search” (Figure 4B(o)).

Case 2: I Want to Find All *D. melanogaster* Regulatory Features within a Genomic Region

It is also simple to determine what regulatory features are present in a given region of the genome, of arbitrary size, without needing to specify a particular gene or genes within the region:

(a) Make sure that both the "sequence from" and "assayed in" fields (Figure 4A(c)) are set to *Drosophila melanogaster*;
(b) Open the Advanced Search box (Figure 4A(e));
(c) Set the "Chromosome", "Start Coord.", and "End Coord." fields (Figure 4B(j)) to reflect your region of interest (make sure that the chromosomes are from the correct species, e.g., "3R (dmel)";
(d) Click on "search" (Figure 4B(o)).

Case 3: I Want to Find All *Anopheles gambiae* Sequences Tested for Regulatory Activity Using a Transgenic Anopheles Gambiae Assay

As REDfly begins to curate regulatory data from a wider variety of species, it become important to be able to isolate data for a species of interest, as well as to be able to separate out data obtained by direct observation in the desired species, versus heterologous testing in a host organism or cells of a different species. The two "species" search fields allow for these distinctions.

(a) Set the "sequence from" and "assayed in" fields (Figure 4A(c)) to "*Anopheles gambiae*";
(b) Click on "search" (Figure 4A(d)).

Case 4: I Want to Find All *Aedes aegypti* Sequences Tested for Regulatory Activity Using a Transgenic Drosophila Assay

(a) Set the "sequence from" field (Figure 4A(c)) to "*Aedes aegypti*";
(b) Set the "assayed in" field "*Drosophila melanogaster*";
(c) Click on "search" (Figure 4A(d)).

Case 5: I Want to Find All *D. melanogaster* Sequences That Were Tested In Vivo and Found to Be Negative for Regulatory Activity

It is often desirable to know that a sequence has been tested for regulatory activity and demonstrated not to be a CRM, at least in a particular context. Since REDfly curates all experimental data, regardless of outcome, such a search is straightforward.

(a) Make sure that both the "sequence from" and "assayed in" fields (Figure 4A(c)) are set to *Drosophila melanogaster*;
(b) Make sure the "Exclude Cell Culture Only" box is checked (Figure 4A(d));
(c) Open the Advanced Search box (Figure 4A(e));
(d) In the "Restrictions" section (Figure 4B(i)), check the "Negative Expression Only" box;
(e) Click on "search" (Figure 4B(o)).

Case 6: I Want to Find All *D. melanogaster* Regulatory Sequences Shorter Than 1 kb in Length Discovered Using Reporter Gene Assays in Cell Culture (Excluding Results from STARR-Seq Assays)

Investigators test sequences of greatly varying size for regulatory activity, but, for many uses, one might want to focus on sequences of only a certain length. We show how to do that in this use case, along with an illustration of how to confine results to those obtained in cell culture reporter gene assays other than STARR-seq [72].

(a) Make sure that both the "sequence from" and "assayed in" fields (Figure 4A(c)) are set to *Drosophila melanogaster*;
(b) Uncheck the "Exclude Cell Culture" box (Figure 4A(d));
(c) Open the Advanced Search box (Figure 4A(e));
(d) Type "1000" in the "Maximum Size" box (Figure 4B(k));
(e) In the "Restrict evidence to" dropdown select "reporter construct (cell culture)" (Figure 4B(l));
(f) Click on "search" (Figure 4B(o)).

Case 7: I Want to Find All *D. melanogaster* Regulatory Sequences That Drive Reporter Gene Expression in the Larval Wing Disc as a Response to Injury

Some of the most powerful uses for REDfly data are to assemble tissue-specific or function-specific sets of CRMs for experimental use or bioinformatic analysis. The use of anatomy, staging, and biological process ontologies provides for great flexibility and allows the user to determine how granular of a search to conduct.

(a) Make sure that both the "sequence from" and "assayed in" fields (Figure 4A(c)) are set to *Drosophila melanogaster*;
(b) Open the Advanced Search box (Figure 4A(e));
(c) In the "Anatomical Expression Term" box (Figure 4B(m)), begin typing "wing disc" and select this term when it appears in the completion list;
(d) In the "Biological Process Term" box (Figure 4B(n)), begin typing "wound healing" and select "Response to wounding" from the completion list;
(e) Click on "search" (Figure 4B(o)).

The above use cases illustrate only some of the many types of ways to access the REDfly regulatory data. Moreover, as powerful as these search capabilities are, they have not fully kept up with the growth of REDfly's data model over the years, such that the data contained in REDfly are in fact richer than what can be easily searched and downloaded. A major development priority over the next year will be to introduce new and more flexible search and download capabilities along with improved integration of the different data types and species curated by REDfly. We also expect there to be considerable growth in several data categories. These include *CRM_segments*, as researchers increasingly turn to CRISPR/Cas9-mediated deletion of regulatory sequences, and *pCRMs* from multiple new species as experimental methods, such as single-cell ATAC-seq [73] and computational CRM discovery methods, such as SCRMshaw (reviewed by [54]) continue to be applied to sequenced insect species at a rapid rate.

8. Access to REDfly

REDfly is freely available to the public at http://redfly.ccr.buffalo.edu. News about database updates and new features can be obtained from our Twitter feed @REDfly_database. Subsets of REDfly data can also be obtained from FlyBase [74] and FlyMine [75].

9. Data Submission

REDfly is a curated resource, with all data entry handled by our team of biocurators. The biocuration team engages in frequent back-and-forth communication with authors to ensure that published regulatory data are accurately reflected in REDfly's records. This includes a post-curation author review step, through which an e-mail is automatically sent inviting the corresponding author of each newly curated paper to review and, if necessary, correct the REDfly annotation. Although there is currently no mechanism for direct outside data submission, we encourage members of the community to use the "contact us" function on the website to alert us to missing or incorrect annotations and to work with us on making sure their data are included in REDfly in a timely manner.

Funding: REDfly is supported by NIH grants U24 GM142435 and R01 GM114067 and NSF grant DBI-1758252. The Center for Computational Research is supported by NIH award S10OD024973 and NSF award 1724891.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are publicly available at http://redfly.ccr.buffalo.edu and as a permanent archive in the University at Buffalo Institutional Repository at http://hdl.handle.net/10477/82107.

Acknowledgments: We thank Kimberley Loguirato and the students and faculty from Baruch College BIO 4015 and 3005 for assistance with curation, Gil dos Santos and FlyBase for their cooperation and efforts, past and present members of the REDfly programming team, and the many members of the insect-research community who generate the data curated by REDfly. REDfly utilizes the staff and servers of the University at Buffalo Center for Computational Research (http://hdl.handle.net/10477/79221).

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

References

1. Adams, M.D.; Celniker, S.E.; Holt, R.A.; Evans, C.A.; Gocayne, J.D.; Amanatides, P.G.; Scherer, S.E.; Li, P.W.; Hoskins, R.A.; Galle, R.F.; et al. The genome sequence of Drosophila melanogaster. *Science* **2000**, *287*, 2185–2195. [CrossRef] [PubMed]
2. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **1998**, *282*, 2012–2018. [CrossRef] [PubMed]
3. Holt, R.A.; Subramanian, G.M.; Halpern, A.; Sutton, G.G.; Charlab, R.; Nusskern, D.R.; Wincker, P.; Clark, A.G.; Ribeiro, J.M.; Wides, R.; et al. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **2002**, *298*, 129–149. [CrossRef] [PubMed]
4. Waterston, R.H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J.F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*, 520–562. [CrossRef]
5. The Alliance of Genome Resources Consortium. The Alliance of Genome Resources: Building a Modern Data Ecosystem for Model Organism Databases. *Genetics* **2019**, *213*, 1189–1196. [CrossRef]
6. Grosveld, F.; van Staalduinen, J.; Stadhouders, R. Transcriptional Regulation by (Super)Enhancers: From Discovery to Mechanisms. *Annu. Rev. Genom. Hum. Genet.* **2021**, *22*, 127–146. [CrossRef]
7. Chen, D.; Lei, E.P. Function and regulation of chromatin insulators in dynamic genome organization. *Curr. Opin. Cell Biol.* **2019**, *58*, 61–68. [CrossRef]
8. Segert, J.A.; Gisselbrecht, S.S.; Bulyk, M.L. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends Genet.* **2021**, *37*, 514–527. [CrossRef]
9. Batut, P.J.; Bing, X.Y.; Sisco, Z.; Raimundo, J.; Levo, M.; Levine, M.S. Genome organization controls transcriptional dynamics during development. *Science* **2022**, *375*, 566–570. [CrossRef]
10. Kassis, J.A.; Brown, J.L. Polycomb group response elements in Drosophila and vertebrates. *Adv. Genet.* **2013**, *81*, 83–118. [CrossRef]
11. Atkinson, T.J.; Halfon, M.S. Regulation of Gene Expression in the Genomic Context. *Comput. Struct. Biotechnol. J.* **2014**, *9*, e201401001. [CrossRef] [PubMed]
12. Gallo, S.M.; Gerrard, D.T.; Miner, D.; Simich, M.; Des Soye, B.; Bergman, C.M.; Halfon, M.S. REDfly v3.0: Toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.* **2011**, *39*, D118–D123. [CrossRef] [PubMed]
13. Gallo, S.M.; Li, L.; Hu, Z.; Halfon, M.S. REDfly: A Regulatory Element Database for Drosophila. *Bioinformatics* **2006**, *22*, 381–383. [CrossRef] [PubMed]
14. Halfon, M.S.; Gallo, S.M.; Bergman, C.M. REDfly 2.0: An integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucl. Acids Res.* **2008**, *36*, D594–D598. [CrossRef] [PubMed]
15. Rivera, J.; Keranen, S.V.E.; Gallo, S.M.; Halfon, M.S. REDfly: The transcriptional regulatory element database for Drosophila. *Nucleic Acids Res.* **2018**, *47*, D828–D834. [CrossRef]
16. Abnizova, I.; te Boekhorst, R.; Walter, K.; Gilks, W.R. Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: The fluffy-tail test. *BMC Bioinform.* **2005**, *6*, 109. [CrossRef]
17. Arnone, M.I.; Davidson, E.H. The hardwiring of development: Organization and function of genomic regulatory systems. *Development* **1997**, *124*, 1851–1864. [CrossRef]
18. Lifanov, A.P.; Makeev, V.J.; Nazina, A.G.; Papatsenko, D.A. Homotypic regulatory clusters in Drosophila. *Genome Res.* **2003**, *13*, 579–588. [CrossRef]
19. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
20. 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]
21. Papatsenko, D.; Goltsev, Y.; Levine, M. Organization of developmental enhancers in the Drosophila embryo. *Nucleic Acids Res.* **2009**, *37*, 5665–5677. [CrossRef] [PubMed]
22. 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]
23. Soluri, I.V.; Zumerling, L.M.; Parra, O.A.P.; 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*, e53916. [CrossRef] [PubMed]
24. 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]

25. Halfon, M.S. Silencers, Enhancers, and the Multifunctional Regulatory Genome. *Trends Genet.* **2020**, *36*, 149–151. [CrossRef]
26. Erceg, J.; Pakozdi, T.; Marco-Ferreres, R.; Ghavi-Helm, Y.; Girardot, C.; Bracken, A.P.; Furlong, E.E.M. Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements. *Genes Dev.* **2017**, *31*, 590–602. [CrossRef]
27. 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]
28. 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]
29. 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]
30. 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]
31. 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]
32. Bonn, S.; Zinzen, R.P.; Girardot, C.; Gustafson, E.H.; Perez-Gonzalez, A.; Delhomme, N.; Ghavi-Helm, Y.; Wilczynski, B.; Riddell, A.; Furlong, E.E. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **2012**, *44*, 148–156. [CrossRef]
33. 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]
34. 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] [PubMed]
35. 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] [PubMed]
36. 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]
37. Ghavi-Helm, Y.; Klein, F.A.; Pakozdi, T.; Ciglar, L.; Noordermeer, D.; Huber, W.; Furlong, E.E.M. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **2014**, *512*, 96–100. [CrossRef]
38. Li, X.; Zhou, B.; Chen, L.; Gou, L.T.; Li, H.R.; Fu, X.D. GRID-seq reveals the global RNA-chromatin interactome. *Nat. Biotechnol.* **2017**, *35*, 940–950. [CrossRef]
39. 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]
40. 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]
41. 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]
42. Cusanovich, D.A.; Reddington, J.P.; Garfield, D.A.; Daza, R.M.; Aghamirzaie, D.; Marco-Ferreres, R.; Pliner, H.A.; Christiansen, L.; Qiu, X.J.; Steemers, F.J.; et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **2018**, *555*, 538–542. [CrossRef] [PubMed]
43. 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] [PubMed]
44. Kantorovitz, M.R.; Kazemian, M.; Kinston, S.; Miranda-Saavedra, D.; Zhu, Q.; Robinson, G.E.; Gottgens, 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]
45. 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]
46. 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]
47. 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]
48. Brody, T.; Yavatkar, A.S.; Kuzin, A.; Kundu, M.; Tyson, L.J.; Ross, J.; Lin, T.Y.; Lee, C.H.; Awasaki, T.; Lee, T.; et al. Use of a Drosophila genome-wide conserved sequence database to identify functionally related cis-regulatory enhancers. *Dev. Dyn.* **2012**, *241*, 169–189. [CrossRef]
49. 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]
50. Guo, H.T.; Huo, H.W.; Yu, Q. SMCis: An Effective Algorithm for Discovery of Cis-Regulatory Modules. *PLoS ONE* **2016**, *11*, e0162968. [CrossRef]

51. Asma, H.; Halfon, M.S. Computational enhancer prediction: Evaluation and improvements. *BMC Bioinform.* **2019**, *20*, 174. [CrossRef] [PubMed]
52. Su, J.; Teichmann, S.A.; Down, T.A. Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol.* **2010**, *6*, e1001020. [CrossRef] [PubMed]
53. Kazemian, M.; Suryamohan, K.; Chen, J.Y.; Zhang, Y.; Samee, M.A.; 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] [PubMed]
54. Asma, H.; Halfon, M.S. Annotating the Insect Regulatory Genome. *Insects* **2021**, *12*, 591. [CrossRef] [PubMed]
55. Clark, A.G.; Eisen, M.B.; Smith, D.R.; Bergman, C.M.; Oliver, B.; Markow, T.A.; Kaufman, T.C.; Kellis, M.; Gelbart, W.; Iyer, V.N.; et al. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **2007**, *450*, 203–218. [CrossRef]
56. 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]
57. 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]
58. 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]
59. 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]
60. 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]
61. Khoueiry, P.; Girardot, C.; Ciglar, L.; Peng, P.C.; Gustafson, E.H.; Sinha, S.; Furlong, E.E.M. Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *Elife* **2017**, *6*, e28440. [CrossRef] [PubMed]
62. Wang, X.F.; Zhou, T.Y.; Wunderlich, Z.; Maurano, M.T.; DePace, A.H.; Nuzhdin, S.V.; Rohs, R. Analysis of Genetic Variation Indicates DNA Shape Involvement in Purifying Selection. *Mol. Biol. Evol.* **2018**, *35*, 1958–1967. [CrossRef] [PubMed]
63. Peng, P.C.; Khoueiry, P.; Girardot, C.; Reddington, J.P.; Garfield, D.A.; Furlong, E.E.M.; Sinha, S. The Role of Chromatin Accessibility in cis-Regulatory Evolution. *Genome Biol. Evol.* **2019**, *11*, 1813–1828. [CrossRef] [PubMed]
64. Benton, M.L.; Talipineni, S.C.; Kostka, D.; Capra, J.A. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genom.* **2019**, *20*, 511. [CrossRef]
65. Halfon, M.S. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends Genet.* **2019**, *35*, 93–103. [CrossRef] [PubMed]
66. Lindhorst, D.; Halfon, M.S. Reporter gene assays and chromatin-level assays define substantially non-overlapping sets of enhancer sequences. *bioRxiv* **2022**. [CrossRef]
67. Costa, M.; Reeve, S.; Grumbling, G.; Osumi-Sutherland, D. The Drosophila anatomy ontology. *J. Biomed. Semant.* **2013**, *4*, 32. [CrossRef]
68. 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]
69. Gremke, L.; Lord, P.C.; Sabacan, L.; Lin, S.C.; Wohlwill, A.; Storti, R.V. Coordinate regulation of Drosophila tropomyosin gene expression is controlled by multiple muscle-type-specific positive and negative enhancer elements. *Dev. Biol.* **1993**, *159*, 513–527. [CrossRef]
70. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **2002**, *12*, 656–664.
71. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006. [CrossRef] [PubMed]
72. 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] [PubMed]
73. 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]
74. Larkin, A.; Marygold, S.J.; Antonazzo, G.; Attrill, H.; Dos Santos, G.; Garapati, P.V.; Goodman, J.L.; Gramates, L.S.; Millburn, G.; Strelets, V.B.; et al. FlyBase: Updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res.* **2021**, *49*, D899–D907. [CrossRef] [PubMed]
75. Lyne, R.; Smith, R.; Rutherford, K.; Wakeling, M.; Varley, A.; Guillier, F.; Janssens, H.; Ji, W.; McLaren, P.; North, P.; et al. FlyMine: An integrated database for Drosophila and Anopheles genomics. *Genome Biol.* **2007**, *8*, R129. [CrossRef] [PubMed]