Chapter 6
Insect Regulatory Genomics

Kushal Suryamohan and Marc S. Halfon

Abstract Insects are the most diverse and ecologically important group of animals in the animal kingdom, with more than a million species described to date. Whole-genome sequencing, which has revolutionized many areas of biological research, carries significant potential for achieving a deeper understanding of insect development, physiology, and evolution and for facilitating new biotechnological advances in insect management and biocontrol. Comprehensive genome annotation, including not only genes but also regulatory regions, is necessary for realizing the full benefits of this sequencing. However, regulatory element discovery in non-model organisms remains a major challenge as most regulatory sequences have diverged past the point of recognition by standard sequence alignment methods, even for relatively closely related species such as flies and mosquitoes. We review here some of the advances made in insect regulatory genomics and the methods and resources available for identifying regulatory elements in well-studied model insects such as *Drosophila*. We discuss recent efforts to extend these approaches to discovering

K. Suryamohan
Department of Biochemistry, University at Buffalo-State University of New York, 701 Ellicott St, Buffalo, NY 14203, USA

NY State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY 14203, USA
e-mail: kushalsuryamohan@gmail.com

M.S. Halfon (✉)
Department of Biochemistry, University at Buffalo-State University of New York, 701 Ellicott St, Buffalo, NY 14203, USA

Department of Biological Sciences, University at Buffalo-State University of New York, 701 Ellicott St, Buffalo, NY 14203, USA

Department of Biomedical Informatics, University at Buffalo-State University of New York, 701 Ellicott St, Buffalo, NY 14203, USA

Program in Genetics, Genomics and Bioinformatics, University at Buffalo-State University of New York, 701 Ellicott St, Buffalo, NY 14203-1101, USA

NY State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY 14203, USA

Molecular and Cellular Biology Department and Program in Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA
e-mail: mshalfon@buffalo.edu

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*, Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_6

regulatory elements in evolutionarily diverged non-model species and potential applications of the resulting regulatory data.

Abbreviations

B1H	Bacterial one-hybrid
Cas9	CRISPR-associated protein 9
ChIP	Chromatin immunoprecipitation
ChIP–chip	Chromatin immunoprecipitation combined with genome-tiling microarrays
ChIP-seq	Chromatin immunoprecipitation combined with next-generation sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
CRM	*cis*-regulatory module
DNase-seq	DNase I digestion combined with sequencing
DPE	Downstream promoter element
dsRNA	Double-stranded RNA
FACS	Fluorescently activated cell sorting
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements
GFP	Green fluorescent protein
GMOD	Generic Model Organism Database
GTF	General transcription factor
MOD	Model organism database
NCBI	National Center for Biotechnology Information
PBM	Protein-binding microarray
PWM	Position weight matrix
RNA-seq	RNA sequencing
RNAi	RNA interference
STARR-seq	Self-transcribing active regulatory region sequencing
TALENS	Transcription activator-like effector nucleases
TF	Transcription factor
TFBS	Transcription factor binding site
ZFN	Zinc finger nuclease

6.1 The Importance of Regulatory DNA: Why Regulate Genes?

The expression of metazoan protein-coding genes is regulated at several steps in the pathway from DNA to protein, including transcription of DNA to mRNA; mRNA stability, transport, processing, and translation; and posttranslational protein

modification. Such stratified control allows cells exquisite control over which proteins they make, and this confers distinct properties to cells, resulting in cell differentiation and diversity.

The main mechanism by which control of gene expression is achieved is transcriptional regulation. Although a promoter is necessary to initiate gene transcription, a significant part of eukaryotic transcriptional regulation is mediated by distal *cis*-regulatory modules (CRMs), of which the most common forms are known as enhancers: clusters of transcription factor binding sites (TFBS) that act without regard to orientation, distance, or location (up- or downstream) relative to the transcribed gene [1]. Regulation of gene expression is also achieved by additional distal *cis*-acting regulatory elements that include silencers, insulators, and locus control regions.

Next-generation DNA sequencing technologies now enable us to sequence the genomes of many organisms in their entirety relatively rapidly and at constantly decreasing cost. These technological developments have made possible numerous insect genome projects, many of which have now been completed and many more of which are anticipated: the i5K project aims to sequence 5000 insect and other arthropod genomes over the next 5 years [2]. The sequence of a genome, however, is of limited use without its annotation. That is, in addition to the DNA sequence, it is necessary to attach biological information to a genome, including not only identifying protein-coding genes and their coding exons but also defining non-protein-coding genes and, crucially, the different aforementioned regulatory elements—and then assigning function to each.

Historically, annotation of regulatory regions has been a challenge even in well-studied model organisms due to the inherent difficulties involved in regulatory sequence identification. In non-model organisms, where there are few experimental genetic and molecular data, where little is known about most transcription factors and their target binding sites, and where the ability to make transgenic animals is severely limited, genome annotation is particularly difficult. However, the extensive advances achieved by virtue of *Drosophila*'s position as a leading model organism have laid the foundation for molecular and computational tools to study other arthropods. Thus, annotating regulatory regions in other insect species is now becoming a realistic task, one that is essential to understanding the development and physiology of insects and which carries the potential to enable the development of products and techniques to control the harmful aspects of insects on society, as well as to harness the many benefits we derive from them. In this chapter, we discuss salient studies in insect regulatory genomics, focusing mainly on methods developed to identify regulatory elements, and highlight a few key studies on regulatory elements in insects.

6.2 Regulatory Genomic Analysis in Insects

6.2.1 Gene Function

Currently, there are about 200 insect species whose genomes have been sequenced or have begun to be sequenced. Details of each sequencing project and genome data are available at the National Center for Biotechnology Information (NCBI) Entrez Genome Project webpage (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi) and the i5k website (http://www.arthropodgenomes.org/wiki/i5K). Finished insect genome projects include *Drosophila melanogaster* as well as 19 other *Drosophila* species (http://www.flybase.org, ftp://ftp.hgsc.bcm.edu/Dmelanogaster/); medically important species that are vectors for diseases such as malaria (*Anopheles gambiae*), dengue fever (*Aedes aegypti*), and elephantiasis (*Culex pipiens*); agriculturally important species such as the honeybee (*Apis mellifera*) and silkworm (*Bombyx mori*); pests such as the red flour beetle (*Tribolium castaneum*) and the pea aphid (*Acyrthosiphon pisum*) (see Chaps. 4 and 5, in this volume); the parasitoid jewel wasp (*Nasonia vitripennis*); several species of ants and butterflies (see Chap. 3, in this volume); and others. The availability of the sequenced genomes of these insects, combined with the efforts of a diverse group of researchers, has dramatically improved the molecular and genetic tools available to study them with the consequence that many of these are now considered model or emerging-model research organisms.

While a significant proportion of the genes so far identified in insect genomes have been assigned a known or putative function (largely through homology to known genes in *Drosophila* and other organisms), many genes have yet to reveal their function. Fortunately, recent years have seen a surge in methods for the study of formerly non-model insect species, including gene knockdowns by RNA interference and genome engineering using transcription activator-like effector nucleases (TALENs) [3] or the clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9 (CRISPR/Cas9) system [4]. The discovery of transposons such as *piggyBac* [5] and *Hermes* [6] in the moth *Trichoplusia ni* and the house fly *Musca domestica*, respectively, has now made it possible to perform gene perturbation studies in a wide range of insects through the development of transgenic technology (see Sect. 6.3) [7–12].

As with many technologies, the use of RNA interference (RNAi) in insects was pioneered in *Drosophila* [13], but this powerful method was rapidly applied to the red flour beetle *T. castaneum* and many other holometabolous insect species [14, 15], including the malaria vector mosquito *Anopheles gambiae* [16, 17]. The development of parental RNAi techniques [18] meant that gene function in embryos could be disrupted by injecting double-stranded RNA (dsRNA) into pupal or adult females [19], allowing for studies of early insect development [20]. The immense utility of RNAi was also realized when it was used to study *Hox* gene function for the first time in a hemimetabolous species, the bug *Oncopeltus fasciatus* [21]. For an overview of successful applications of RNAi technology to assign functions to genes in various insects, readers are referred to the excellent review article by Xavier Belles [22].

With the recent development of methods such as zinc finger nucleases (ZFNs), TALENs, and CRISPR/Cas9 [3, 23–25], genome editing has become feasible in a broad range of species [26, 27], resulting in a deluge of papers despite the first reported uses of CRISPR taking place only a year ago. While as usual the first applications of CRISPR/Cas9 technology in insects have been in *Drosophila* [24, 25, 28], its success there has now encouraged its use in other insect species [29], and this powerful system is likely to revolutionize experimental studies in model and non-model insects alike. With these new technologies in hand, it is only a matter of time before we have a better understanding of the functions of genes crucial to development, vector biology, and other biological processes in most insect species.

6.2.2 Discovering DNA Regulatory Elements in the Genome

The task of finding regulatory elements in the genome has historically been a challenging one. Approaches can be classified into two broad categories: empirical and computational. Empirical approaches (Fig. 6.1) have traditionally been time-consuming, labor-intensive, and expensive. The genomic era has brought about the development of genome-wide, high-throughput assays and techniques, which has greatly accelerated the pace of regulatory element discovery. However, these methods are not without limitations: they can remain very costly, are often difficult to validate, and typically do not produce comprehensive or fully accurate results. This is a particular problem for *cis*-regulatory modules (CRMs), which may be functional only in certain cell types or under specific conditions.

Computational methods (Fig. 6.2) have provided an attractive complementary approach for regulatory element identification. However, these methods too have drawbacks, including high false-positive prediction rates and the challenges of large-scale empirical validation. Despite this, significant advances have been made in the computational methods for modeling and detection of DNA regulatory elements over the last decade. The availability of complete genome sequences for multiple organisms, whole-transcriptome profiles, high-throughput experimental methods for mapping protein-binding sites in DNA, increased throughput in empirical identification of CRMs, elucidation of higher order structures of the regulatory sequences [30, 31], and more efficient assays for testing putative regulatory regions have all contributed to the development of successful methods. Nevertheless, these approaches have primarily been limited to a few well-understood model organisms and biological systems, where a fair amount of prior knowledge is available, where the organisms are amenable to experimental manipulation, and where there is a large community-driven funding base.

In this section, we will briefly summarize the different empirical and computational approaches used for regulatory element discovery with particular focus on the identification of regulatory elements in insect species (including *D. melanogaster*).

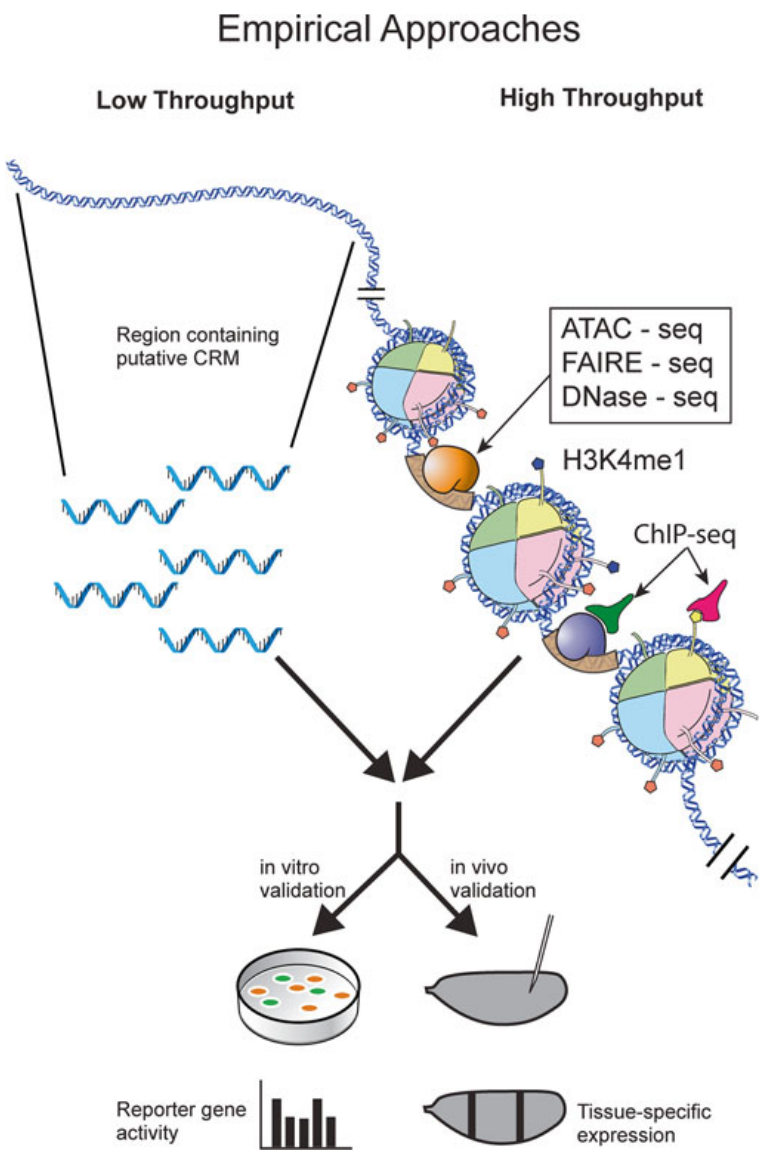

Fig. 6.1 Empirical approaches to CRM discovery. Empirical approaches to CRM discovery can be broadly classified into *low-throughput* (*left*) and *high-throughput* (*right*) methods. In both cases, the putative CRMs are tested in a heterologous reporter system. Low-throughput methods involve testing of isolated regions of DNA that contain putative CRMs in a cell culture or transgenic animal setting (or both). In the former, putative CRMs are transfected into cultured cells and reporter gene activity (e.g., luciferase, GFP) levels are quantified relative to a control vector. In the latter experiment, transgenic animals (here, flies) bearing a reporter gene construct are generated and assayed for tissue-specific expression patterns driven by the putative CRM. High-throughput methods make use of next-generation sequencing to identify potential regions of regulatory DNA. Chromatin immunoprecipitation-based methods use antibodies to detect binding of TFs of interest genome-wide followed by sequencing to identify the bound regions (ChIP-seq). A variant of this uses antibodies against specific chromatin modifications, such as histone methylation (e.g., H3K4me1), that are characteristic of regulatory sequences. A third variant makes use of the fact that regulatory regions have an "open" chromatin configuration, i.e., are depleted of nucleosomes or otherwise more accessible to cleavage by DNase I (DNase-seq) or to transposon insertion (ATAC-seq), or respond differently to chemical fractionation (FAIRE-seq). Additional methods are discussed in the text. Like predictions from low-throughput approaches, results from high-throughput experiments can also be tested using cell culture or transgenic methods, although typically only a fraction of the predictions can be validated

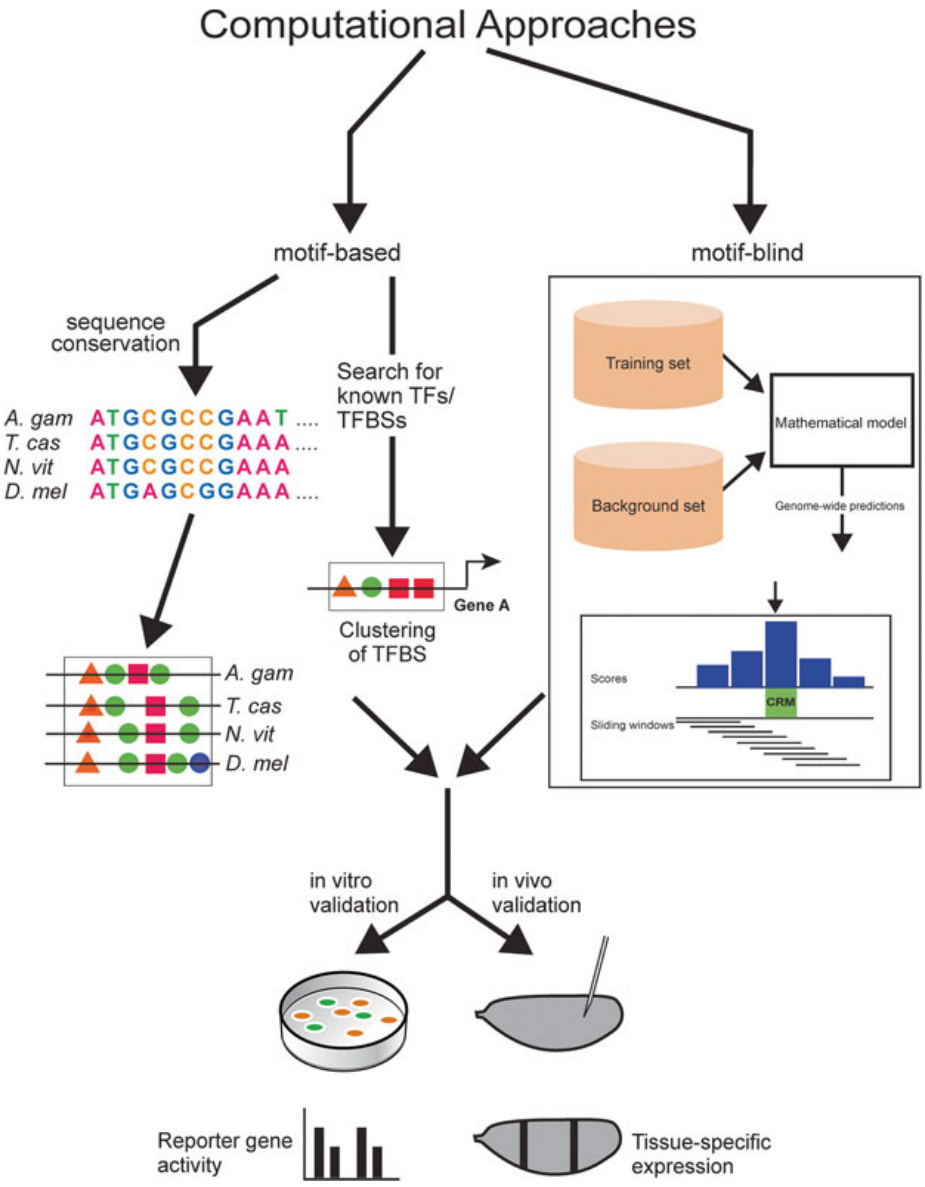

Fig. 6.2 Computational approaches to CRM discovery. Computational approaches can be broadly classified into *motif-based* (*left*) and *motif-blind* (*right*) methods. In the former, all observed instances of a TFBS (which are usually short motifs of ~6–10 bp) are modeled into a position weight matrix (PWM – see Sect. 6.2.2.2). Motif-based methods are predicated upon the knowledge that CRMs consist of clusters of TFBSs in a small region of DNA; these clusters can be searched genome wide (*center*). *Sequence conservation-based* methods (*left*) look for evolutionarily constrained regions of noncoding DNA containing clusters of TFBSs across several closely or distantly related species. The colored triangles, squares, and circles each represent a specific instance of a particular TFBS in a segment of DNA. Motif-blind methods (*right*) are unique in that they do not rely on existing knowledge of TFBSs or TFs and instead make use of the statistical profiles of experimentally validated CRMs (the "training set") against a set of non-CRMs (the "background set"). A statistical model is then used to scan the whole genome of a candidate species using overlapping windows with a score assigned to each window; the highest peaks in the resulting score profile are predicted to be CRMs. As with empirical approaches, predictions from computational methods can then be tested using a variety of cell culture or transgenic validation methods

6.2.2.1 Empirical Discovery of Regulatory Regions

Promoters

Initiation of transcription is achieved by the promoter, which can be viewed as consisting of a core promoter along with a variable number of proximal promoter elements (for variants of promoter architecture, refer to [32]). Together, these regions integrate regulatory inputs and initiate gene transcription. The core promoter consists of TFBSs for general transcription factors (GTFs) necessary to recruit RNA polymerase II (reviewed in [33]) and is typically defined as the ~40 bp region on either side of the transcriptional start site of its gene. While a number of core promoter binding motifs have been defined (e.g., the familiar TATA box and the downstream promoter element (DPE) [34]), there are no universal motifs common to all promoters, and the majority of promoters do not appear to contain any of the well-characterized motifs [35].

In the last decade, several high-throughput next-generation sequencing-based methods have been developed to aid promoter identification, including capture and sequencing of the 5′ ends of mRNA transcripts (CAGE-seq [36], PEAT [37], RAMPAGE [38]) and chromatin immunoprecipitation (ChIP)-based methods (e.g., ChIP of RNA Pol II) ([39–42]; see also Chap. 7, in this volume). In insects, genome-wide characterization of promoters has largely been restricted to *D. melanogaster*, an issue that needs to be addressed for other emerging-model or non-model insect species whose genomes have been sequenced [37, 38, 43, 44].

Cis-Regulatory Modules (CRMs)

Traditional CRM Discovery Methods

Whereas promoters can be identified through capture of 5′ mRNA sequences or RNA Pol II binding, and to a lesser extent by virtue of the presence of defined sequence motifs, discovery of distal CRMs presents a much greater challenge. Unlike promoters, CRMs do not contain broadly recognizable sequence characteristics and do not lend themselves to discovery via simple transcriptional profiling-based methods. Empirical approaches to discovering enhancers have historically involved isolating fragments of DNA containing putative CRMs and cloning them upstream of a minimal promoter fused to a reporter gene to test for transcriptional activity in cell lines or transgenic animals. Although more laborious and expensive to conduct than cell culture assays, transgenic animal studies have the great advantage of providing spatiotemporal expression information simultaneously in all tissues and cell types of an overall wild-type animal. Early empirical approaches were limited in the number of assays that could feasibly be performed. However, the more recent sequencing of the genomes of multiple species, along with the availability of next-generation sequencing strategies, has allowed for the development of higher-throughput methods for regulatory element identification in model organisms such as *Drosophila* and mouse, resulting in an explosion of newly predicted—and in many cases validated—CRMs.

Two efforts in *Drosophila* are notable for both their scope and audacity. Groups at the Howard Hughes Medical Institute's Janelia Farm Research Campus and at the Research Institute of Molecular Pathology in Vienna have taken a genome-tiling approach in which short overlapping segments of noncoding DNA are assayed in a more-or-less unbiased fashion using in vivo reporter gene assays. Collectively, these two groups have generated some 14,000 new reporter lines, increasing in the last few years by 5–7-fold the cumulative efforts of the preceding three decades [45, 46]. It should be noted, however, that many of the tested sequences are on the order of 2–3 kb and as such may contain multiple CRMs (which are frequently less than 500 bp in length). Thus, precise mapping of individual regulatory elements may still require substantial follow-up.

TFBS Discovery

Although such massive undertakings seem an unlikely prospect for extension to other insect species, the rise of microarrays and next-generation sequencing has spawned a growing number of high-throughput yet more broadly accessible methods for both TFBS and CRM discovery. Sensitive, unbiased methods to identify and characterize TFBSs in a systematic manner include SELEX-seq [47], protein-binding microarrays (PBMs) [48], and large-scale bacterial one-hybrid (B1H) assays. The latter are especially advantageous as they can determine the specificities of a TF of interest without requiring purification of the TF [49]. Chromatin immunoprecipitation (ChIP) coupled with genome-tiling microarrays (ChIP–chip [42]), now largely supplanted by next-generation sequencing (ChIP-seq) [39], enables genome-wide identification of regions bound in vivo by a given transcription factor (TF). Regions isolated from ChIP-based assays usually range in size from a few dozen base pairs to a few hundred base pairs. Since the regions obtained from ChIP are larger than the actual TFBSs themselves, additional computational analysis is needed to discover the individual TFBS within these regions. These limitations can be overcome with application of newer methods such as ChIP-exo (in which an exonuclease trims the DNA to give a higher resolution in TFBS mapping) [50] or extremely deep sequencing, which can reveal transcription factor binding sites 10–20 bp long ("digital footprinting") [51]. As always, it is worth bearing in mind the caveat that it cannot always be certain that all observed protein–DNA interactions have an active role in regulation. In at least some instances, substantial in vivo binding has been detected at sequences that do not appear to have regulatory function, and the number of sites bound by a TF can greatly exceed the number of genes the TF is believed to regulate [52–54]. Binding is also cell type specific, meaning that ChIP-based methods are most effective when applied to pure cell populations and provide more limited information when performed on complex tissues or whole embryos. Nevertheless, sufficient data to make reasonable inferences as to probable binding of a given TF at a given locus, through collective application of the discussed approaches, are likely within reach for the majority of TFs in *Drosophila* in the near future. Since TF binding domains have frequently evolved slowly overall [55], in many cases extrapolation to other insect species will also be possible.

CRM Discovery Using Epigenomic Methods

Active regulatory regions tend to be devoid of nucleosomes, a property that can be exploited for regulatory element discovery. Regions of nucleosome-depleted, or "open," chromatin can be identified on a genome-wide scale through methods such as DNase-seq [56], where accessible regions are detected by virtue of higher susceptibility to enzymatic cleavage by DNase I; FAIRE-seq (formaldehyde-assisted isolation of regulatory elements), which separates nucleosome-containing from nucleosome-free DNA using formaldehyde cross-linking followed by phenol extraction [57, 58]; or ATAC-seq [59], in which accessible chromatin is a preferential target for transposon tagging, allowing for direct sequencing of the tagged sequences after DNA isolation. ChIP-seq can also be used for genome-wide CRM discovery. For example, enhancer regions are often associated with the transcriptional cofactor p300/CBP and with components of the Mediator complex [60–62], and active enhancers are associated with specific histone modifications such as histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac), as well as depletion in H3 lysine 4 trimethylation (H3K4me3) [63, 62]. These methods all show great promise, although to date most have not yet produced detailed and well-defined sets of validated CRMs in the way that the traditional reporter gene assays (above) or newer functional assays (below) have done.

Function-Based Methods

The explosion in next-generation sequencing-based technologies has continued in the last few years with the development of new high-throughput function-based methods for enhancer discovery. STARR-seq can directly and quantitatively assess enhancer activity in millions of short sequences (on average ~600 bp in length) drawn from arbitrary sources of DNA to generate an unbiased survey of regulatory sequences active in a given cell line [64]. These sequences are inserted downstream of a minimal promoter and transfected into cells such that each sequence serves as its own reporter; the strength of each regulatory sequence is then assessed by its abundance in a subsequent RNA-seq analysis. When applied to the *Drosophila* genome, STARR-seq identified thousands of cell type-specific enhancers with differing activation strengths. Enhancer-FACS-seq is another method that was developed for identification of enhancers in *Drosophila*, where developmentally relevant, tissue-specific enhancers were detected within developing *Drosophila* embryos using a two-color FACS (fluorescently activated cell sorting)-based filtering: one color is used to register reporter gene activity and the other to mark cell types of interest [65]. This is an innovative method in that it eliminates the initial need to screen individual enhancer constructs in transgenic animals and allows for simultaneous testing of multiple pooled putative regulatory sequences, although full characterization of identified CRMs still requires subsequent generation of a new transgenic line. FIREWACh (Functional Identification of Regulatory Elements Within Active Chromatin) [66] and SIF-seq (site-specific integration fluorescence-activated cell sorting followed by sequencing) [67] also identify regulatory elements

by monitoring activity during initial screening assays using FACS sorting. Although they have not to date been applied to insect models, nothing specifically precludes their use for this purpose.

While these methods are elegant and high-throughput and demonstrate successful CRM discovery, they do have limitations. In particular, with respect to insect regulatory genomics, each depends either on the availability of a reasonable selection of cell lines or on the capacity to generate transgenic animals in an efficient and scalable manner—capabilities that for the most part are absent for insect species other than *D. melanogaster*.

Generality of Assays and Results

Although genome-wide maps of accessible chromatin, epigenetic marks, TF binding, and even regulatory function serve as a useful starting point, a significant challenge remains in that many regulatory regions function only in specific cell types and thus can only be identified when assays are performed using those cells. Each of these features must therefore be assessed in multiple tissues over many developmental time points and/or under varying environmental conditions in order to achieve comprehensive CRM discovery. This is a difficult goal for a variety of reasons, not least of which is financial, as well as obtaining sufficiently large homogeneous pools of each cell type at different time points and, more importantly, addressing depth of coverage in terms of the number of TFs and histone modifications to assay. These issues are especially acute in studying insects, which are anatomically small, thereby making it hard to isolate specific tissues in adequate amounts. In this regard, it is encouraging that DNaseI-seq at least appears to be reasonably robust in the sense that open chromatin regions are detected even when present in a limited fraction of overall embryonic cells [68]. Moreover, given the rate of technological progress, many of the logistical hurdles may soon be overcome as assays for small numbers of or even single cells are perfected [69, 70], and these methods will continue to aid in painting a more complete picture of the regulatory landscape of many cell types.

Assigning CRMs to Target Promoters

Once a CRM is identified, a major hurdle still often lies in assigning it to the appropriate target gene (or genes). Although many studies use "the closest active gene" theory to assign target genes, this clearly does not always lead to accurate assignment. Genes can lie hundreds of kilobases away from their cognate enhancers, and there can even be additional, separately regulated genes lying between a CRM-promoter pair. High-throughput versions of chromosome conformation capture technologies yield three-dimensional interaction maps that are providing exciting new insights into how distal CRMs interact with target promoters and can aid in CRM target gene assignment [30, 41, 71, 72], although these assays are technically challenging and artifact prone. Computational methods that make use of multiple

sources of more readily available data—histone modifications, RNA-seq, sequence conservation, etc.—will also be a valuable aid for determining CRM targets [72].

6.2.2.2 Computational Approaches to CRM Discovery

Even with the current trend of decreasing costs for empirical high-throughput experiments, the methods discussed in the preceding section remain prohibitively expensive and technically challenging for many emerging/non-model organisms, especially if considering extensive assaying of the genome under multiple conditions or at many developmental stages. Computational methods provide an attractive complement to experimental approaches and can often precede them as a first step in identifying regulatory regions, to be followed later by in vivo validation. Computational analysis can also help to refine or increase the predictive power of results obtained by empirical assays. In many cases, when working with non-model organisms with limited amenability to molecular genetic approaches, these methods may be essential for successful discovery and understanding of transcriptional regulatory elements.

Computational methods for CRM discovery can be broadly classified into three major categories: (a) comparative genomics, based on searching for regions of conserved noncoding DNA sequences across related species; (b) motif-based methods, which search for short genomic regions containing clusters of transcription factor binding sites; and (c) "motif-blind" approaches, which require no a priori knowledge of TFs or TFBSs.

Comparative Genomic Approaches

Comparative genomic approaches look for regions in the genome that are conserved between species. The underlying assumption is that there is likely to be a high degree of conservation of functionally important sequence elements (both coding and noncoding) between related species, an assumption that has frequently, although not universally, been shown to be true (e.g., [73] and references therein). There is mixed evidence as to whether or not attempting to discriminate CRMs from non-CRMs based solely on sequence conservation is effective. Li et al. [74] showed that while in the aggregate CRMs are more highly conserved, comparison among eight sequenced drosophilids gave poor predictive value for any particular sequence when assessing overall percentage of conserved bases. However, a more recent study found that reasonable discriminative performance could be achieved on a similar set of CRMs using a windowed version of the PhastCons conservation score (although on other data sets, this method performed less well) [75].

Less important than overall conservation of CRM sequence appears to be the conservation of CRM content, i.e., maintenance of a similar complement of TFBSs, although the number and organization of these sites can vary widely [76, 77]. As a result, sequence conservation is more clearly of utility when mixed with identification

of TFBSs, either to reduce false-positive identification of bona fide binding sites or to predict CRMs based on TFBS composition. For instance, the specificity of motif-based CRM prediction (see following section) can be improved by restricting TFBS motif instances to those that are also conserved in other species [78, 79]. Many in vivo-bound TFBS motifs are conserved among *Drosophila* species and other insect species [80–83]. Regulatory regions have been identified in several *Drosophila* genomes [80, 84, 85] as well as in other dipterans, including the malaria mosquito *An. gambiae*, the distant drosophilid *Scaptodrosophila lebanonensis*, and the fly *Calliphora vicina* [86–88], by looking for conservation of validated or predicted TFBSs when compared against the *D. melanogaster* genome. The enhancers from various species identified in this manner function as expected when tested in transgenic *Drosophila* [89–93]. Nevertheless, care must be taken when imputing function based on overall conservation of either sequence or binding site content. Studies of CRM evolution have revealed large-scale turnover of TFBSs despite the CRMs having maintained their function across multiple species of *Drosophila* [94, 95]. A landmark study by Ludwig et al. demonstrated that the *eve_stripe2* CRMs from *D. melanogaster* and *D. pseudoobscura*, which show clear sequence conservation as well as conservation of function, are completely nonfunctional as a chimera consisting of the 5′ half of one CRM and the 3′ half of the other [96]. Thus, TFBS turnover and compensatory evolutionary adaptations in the individual CRMs play a significant role in shaping their respective functions despite overall sequence-level conservation. Moreover, extensive enhancer mutagenesis has shown that simple scrambling of a CRM sequence can confer new tissue specificity to its output, and minor changes in motif positioning can affect CRM function in a tissue-specific manner [97, 98]. Merely possessing the same TFBSs, therefore, does not guarantee conservation of CRM function.

A significant limitation to sequence conservation as a means of CRM discovery, especially within the insects, is that noncoding sequences have evolved rapidly. Indeed, even within the Diptera, regulatory sequences have frequently diverged beyond the point of recognition by standard alignment methods ([96, 99–102] and M. Kazemian, S. Sinha, K.S and M.S.H., unpublished data). Moreover, sequence conservation cannot not reveal lineage-specific, recently evolved CRMs. Nevertheless, given the generality of the methods and the lack of need for any a priori knowledge of TFBS or TFs, comparative genomic approaches will remain a useful tool—at least for closely related species—for identification of putative CRMs.

Motif-Based CRM Discovery

In essence, CRMs are composed of a set of specific TFBSs spread over up to a few hundred nucleotides [103]. When these TFBSs are known or can be inferred, motif-based approaches for predicting enhancers and promoters can be applied. These approaches predict CRMs based on their DNA sequence and searchable representations of the TFBSs. Most typically, TFBSs are modeled in the form of position weight matrices (PWMs) [104], although alternate representations such as

degenerate consensus sequences or hidden Markov models have also been used [105, 106].

Motif-based CRM discovery was first conducted in mammals in the late 1990s in seminal work by Wasserman and Fickett [107]. Prior knowledge of the transcription factors that regulate expression of genes controlling muscle development, and their cognate TFBS motifs, was successfully used to look for clusters of those TFBS motifs elsewhere in the human genome. This approach was seized on by *Drosophila* researchers upon publication of the fly genome in early 2000 [108–113]. The extensive existing knowledge of early developmental CRMs—e.g., the "stripe" enhancers of the pair-rule genes—provided a rich set of TFBS motifs as well as a validation enhancer set to gauge sensitivity, and the tendency toward homotypic clustering of TFBSs within these CRMs allowed for simple "motif clustering" algorithms to be successful. All of these analyses found at least one novel enhancer that was active in transgenic flies, but in general suffered from low predictive power. A subsequent generation of algorithms incorporated probabilistic searching and sequence conservation between related species [114, 115], which helped to reduce false-positive rates; however, false-positive results continue to plague most motif-based CRM discovery methods, which are consistently outperformed in head-to-head comparisons of methods [75, 116, 117]. The high false-positive rates are likely a consequence of several factors, one of the largest being the fact that TFBS prediction itself is highly error-prone [118]. TFBS motifs are degenerate, and our knowledge of the full range of sequences capable of being bound by a given TF is usually incomplete, especially with respect to *in vivo* versus *in vitro* binding.

It is worth noting that motif-based methods rely on an important biological assumption: that genes that are expressed in a similar pattern are regulated by a similar complement of TFs. While no doubt this does not hold universally, the fact that these methods consistently work—albeit with high false-positive rates—supports the assumption. Nor is this confined only to the presence of highly tissue-specific TFs, broadly general expression patterns, or highly clustered binding sites. For instance, Halfon et al. [112] demonstrated that motif-based searching could identify CRMs driving a tightly restricted expression pattern in a small subset of cells and regulated by a combination of widely expressed TFs, some of which bound only once or twice in the CRM. On the other hand, while it is clear that identification of (usually conserved) TFBSs can aid in CRM discovery, caution must be taken in ascribing functional roles to each of these sites and/or to their cognate TFs. Previous studies have shown that not all motifs used as input for successful CRM discovery algorithms are functional components of the identified CRMs, and not all important TFBSs are conserved [119, 120].

A different flavor of CRM discovery moves away from clustering of a specific set of TFBSs toward a model of CRM evolution via gain and loss of binding sites. These methods attempt to develop mathematical models that capture the TFBS signatures characteristic of CRMs without assuming direct sequence-level conservation. MorphMS is one such modeling method which identifies candidate CRMs using a pairwise probabilistic alignment method that fits an evolutionary model derived from a set of existing TFBS motifs; it was found to have the best performance

for recovering known *D. melanogaster* CRMs in a comparison of computational approaches [75, 121]. EMMA, an improvement on MorphMS by the same authors, models the evolution of binding sites and allows binding sites to occur in only one species, but not the other (note that both tools construct pairwise alignments) [122]. A similar approach, to account for gain and loss of binding sites, is taken by Majoros and Ohler [123], although its computational complexity precludes it from being implemented on a genome-wide scale. These approaches provide important insights into the potential roles of TFBS turnover in CRM evolution. However, as CRM discovery methods, they still suffer from the requirement of needing knowledge of relevant TFBSs to be effective.

Motif-Blind Approaches

What does one do, then, when not all TFs and/or TFBSs (or sometimes not even one relevant TF and/or its TFBS) are known a priori—the most common situation? In such an event, it becomes impossible to search for CRMs using motif-based methods, and it is necessary to turn to methods that are not limited by current knowledge of TFBS motifs or of the TFs involved in regulating genes of interest. This becomes especially crucial for annotating the genomes of non-model organisms (such as most insect species) where such data are severely lacking.

One approach that has been used is to employ motif discovery and CRM discovery in tandem. An example of this is CisModule, which uses a Bayesian model to simultaneously predict TFBS motifs and CRMs [124]. CisModule showed good specificity in both simulated and applied tests, particularly for its motif-finding phase. However, in other settings, it performed less well for de novo CRM discovery than motif-blind (see below) methods that do not rely on first predicting TFBSs [116].

Better success has been achieved using supervised machine learning methods which search for patterns that can distinguish a training set composed of known CRMs from non-regulatory DNA, using only the DNA sequence itself as input [117, 125, 126]. These methods capture the statistical features inherent in each CRM within the training set without requiring other information, such as TFBSs or TFs, a priori. The genome can then be searched for additional sequence windows containing a similar statistical signature. Kantorovitz et al. [117] first dubbed such methods, which fall into the class of alignment-free sequence comparisons, "motif-blind," as TFBS motifs do not factor into the search algorithms.

One of the most successful examples of motif-blind approaches has come from a collaborative effort between the Sinha and Halfon groups, who in a series of papers have applied their methods to both the *Drosophila* and mouse genomes [117, 126]. This team has developed a computational pipeline, designated "SCRMshaw," that uses multiple machine learning algorithms to search for sequence "words" (i.e., short DNA subsequences) that are overrepresented in a training set of known CRMs (Fig. 6.3). These words (or "*k*-mers") serve as proxies for the unknown and un-modeled TFBSs, but TFBSs themselves, even when known, are not explicitly used by the algorithms. The training sets are constructed from a set of CRMs all

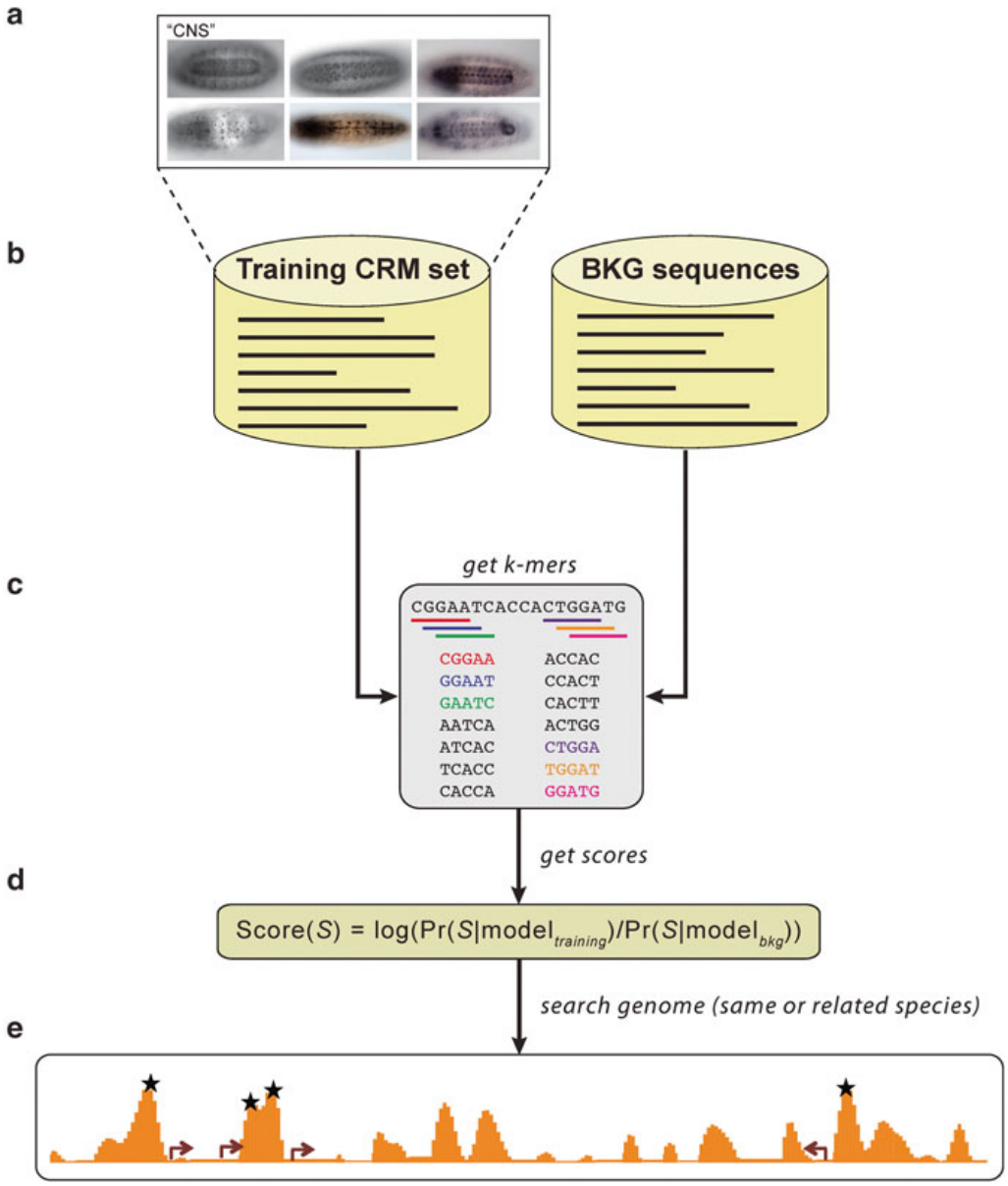

Fig. 6.3 Supervised motif-blind CRM discovery (SCRMshaw). (**a**) A set of CRMs with related activity (shown here, the *Drosophila* embryonic ventral nerve cord, part of the "CNS" CRM set) is selected as a training set. (**b**) The sequences of the training CRMs and of a set of similarly sized "background" non-CRMs (BKG) serve as input to the algorithm. The training set can also include orthologous sequences from related species. (**c**) The k-mer profile of the sequence sets is obtained and used to train one of the several statistical models. (**d**) The score for a given sequence S is the log-likelihood ratio of the models for the positive ("training") and negative ("background") sets on S. (**e**) Overlapping sequence windows are scored throughout the genome. High-scoring windows (*stars*) are predicted CRMs. The genome being searched can be from the same species as the training data (e.g., *Drosophila melanogaster*) or from a more distantly related insect species (e.g., *Apis mellifera*)

demonstrated to drive a related expression pattern and do not need to be large; sets as small as six known CRMs have provided successful CRM discovery results. Comparisons with motif-based methods for CRM sets where there is good knowledge of TFBSs—e.g., the *Drosophila* stripe enhancers referred to above—demonstrated that SCRMshaw consistently performed as good or better and was able to reach unprecedentedly high (80 % or better) success rates [117].

CRM Discovery in Insects Other Than *D. melanogaster*

The number of characterized non-*Drosophila* insect CRMs is small but growing. Computational methods based on motif clustering have proven effective in discovering CRMs in insects other than *Drosophila*, usually using PWMs derived from *Drosophila* TF binding studies and relying on the assumption that the binding sites for orthologous TFs would have similar sequences [101, 127–130]. These studies have mainly focused on well-described developmental systems, in particular early anterior–posterior and dorsal–ventral patterning, where there is extensive knowledge of TFs, their binding motifs, and similar *Drosophila* CRMs.

We have recently determined that *Drosophila* CRM training data can be used to apply the SCRMshaw method for motif-blind supervised CRM discovery to a broad range of holometabolous insects with success rates comparable to those obtained when conducting *Drosophila*-specific CRM discovery [131]. By using the same methods and training sets used for within-species CRM discovery [117, 126] but searching the genomes of *An. gambiae*, *T. castaneum*, *A. mellifera*, and *N. vitripennis* instead of that of *D. melanogaster* (Fig. 6.3), we were able to rapidly almost double the collective number of in vivo validated CRMs for these species and predict some 7000 more [131]. This is a significant advance given that the genomes of these species are highly diverged—substantially more so than human-to-fish for Diptera-to-Hymenoptera, for example [132]—to the point where alignment of noncoding sequences to the *Drosophila* genome is for the most part not possible. Successful application of supervised motif-blind CRM discovery therefore suggests that not only is regulatory sequence annotation in diverged insect species an attainable goal but also that it is one that can progress without requiring extensive new experimental data to be generated for each newly sequenced genome.

6.2.2.3 Database Resources for Insect Genomic Data

Biological databases are an essential part of any research project undertaken today. The ever-increasing amounts of data collected from biological experiments, especially high-throughput experiments such as genome sequencing and annotation, protein and gene interaction studies, protein structure determination, and the like, make these databases invaluable for managing information and making it easily accessible. Several dedicated databases have been developed for insect-specific research and are briefly reviewed below.

Model Organism Databases

Many of the insects that have been sequenced within the last decade now have dedicated model organism databases (MODs) [133–141]. The MODs constitute a tremendously valuable resource and serve as clearinghouses for much of the available data on the genetics and genomics of the covered model organisms. This is most evident for *Drosophila*, where FlyBase, one of the first MODs to be developed, maintains not just the genome annotation but also allele descriptions, gene expression pattern data, transcriptomic data, cytogenetic maps, and much of the other collected information from over a century of *Drosophila* research [133].

While the MODs are crucial for allowing researchers to access sequence data and genome annotations, a problem often encountered with species-specific databases is that of interoperability. Different interfaces and data formats make it complicated for users to move about through the different databases, and the databases include widely varying degrees of information on homologous sequences in other species, tools for pathway analysis, gene ontology annotations, protein domain annotation (e.g., InterPro), and functional pathway annotation (e.g., KEGG). In this regard, the Hymenoptera Genome Database stands out as a truly multispecies genome database for representatives of the over 115,000 insects in the Hymenopteran order [134]. Combining information on all these species into a single database provides an enormously useful resource for researchers interested in comparing and studying the Hymenoptera. The combination of numerous pest species into the AgripestBase (www.agripestbase.org) framework is another positive step in the direction of interoperability. As various species are becoming sequenced through the i5K project [2], many of the assemblies and early annotation are being housed through the National Agricultural Library's "i5K Workspace" (http://i5k.nal.usda.gov/), which provides a hosting framework for species not backed by a large, organized research community. The i5K Workspace and many of the MODs are built using components from the Generic Model Organism Database (GMOD) toolkit [142], a powerful resource for researchers who wish to provide bioinformatic tools for accessing whole-genome data. The use of GMOD components by a broad selection of MODs provides a familiar interface and a degree of interoperability for users of multiple genome databases. A holdout in this regard is VectorBase [143], which is built using the ENSEMBL framework rather than GMOD. Although this design choice has many positive features—the versatile and intuitive BioMart [144] is a particularly useful tool—it places VectorBase somewhat at odds with the other MODs and complicates cross-organism comparisons. Some of the more traditional model insect species (several *Drosophila*, *An. gambiae*, *A. mellifera*) can also be found in the UCSC Genome Browser [145], allowing access to the powerful tools, and integration with the many other genomes, covered by that major resource. Similarly, many insect species are also accessible via ENSEMBL (http://metazoa.ensembl.org/index.html). *Drosophila* and *An. gambiae* data can be found in FlyMine, a data warehouse with a powerful search interface that integrates genomic and proteomic data for these two species [146]. While these latter three databases offer the advantages of data integration and standard included tools, it should be

noted that the primary genome sequences and annotations are still imported from the MODs.

Gene Expression Resources

Several resources are devoted to gene expression data. The Berkeley *Drosophila* Genome Project (BDGP) contains genome-wide expression profiles of over 6000 genes in *D. melanogaster* embryos as determined by whole-mount in situ hybridization over all embryonic developmental stages and documented in over 70,000 images [147–149]. FlyExpress is another such resource that catalogues the spatial expression domains of over 4000 genes via a series of over 100,000 images and allows for pattern-based searching of the database [150]. FlyAtlas [151] provides transcriptional profiles for dissected *D. melanogaster* adult and larval tissues, and modENCODE [152] has produced time-course expression data for all stages of the fly life cycle as well as a limited number of dissected tissues. Many of these data are also mirrored in FlyBase. Many of the other MODs also include gene expression data, either from EST sequencing, microarray, or RNA-seq studies (see Table 6.1).

6.2.2.4 Regulatory DNA Element and Transcription Factor Databases

Resources related to insect gene regulation are primarily directed toward *Drosophila*, where the bulk of the existing work on regulatory element discovery has been performed. The most comprehensive regulatory genomics database available for insects—in fact, for any metazoan—is REDfly, the Regulatory Element Database for *Drosophila* [153]. REDfly is a highly curated portal for *Drosophila cis-regulatory* data containing records for empirically validated CRMs and TFBSs obtained from the published literature. This single searchable database of CRMs enables researchers to search for all experimentally verified fly regulatory elements along with their DNA sequence, their associated genes, and the expression patterns they direct (Fig. 6.4). REDfly serves as an important source of data for both validation and generation of hypotheses about gene regulation and has been particularly important for facilitating studies of CRM evolution and development of methods for CRM discovery.

The JASPAR [154] and TRANSFAC [155] databases are a major source of TFBS data, but although they contain TFBSs from *Drosophila*, they are not limited to insects, and most of their data are from vertebrate species. FlyFactorSurvey, on the other hand, contains *D. melanogaster* TF binding specificities as determined by bacterial one-hybrid assays, SELEX, or DNase I footprinting. The database contains PWMs associated with over 300 TFs and computational tools for identifying motifs within new candidate sequences [156]. The related Genome Surveyor [157] is a web-based tool for CRM discovery and analysis in a growing number of species; covered insect species include *D. melanogaster*, *Ae. aegypti*, *An. gambiae*, *N. vitripennis*, *A. mellifera*, and *T. castaneum*. Using the motifs contained within

Table 6.1 Database resources for insect regulatory genomics

Resources	Description	Species included (as of June 2015)	Link
AgripestBase	A comprehensive model organism database for agricultural pests	The Hessian fly, *Mayetiola destructor*; the tobacco hornworm, *Manduca sexta*; and the red flour beetle, *Tribolium castaneum*	http://agripestbase.org/
AphidBase	Model organism database	The pea aphid *Acyrthosiphon pisum*	http://www.aphidbase.com/
Berkeley Drosophila Genome Project in situ database	Contains genome-wide spatial expression profiles of 7917 genes during embryogenesis	*Drosophila melanogaster*	http://insitu.fruitfly.org/
Ensembl–Metazoa	A database for genomes of metazoan species including a number of insect species, with tools for querying and extracting features of each genome such as sequence variation, annotation, and protein homologies	Genomes of 17 dipteran, 4 hymenopteran, 4 lepidopteran, 2 coleopteran, 2 hemipteran, 1 isopteran, and 1 pthirapteran species	http://metazoa.ensembl.org/index.html
FlyAtlas 2	Transcriptional profiles of genes in multiple tissues at multiple larval through adult developmental stages	*D. melanogaster*	http://flyatlas.gla.ac.uk/flyatlas/index.html
FlyBase	Model organism database	Twelve species in the genus *Drosophila*	http://flybase.org/
FlyExpress	Digital library capturing the spatiotemporal expression patterns of thousands of genes during development. Can be used to match/search for specific expression patterns of interest	*D. melanogaster*	http://www.flyexpress.net/
FlyMine	An integrated resource for multiple types of genomic and proteomic data for *Drosophila* and *Anopheles*	*D. melanogaster*, *An. gambiae*	http://www.flymine.org/
FlyTF	An integrated database of data for *Drosophila* transcription factors	*D. melanogaster*	http://www.flytf.org/

(continued)

Table 6.1 (continued)

Resources	Description	Species included (as of June 2015)	Link
Genome Surveyor	A web-based tool for discovery and analysis of *cis-regulatory* elements in *Drosophila* and other organisms. Provides prediction and visualization of putative CRMs and TFBSs	*D. melanogaster*, *An. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum*	http://veda.cs.uiuc.edu/cgi-bin/gb2/gbrowse/Dmel5/
Hymenoptera Genome Database	Model organism database	Species in the order Hymenoptera including 3 *Nasonia* species, 4 bee species, and 8 ant species; also the genomes of 4 *Apis mellifera* pests and pathogens	http://hymenopteragenome.org
JASPAR	An excellent resource for a curated, nonredundant set of TF binding profiles, derived from published collections of experimentally defined transcription factor binding sites for several organisms; tools for querying DNA sequences of interest for instances of TFBS in the database	*D. melanogaster*; various noninsects	http://jaspar.genereg.net/
LocustDB	A transcriptomic database with a library of ESTs	The migratory locust *Locusta migratoria*	http://locustdb.genomics.org.cn/
modENCODE	Data access portal for genomic and epigenomic data from the modENCODE project	*D. melanogaster* and several other *Drosophila* species; *Caenorhabditis* species	http://www.modencode.org/
MyzusDB	A preliminary database resource with whole genome as well as comparative genome analyses	The green peach aphid *Myzus persicae*	http://www.aphidbase.com/node_94263/Myzus-DB

(continued)

Table 6.1 (continued)

Resources	Description	Species included (as of June 2015)	Link
ORegAnno	An open database that allows users to manually curate and annotate regulatory elements as well as visualize and access the annotated regulatory elements	*D. melanogaster* but mostly noninsect species	http://www.oreganno.org/oregano/
REDfly	Comprehensive database of over 5000 experimentally validated CRMs and TFBSs along with accompanying information such as expression patterns, sequences, and target genes	*D. melanogaster*	http://redfly.ccr.buffalo.edu/
SilkDB	Model organism database	*Bombyx mori*	http://www.silkdb.org/silkdb/
SpodoBase	Model organism database	*Spodoptera frugiperda* (fall army worm)	http://bioweb.ensam.inra.fr/spodobase/
TRANSFAC	A manually curated database of eukaryotic transcription factors, their genomic binding sites, and DNA binding profiles	*D. melanogaster*; many noninsects	http://www.gene-regulation.com/pub/databases.html
UCSC Genome Browser	A comprehensive resource for all genomic information as well as proteomic information for several model and non-model insect species	11 drosophilids, *An. gambiae*, and *A. mellifera*; many noninsects	https://genome.ucsc.edu/
VectorBase	Model organism database	19 *Anopheline* species, *Ae. aegypti*, *C. quinquefasciatus*, the Tsetse fly *G. morsitans*, many others	https://www.vectorbase.org/

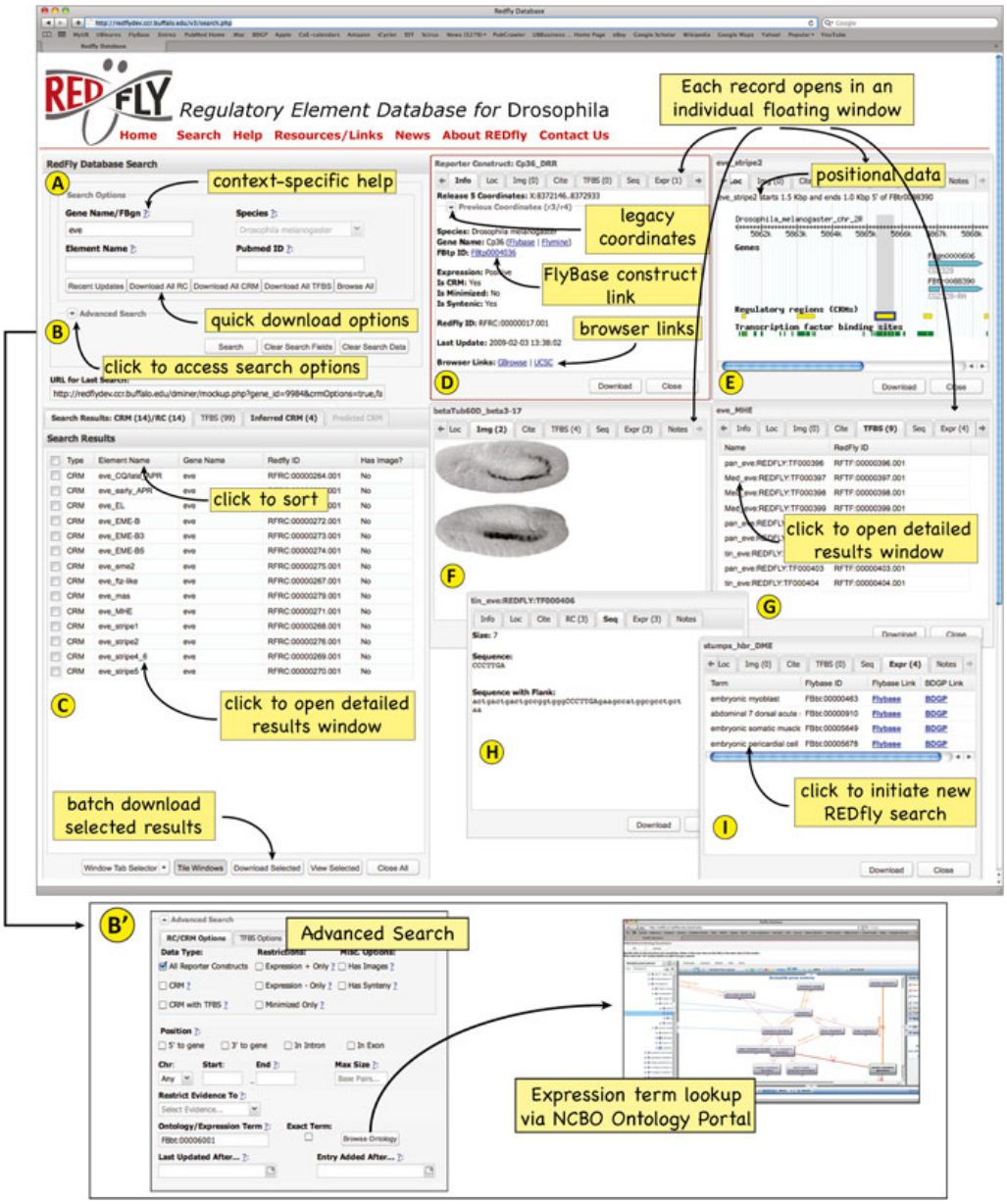

Fig. 6.4 The REDfly database REDfly is a comprehensive CRM and TFBS database for *Drosophila*. Search options (*A*, *B*), results overview (*C*), and detailed results (D–I) are all displayed within a single web browser window. Advanced search options (*B'*) include the ability to search based on ability of a tested genomic sequence to regulate/not regulate gene expression, position of a CRM relative to the transcription start site of the gene, and pattern of expression regulated by the CRM. For the latter function, an anatomy ontology browser can be used to select desired search terms (*right-hand panel*). The detailed results (*D–I*) are displayed as individual floating windows that can be stacked or tiled to facilitate comparison of multiple CRMs (Adapted from Gallo et al. [153])

FlyFactorSurvey, Genome Surveyor can predict TFBSs and CRMs in *Drosophila* using several different methods, including the supervised motif-blind CRM discovery method from Kantorovitz et al. [117]. FlyTF [158] allows for query-based retrieval of curated TFs for several *Drosophila* species that have been identified using different biological assays such as footprinting and chromatin interaction assays, and their target genes, although at present it is no longer being actively maintained. A newer resource for *D. melanogaster* TFs, OnTheFly, includes TFs, their binding sites, and annotation of their DNA-binding domains with structural properties and evolutionary homology [159].

The rapid accumulation of experimental data in the field of insect genomics highlights the need for databases that include interactive web-based computational analysis tools to simplify integration of different types of data such as genome-wide high-throughput genomic data, proteomic data, transcriptome data, and RNAi data with genome annotations for transcripts and regulatory elements. FlyMine does much of this for the two species it covers, but does not currently provide a home for additional insect species. REDfly would be a natural repository for regulatory-specific data from across the Insecta as they become available and was designed with this goal in mind, although to date no non-*Drosophila* data have been incorporated. Galaxy [160–162] provides a user-friendly platform for conducting many types of genomic analysis and has potential as a unifying tool for bringing together different data sources [163]. Although having a single consolidated resource for insect genomics and proteomics would greatly facilitate research and reduce the need to navigate multiple different database implementations, developing tools and methods to better take advantage of existing resources for data integration and analysis may prove the most feasible and cost-effective strategy.

6.3 Insect Transgenesis: Historical Perspective and Current State

6.3.1 Application to Understanding Gene Regulation

The ability to transform foreign DNA into a host genome has proven to be a powerful tool for genetic analysis and manipulation and is instrumental for studies of gene regulation. Transgenesis allows for in vivo reporter gene analysis, essential for characterizing regulatory sequences, as well as for generating cell- and tissue-specific markers, determining cell lineages, ablating specific cells, and marking chromosomes for genetic studies.

Genetic transformation was first applied to insects almost half a century ago in the flour moth *Ephestia kuehniella* [164]. In this experiment, larvae with mutant wing scales were injected with wild-type DNA, with some developing into adults with rescue of the phenotype from integrated DNA. Microinjection of DNA into embryos began in the late 1970s with mutant rescue experiments in *D. melanogaster* [165], but *Drosophila* transgenesis did not really take off until the seminal development

by Rubin and Spradling of stable germline transformation through the use of the *P* element transposon [166, 167]. This marked the first instance of mutant rescue in an animal model by heritable gene transfer and laid the foundation for germline transformation in many other model organisms. Although *P*-based transformation proved ineffective for other insect species, a number of other transposable elements with broad efficacy in insects have since been identified. *Minos*, originally discovered in *D. hydei*, was the first transposon vector that was successful in transformation of a non-drosophilid, the medfly *Ceratitis capitata* [168]. The *Minos* transposon is especially useful because of its low insertional bias and high-frequency transformation rates and has thus seen wide use in vertebrate and invertebrate model organisms alike [169]. A second transposon, *piggyBac*, discovered in the cabbage looper moth *Trichoplusia ni* [170], is perhaps the most widely used transposon vector to date and has seen use for transgenesis in many eukaryotic systems [171], including insects [172] and even human cells [173]. *piggyBac* has been used to extend enhancer trapping strategies for identification and functional analysis of genes in both *B. mori* and *T. castaneum* [9, 174–179] and has also been used to transform the genomes of the butterfly *Bicyclus anynana* and the honey bee *A. mellifera* [180, 181]. *Bicyclus* has also been transformed with the transposon *Hermes* [180].

More recently, the use of site-specific recombinases has allowed for reproducible insertion into specific loci. One of the most highly used systems is the φC31 integrase [182–187]. A high and stable integration frequency coupled with its ability to accept integration of large inserts (over 100 kb) has made this a method of choice for many *Drosophila* applications. Subsequent reports have demonstrated the utility of this integrase system in *Ae. aegypti*, *Ae. albopictus*, *An. gambiae*, *C. capitata*, and *B. mori* [188–194]. Although φC31 integration shows great promise for facilitating efficient transformation in diverse insect species, an important caveat is that its use requires prior engineering of the host genome to insert an *attP* landing site for the integration event; multiple landing site choices are desirable as not all sites may prove effective for all applications. Therefore, φC31-mediated transgenesis has been restricted to species for which at least one other method for germline transformation is already available, so that landing site strains can be constructed. However, the relative ease of CRISPR-/Cas9-based genome engineering may soon make it possible to readily add integration landing sites or to simply insert transgenes at a desired location, in a species of choice. Indeed, while the genomes of most insects historically have been refractory to manipulation, the i5K project has provided an impetus to develop and apply efficient transgenic technology to better take advantage of the wealth of accumulating sequence data, and it is likely that we will soon see rapid improvements in strategies for insect transgenesis.

6.3.2 Biotechnological Applications

Effective insect transgenesis will be instrumental to further studies of insect biology and to the understanding of insect gene regulation, and the ability to combine transgenic technologies with a firm understanding of regulatory genomics carries

exciting potential for developing improved methods for insect management and control. For example, the ability to drive gene expression in the adult female salivary glands, midgut, and fat body of *Anopheline* mosquitoes (tissues that play critical roles during infection by and transmission of malaria-causing *Plasmodium* parasites) should greatly facilitate studies of mosquito/parasite interactions and may eventually lead to improved strategies for malaria mitigation [195]. A CRM of the *nanos* gene has been used to drive gene expression in female germ cells of *Ae. aegypti*, a key innovation in mosquito transgenic technology with major implications for future genetic engineering and improved population control of this important disease vector [196]. Similarly, the recent development of female-flightless transgenic control strategies for *Ae. aegypti* uses a muscle-specific CRM to ablate flight muscles in adult females, leading to flightless and therefore effectively sterile mosquitoes [194, 197]. In a materials science rather than a disease vector control setting, application of transgenic technology has been used in the silkworm *B. mori* to produce a variety of biomaterials including expression of the spider silk protein *MaSp1* driven by the *B. mori Ser1* promoter, resulting in silkworm-produced silk with the same unparalleled tensile and structural properties as spider dragline silk [198, 199] (see volume 2, Chap. 9, in this series). Elucidation of additional species-specific and tissue-specific regulatory elements, coupled with improved ability to construct transgenic insects, promises many more advances along these lines in the years to come.

6.4 Prospects for Studying Evolution

Changes in CRMs alter the structure and function of gene regulatory networks, making CRM evolution a major driving force of the morphological diversity seen in metazoan body plans [200–203]. New regulatory functions may be acquired not just by changes in existing CRMs [204–208] but also by the gain of entirely new enhancers, which can arise de novo from nucleotide substitution, deletion, insertion, transposition, or duplication. The details of these processes, as well as the relative frequency of CRM repurposing versus CRM creation, are not yet well understood. Insects are an ideal class of animals in which to study regulatory evolution due to their tremendous diversity and the growing number of species becoming amenable to experimental manipulation. The incredible morphological specialization found within insects even at the family and subfamily level provides us with the opportunity to build a comprehensive comparative developmental framework and to elucidate the genetic and molecular mechanisms behind the vast insect radiation.

6.5 Concluding Remarks

The past decade has witnessed dramatic progress in the area of regulatory genomics, driven by developments in genome sequencing and analysis. Insects, spearheaded by the model research animal *D. melanogaster*, have played a major role in these

advances. The next several years will see completion of the full genome sequencing of a large number of insects across a wide evolutionary spectrum. A great challenge will thus be annotating the regulatory genomes of these diverse sequenced species. Fortunately, the outlook is bright for non-model insects. Decreasing costs for genomic assays and the ability to apply them to increasingly small numbers of—or even single—cells [59, 69, 70, 209] raise the hope that direct empirical studies will become feasible for many different species. Similarly, methods such as RNAi and CRISPR-/Cas9-based genome engineering open up traditionally nongenetic systems to experimental analysis. Computational methods, which have matured greatly over the last dozen years, can predict with growing accuracy CRMs in model and non-model organisms alike. It is thus with great anticipation that we look forward to seeing the power of the computational and empirical methods developed for studying regulatory genomics applied broadly to the insects, with their enormous diversity and tremendous impact on human health and agriculture.

6.6 Further Reading

For a general treatment of transcriptional regulation in eukaryotes, see the detailed review by Maston et al. [33].

For a current perspective on enhancer biology and the implications of the myriad studies on the role of enhancers in development, disease, and evolution, see the recent set of commentaries by several prominent researchers in *Nature Genetics* [210].

For more on how TFBSs can be represented and the basis for such representations, see [104]. The review by Stormo [104] remains one of the best gentle introductions to PWM-based TFBS representation. For a brief and accessible yet thorough treatment, see [211].

For more on computational tools available for motif discovery, readers are referred to the excellent reviews by Zambelli et al. [212], a commentary on the different methods for TFBS discovery before and after the advent of next-generation sequencing, and MacIsaac et al. [213], which describes strategies for using motif-based methods and tools.

Reviews on methods for CRM discovery (both empirical and computational) include Haeussler and Joly's review on strategies and methods to choose when embarking on a CRM discovery project [214] and overviews of the many computational methods for CRM discovery by Van Loo and Marynen [215] and Aerts [216].

For a review on the numerous next-generation technologies currently available to aid functional genomics studies, readers are referred to the excellent commentary by Wold and Myers [217]. Zentner et al. [218] discuss using chromatin features to identify enhancers, and Shyueva et al. [219] provide a recent review on current technologies available for large-scale annotation of regulatory elements.

Resources for *Drosophila*-specific genomics are comprehensively reviewed by Mohr et al. [220].

For reviews on the role on enhancers and CRMs in evolution, see [73, 200, 203, 221]. For thorough coverage of this subject, two major treatments are the books by Eric Davidson [103] and Sean Carroll [222].

Acknowledgments The authors are grateful to John Nyquist at the University at Buffalo for help with the illustrations contained in this chapter. The authors are supported by USDA grant 2011-04656.

References

1. Banerji J, Rusconi S, Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27(2 Pt 1):299–308
2. i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104(5):595–600
3. Liu J, Li C, Yu Z et al (2012) Efficient and specific modifications of the *Drosophila* genome by means of an easy TALEN strategy. J Genet Genomics 39(5):209–215
4. Gilbert LA, Larson MH, Morsut L et al (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154(2):442–451
5. Elick TA, Bauser CA, Fraser MJ (1996) Excision of the piggyBac transposable element *in vitro* is a precise event that is enhanced by the expression of its encoded transposase. Genetica 98(1):33–41
6. Sarkar A, Coates CJ, Whyard S et al (1997) The Hermes element from *Musca domestica* can transpose in four families of cyclorrhaphan flies. Genetica 99(1):15–29
7. Berghammer AJ, Klingler M, Wimmer EA (1999) A universal marker for transgenic insects. Nature 402(6760):370–371
8. Peloquin JJ, Thibault ST, Staten R et al (2000) Germ-line transformation of pink bollworm (Lepidoptera: Gelechiidae) mediated by the piggyBac transposable element. Insect Mol Biol 9(3):323–333
9. Tamura T, Thibert C, Royer C et al (2000) Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. Nat Biotechnol 18(1):81–84
10. Pavlopoulos A, Berghammer AJ, Averof M et al (2004) Efficient transformation of the beetle *Tribolium castaneum* using the Minos transposable element: quantitative and qualitative analysis of genomic integration events. Genetics 167(2):737–746
11. Nakamura T, Yoshizaki M, Ogawa S et al (2010) Imaging of transgenic cricket embryos reveals cell movements consistent with a syncytial patterning mechanism. Curr Biol 20(18):1641–1647
12. Warren IA, Fowler K, Smith H (2010) Germline transformation of the stalk-eyed fly, *Teleopsis dalmanni*. BMC Mol Biol 11:86
13. Kennerdell JR, Carthew RW (2000) Heritable gene silencing in *Drosophila* using double-stranded RNA. Nat Biotechnol 18(8):896–898
14. Brown S, Holtzman S, Kaufman T et al (1999) Characterization of the *Tribolium Deformed* ortholog and its ability to directly regulate *Deformed* target genes in the rescue of a *Drosophila Deformed* null mutant. Dev Genes Evol 209(7):389–398
15. Terenius O, Papanicolaou A, Garbutt JS et al (2011) RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. J Insect Physiol 57(2):231–245
16. Blandin S, Moita LF, Kocher T et al (2002) Reverse genetics in the mosquito *Anopheles gambiae*: targeted disruption of the *Defensin* gene. EMBO Rep 3(9):852–856
17. Osta MA, Christophides GK, Kafatos FC (2004) Effects of mosquito genes on Plasmodium development. Science 303(5666):2030–2032

18. Bucher G, Scholten J, Klingler M (2002) Parental RNAi in *Tribolium* (Coleoptera). Curr Biol 12(3):R85–R86
19. Lynch JA, Desplan C (2006) A method for parental RNA interference in the wasp *Nasonia vitripennis*. Nat Protoc 1(1):486–494
20. Lynch JA, Roth S (2011) The evolution of dorsal-ventral patterning mechanisms in insects. Genes Dev 25(2):107–118
21. Hughes CL, Kaufman TC (2000) RNAi analysis of Deformed, proboscipedia and Sex combs reduced in the milkweed bug *Oncopeltus fasciatus:* novel roles for Hox genes in the hemipteran head. Development 127(17):3683–3694
22. Belles X (2010) Beyond *Drosophila*: RNAi in vivo and functional genomics in insects. Annu Rev Entomol 55:111–128
23. Bibikova M, Golic M, Golic KG et al (2002) Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. Genetics 161(3):1169–1175
24. Bassett AR, Tibbit C, Ponting CP et al (2013) Highly efficient targeted mutagenesis of Drosophila with the CRISPR/Cas9 system. Cell Rep 4(1):220–228
25. Gratz SJ, Cummings AM, Nguyen JN et al (2013) Genome engineering of *Drosophila* with the CRISPR RNA-guided *Cas9* nuclease. Genetics 194(4):1029–1035
26. Richter H, Randau L, Plagens A (2013) Exploiting CRISPR/Cas: interference mechanisms and applications. Int J Mol Sci 14(7):14518–14531
27. Bassett AR, Liu JL (2014) CRISPR/Cas9 and genome editing in *Drosophila*. J Genet Genomics 41(1):7–19
28. Yu Z, Ren M, Wang Z et al (2013) Highly efficient genome modifications mediated by CRISPR/Cas9 in *Drosophila*. Genetics 195(1):289–291
29. Ma S, Chang J, Wang X et al (2014) CRISPR/Cas9 mediated multiplex genome editing and heritable mutagenesis of BmKu70 in *Bombyx mori*. Sci Rep 4:4489
30. Dostie J, Richmond TA, Arnaout RA et al (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16(10):1299–1309
31. Dekker J, Rippe K, Dekker M et al (2002) Capturing chromosome conformation. Science 295(5558):1306–1311
32. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13(4):233–245
33. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29–59
34. Kutach AK, Kadonaga JT (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. Mol Cell Biol 20(13):4754–4764
35. Zhu Q, Halfon MS (2009) Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in *Drosophila melanogaster*. BMC Genomics 10:9
36. Shiraki T, Kondo S, Katayama S et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100(26):15776–15781
37. Ni T, Corcoran DL, Rach EA et al (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7(7):521–527
38. Batut P, Gingeras TR (2013) RAMPAGE: promoter activity profiling by paired-end sequencing of 5′-complete cDNAs. Curr Protoc Mol Biol 104:Unit 25B 11
39. Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4(8):613–614
40. Collas P, Dahl JA (2008) Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Front Biosci 13:929–943
41. Fullwood MJ, Ruan Y (2009) ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 107(1):30–39
42. Pillai S, Chellappan SP (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. Methods Mol Biol 523:341–366

43. Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. Science 327(5963):335–338
44. Hoskins RA, Landolin JM, Brown JB et al (2011) Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21(2):182–192
45. Kvon EZ, Kazmar T, Stampfel G et al (2014) Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 512(7512):91–95
46. Jory A, Estella C, Giorgianni MW et al (2012) A survey of 6,300 genomic fragments for *cis-regulatory* activity in the imaginal discs of *Drosophila melanogaster*. Cell Rep 2(4):1014–1024
47. Jolma A, Kivioja T, Toivonen J et al (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res 20(6):861–873
48. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc 4(3):393–411
49. Meng X, Brodsky MH, Wolfe SA (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol 23(8):988–994
50. Rhee HS, Pugh BF (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr Protoc Mol Biol 100:21.24:21.24.1–21.24.14
51. Hesselberth JR, Chen X, Zhang Z et al (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. Nat Methods 6(4):283–289
52. Cao Y, Yao Z, Sarkar D et al (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell 18(4):662–674
53. Fisher WW, Li JJ, Hammonds AS et al (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. Proc Natl Acad Sci U S A 109(52):21330–21335
54. Li XY, MacArthur S, Bourgon R et al (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. PLoS Biol 6(2):e27
55. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25–36
56. Boyle AP, Davis S, Shulha HP et al (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132(2):311–322
57. Giresi PG, Kim J, McDaniell RM et al (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17(6):877–885
58. Giresi PG, Lieb JD (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods 48(3):233–239
59. Buenrostro JD, Giresi PG, Zaba LC et al (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10(12):1213–1218
60. Whyte WA, Orlando DA, Hnisz D et al (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153(2):307–319
61. Visel A, Blow MJ, Li Z, Zhang T et al (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457(7231):854–858
62. Heintzman ND, Stuart RK, Hon G et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39(3):311–318
63. Heintzman ND, Hon GC, Hawkins RD et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459(7243):108–112
64. Arnold CD, Gerlach D, Stelzer C et al (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339(6123):1074–1077
65. Gisselbrecht SS, Barrera LA, Porsch M et al (2013) Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. Nat Methods 10(8):774–780

66. Murtha M, Tokcaer-Keskin Z, Tang Z et al (2014) FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat Methods 11(5):559–565
67. Dickel DE, Zhu Y, Nord AS et al (2014) Function-based identification of mammalian enhancers using site-specific integration. Nat Methods 11(5):566–571
68. Thomas S, Li XY, Sabo PJ et al (2011) Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. Genome Biol 12(5):R43
69. Adli M, Bernstein BE (2011) Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. Nat Protoc 6(10):1656–1668
70. Nagano T, Lubling Y, Stevens TJ et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502(7469):59–64
71. Goh Y, Fullwood MJ, Poh HM et al (2012) Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. J Vis Exp 62:e3770
72. He B, Chen C, Teng L et al (2014) Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A 111(21):E2191–E2199
73. Wittkopp PJ (2006) Evolution of cis-regulatory sequence and function in Diptera. Heredity 97(3):139–147
74. Li L, Zhu Q, He X, Sinha S, Halfon MS (2007) Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol 8(6):R101
75. Su J, Teichmann SA, Down TA (2010) Assessing computational methods of cis-regulatory module prediction. PLoS Comput Biol 6(12):e1001020
76. Swanson CI, Schwimmer DB, Barolo S (2011) Rapid evolutionary rewiring of a structurally constrained eye enhancer. Curr Biol 21(14):1186–1196
77. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EE (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148(3):473–486
78. Berman BP, Pfeiffer BD, Laverty TR et al (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol 5(9):R61
79. Kim J, Sinha S (2010) Towards realistic benchmarks for multiple alignments of non-coding sequences. BMC Bioinf 11:54
80. Kheradpour P, Stark A, Roy S et al (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17(12):1919–1931
81. Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450(7167):203–218
82. Sieglaff DH, Dunn WA, Xie XS et al (2009) Comparative genomics allows the discovery of *cis*-regulatory elements in mosquitoes. Proc Natl Acad Sci U S A 106(9):3053–3058
83. Kim J, Cunningham R, James B et al (2010) Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances. PLoS Comput Biol 6(1):e1000652
84. Stark A, Lin MF, Kheradpour P et al (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450(7167):219–232
85. Brody T, Rasband W, Baler K et al (2007) cis-Decoder discovers constellations of conserved DNA sequences shared among tissue-specific enhancers. Genome Biol 8(5):R75
86. Papatsenko D, Levine M (2005) Computational identification of regulatory DNAs underlying animal development. Nat Methods 2(7):529–534
87. Papaceit M, Orengo D, Juan E (2004) Sequences upstream of the homologous cis-elements of the Adh adult enhancer of *Drosophila* are required for maximal levels of *Adh* gene transcription in adults of *Scaptodrosophila lebanonensis*. Genetics 167(1):289–299
88. Gibert JM, Simpson P (2003) Evolution of cis-regulation of the proneural genes. Int J Dev Biol 47(7–8):643–651

89. Mitsialis SA, Kafatos FC (1985) Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature 317(6036):453–456
90. Langeland JA, Carroll SB (1993) Conservation of regulatory elements controlling hairy pair-rule stripe formation. Development 117(2):585–596
91. Lukowitz W, Schroder C, Glaser G et al (1994) Regulatory and coding regions of the segmentation gene hunchback are functionally conserved between *Drosophila virilis* and *Drosophila melanogaster*. Mech Dev 45(2):105–115
92. Ludwig MZ, Patel NH, Kreitman M (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. Development 125(5):949–958
93. Wittkopp PJ, Vaccaro K, Carroll SB (2002) Evolution of yellow gene regulation and pigmentation in Drosophila. Curr Biol 12(18):1547–1556
94. Paris M, Kaplan T, Li XY et al (2013) Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. PLoS Genet 9(9):e1003748
95. Moses AM, Pollard DA, Nix DA et al (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. PLoS Comput Biol 2(10):e130
96. Ludwig MZ, Bergman C, Patel NH et al (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403(6769):564–567
97. Swanson CI, Evans NC, Barolo S (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev Cell 18(3):359–370
98. Erceg J, Saunders TE, Girardot C et al (2014) Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. PLoS Genet 10(1):e1004060
99. Richards S, Liu Y, Bettencourt BR et al (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res 15(1):1–18
100. Hare EE, Peterson BK, Iyer VN et al (2008) Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet 4(6):e1000106
101. Cande J, Goltsev Y, Levine MS (2009) Conservation of enhancer location in divergent insects. Proc Natl Acad Sci U S A 106(34):14414–14419
102. Ludwig MZ, Palsson A, Alekseeva E et al (2005) Functional evolution of a *cis*-regulatory module. PLoS Biol 3(4):e93
103. Davidson EH (2006) The regulatory genome: gene regulatory networks in development and evolution. Academic, Burlington/San Diego
104. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16(1):16–23
105. Marinescu VD, Kohane IS, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinf 6:79
106. Cave JW, Loh F, Surpris JW et al (2005) A DNA transcription code for cell-specific gene activation by notch signaling. Curr Biol 15(2):94–104
107. Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278(1):167–181
108. Rebeiz M, Reeves NL, Posakony JW (2002) SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. Proc Natl Acad Sci U S A 99(15):9888–9893
109. Berman BP, Nibu Y, Pfeiffer BD et al (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci U S A 99(2):757–762
110. Rajewsky N, Vergassola M, Gaul U et al (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. BMC Bioinf 3:30
111. Markstein M, Markstein P, Markstein V et al (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc Natl Acad Sci U S A 99(2):763–768

112. Halfon MS, Grad Y, Church GM et al (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Genome Res 12(7):1019–1028
113. Schroeder MD, Pearce M, Fak J et al (2004) Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biol 2(9):E271
114. Grad YH, Roth FP, Halfon MS et al (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*. Bioinformatics 20(16):2738–2750
115. Sinha S, Schroeder MD, Unnerstall U et al (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in *Drosophila*. BMC Bioinf 5:129
116. Ivan A, Halfon MS, Sinha S (2008) Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs. Genome Biol 9(1):R22
117. Kantorovitz MR, Kazemian M, Kinston S et al (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in *Drosophila* and mouse. Dev Cell 17(4):568–579
118. Tompa M, Li N, Bailey TL et al (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23(1):137–144
119. Halfon MS, Zhu Q, Brennan ER et al (2011) Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. BMC Genomics 12:578
120. Kahana S, Pnueli L, Kainth P et al (2010) Functional dissection of IME1 transcription using quantitative promoter-reporter screening. Genetics 186(3):829–841
121. Sinha S, He X (2007) MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput Biol 3(11):e216
122. He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. PLoS Comput Biol 5(3):e1000299
123. Majoros WH, Ohler U (2010) Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. PLoS Comput Biol 6(12):e1001037
124. Zhou Q, Wong WH (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A 101(33):12114–12119
125. Arunachalam M, Jayasurya K, Tomancak P et al (2010) An alignment-free method to identify candidate orthologous enhancers in multiple *Drosophila* genomes. Bioinformatics 26(17):2109–2115
126. Kazemian M, Zhu Q, Halfon MS et al (2011) Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. Nucleic Acids Res 39(22):9463–9472
127. Wolff C, Schroder R, Schulz C et al (1998) Regulation of the *Tribolium* homologues of caudal and hunchback in *Drosophila*: evidence for maternal gradient systems in a short germ embryo. Development 125(18):3645–3654
128. Erives A, Levine M (2004) Coordinate enhancers share common organizational features in the Drosophila genome. Proc Natl Acad Sci U S A 101(11):3851–3856
129. Zinzen RP, Cande J, Ronshaugen M et al (2006) Evolution of the ventral midline in insect embryos. Dev Cell 11(6):895–902
130. Goltsev Y, Fuse N, Frasch M et al (2007) Evolution of the dorsal-ventral patterning network in the mosquito, *Anopheles gambiae*. Development 134(13):2415–2424
131. Kazemian M, Suryamohan K, Chen JY et al (2014) Evidence for deep regulatory similarities in early developmental programs across highly diverged insects. Genome Biol Evol 6(9):2301–2320
132. Zdobnov EM, Bork P (2007) Quantification of insect genome divergence. Trends Genet 23(1):16–20
133. FlyBase Consortium (2002) The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 30(1):106–108

134. Munoz-Torres MC, Reese JT, Childers CP et al (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Res 39(Database issue):D658–D662
135. Legeai F, Shigenobu S, Gauthier JP et al (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol 19(Suppl 2):5–12
136. Wang L, Wang S, Li Y et al (2007) BeetleBase: the model organism database for *Tribolium castaneum*. Nucleic Acids Res 35(Database issue):D476–D479
137. Papanicolaou A, Gebauer-Jung S, Blaxter ML et al (2008) ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Res 36(Database issue):D582–D587
138. Wang J, Xia Q, He X et al (2005) SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res 33(Database issue):D399–D402
139. Negre V, Hotelier T, Volkoff AN et al (2006) SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. BMC Bioinf 7:322
140. Megy K, Emrich SJ, Lawson D et al (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. Nucleic Acids Res 40(Database issue): D729–D734
141. Ma Z, Yu J, Kang L (2006) LocustDB: a relational database for the transcriptome and biology of the migratory locust (*Locusta migratoria*). BMC Genomics 7:11
142. Papanicolaou A, Heckel DG (2010) The GMOD Drupal bioinformatic server framework. Bioinformatics 26(24):3119–3124
143. Lawson D, Arensburger P, Atkinson P et al (2009) VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Res 37(Database issue):D583–D587
144. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database 2011:bar049
145. Karolchik D, Hinrichs AS, Kent WJ (2009) The UCSC Genome Browser. Curr Protoc Bioinformatics 40:1.4:1.r.1-4.33
146. Lyne R, Smith R, Rutherford K, Wakeling M et al (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. Genome Biol 8(7):R129
147. Hammonds AS, Bristow CA, Fisher WW et al (2013) Spatial expression of transcription factors in *Drosophila* embryonic organ development. Genome Biol 14(12):R140
148. Tomancak P, Beaton A, Weiszmann R et al (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 3(12):RESEARCH0088
149. Tomancak P, Berman BP, Beaton A et al (2007) Global analysis of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 8(7):R145
150. Kumar S, Konikoff C, Van Emden B et al (2011) FlyExpress: visual mining of spatiotemporal patterns for genes and publications in *Drosophila* embryogenesis. Bioinformatics 27(23):3319–3320
151. Robinson SW, Herzyk P, Dow JA et al (2013) FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*. Nucleic Acids Res 41(Database issue):D744–D750
152. modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330(6012):1787–1797
153. Gallo SM, Gerrard DT, Miner D et al (2011) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. Nucleic Acids Res 39(Database issue):D118–D123
154. Sandelin A, Alkema W, Engstrom P et al (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32(Database issue):D91–D94
155. Wingender E, Chen X, Hehl R et al (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28(1):316–319
156. Zhu LJ, Christensen RG, Kazemian M et al (2011) FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res 39(Database issue):D111–D117
157. Kazemian M, Brodsky MH, Sinha S (2011) Genome Surveyor 2.0: cis-regulatory analysis in Drosophila. Nucleic Acids Res 39(Web Server issue):W79–W85

158. Adryan B, Teichmann SA (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. Bioinformatics 22(12):1532–1533
159. Shazman S, Lee H, Socol Y et al (2014) OnTheFly: a database of *Drosophila melanogaster* transcription factors and their binding sites. Nucleic Acids Res 42(Database issue): D167–D171
160. Giardine B, Riemer C, Hardison RC et al (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455
161. Blankenberg D, Von Kuster G, Coraor N et al (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 89:19.10:19.10.1–19.10.21
162. Goecks J, Nekrutenko A, Taylor J et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86
163. Blankenberg D, Coraor N, Von Kuster G et al (2011) Integrating diverse databases into an unified analysis framework: a Galaxy approach. Database 2011:bar011
164. Caspari EW, Nawa S (1965) A method to demonstrate transformation in *Ephestia*. Z Naturforsch 20b:281–284
165. Germeraad S (1976) Genetic transformation in *Drosophila* by microinjection of DNA. Nature 262(5565):229–231
166. Rubin GM, Spradling AC (1982) Genetic transformation of *Drosophila* with transposable element vectors. Science 218(4570):348–353
167. Spradling AC, Rubin GM (1982) Transposition of cloned P elements into *Drosophila* germ line chromosomes. Science 218(4570):341–347
168. Loukeris TG, Livadaras I, Arca B et al (1995) Gene transfer into the medfly, *Ceratitis capitata*, with a *Drosophila hydei* transposable element. Science 270(5244):2002–2005
169. Pavlopoulos A, Oehler S, Kapetanaki MG et al (2007) The DNA transposon Minos as a tool for transgenesis and functional genomic analysis in vertebrates and invertebrates. Genome Biol 8(Suppl 1):S2
170. Handler AM, McCombs SD, Fraser MJ et al (1998) The lepidopteran transposon vector, piggyBac, mediates germ-line transformation in the Mediterranean fruit fly. Proc Natl Acad Sci U S A 95(13):7520–7525
171. Balu B, Shoue DA, Fraser MJ Jr et al (2005) High-efficiency transformation of *Plasmodium falciparum* by the lepidopteran transposable element piggyBac. Proc Natl Acad Sci U S A 102(45):16391–16396
172. Handler AM (2002) Use of the piggyBac transposon for germ-line transformation of insects. Insect Biochem Mol Biol 32(10):1211–1220
173. Wilson MH, Coates CJ, George AL Jr (2007) PiggyBac transposon-mediated gene transfer in human cells. Mol Ther 15(1):139–145
174. Berghammer A, Bucher G, Maderspacher F et al (1999) A system to efficiently maintain embryonic lethal mutations in the flour beetle *Tribolium castaneum*. Dev Genes Evol 209(6):382–389
175. Schinko JB, Weber M, Viktorinova I et al (2010) Functionality of the GAL4/UAS system in *Tribolium* requires the use of endogenous core promoters. BMC Dev Biol 10:53
176. Berghammer AJ, Weber M, Trauner J et al (2009) Red flour beetle (*Tribolium*) germline transformation and insertional mutagenesis. Cold Spring Harb Protoc 2009(8):pdb prot5259
177. Trauner J, Schinko J, Lorenzen MD et al (2009) Large-scale insertional mutagenesis of a coleopteran stored grain pest, the red flour beetle *Tribolium castaneum*, identifies embryonic lethal mutations and enhancer traps. BMC Biol 7:73
178. Eckert C, Aranda M, Wolff C et al (2004) Separable stripe enhancer elements for the pair-rule gene hairy in the beetle *Tribolium*. EMBO Rep 5(6):638–642
179. Uchino K, Sezutsu H, Imamura M et al (2008) Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori*. Insect Biochem Mol Biol 38(12):1165–1173
180. Marcus JM, Ramos DM, Monteiro A (2004) Germline transformation of the butterfly *Bicyclus anynana*. Proc Roy Soc 271(Suppl 5):S263–S265

181. Schulte C, Theilenberg E, Muller-Borg M et al (2014) Highly efficient integration and expression of piggyBac-derived cassettes in the honeybee (*Apis mellifera*). Proc Natl Acad Sci U S A 111(24):9003–9008
182. Groth AC, Fish M, Nusse R et al (2004) Construction of transgenic *Drosophila* by using the site-specific integrase from phage phiC31. Genetics 166(4):1775–1782
183. Oberstein A, Pare A, Kaplan L et al (2005) Site-specific transgenesis by Cre-mediated recombination in *Drosophila*. Nat Methods 2(8):583–585
184. Horn C, Handler AM (2005) Site-specific genomic targeting in *Drosophila*. Proc Natl Acad Sci U S A 102(35):12483–12488
185. Bateman JR, Lee AM, Wu CT (2006) Site-specific transformation of *Drosophila* via phiC31 integrase-mediated cassette exchange. Genetics 173(2):769–777
186. Venken KJ, He Y, Hoskins RA et al (2006) P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. Science 314(5806):1747–1751
187. Bischof J, Maeda RK, Hediger M et al (2007) An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A 104(9):3312–3317
188. Amenya DA, Bonizzoni M, Isaacs AT et al (2010) Comparative fitness assessment of *Anopheles stephensi* transgenic lines receptive to site-specific integration. Insect Mol Biol 19(2):263–269
189. Meredith JM, Basu S, Nimmo DD et al (2011) Site-specific integration and expression of an anti-malarial gene in transgenic *Anopheles gambiae* significantly reduces *Plasmodium* infections. PLoS One 6(1):e14587
190. Nakayama G, Kawaguchi Y, Koga K et al (2006) Site-specific gene integration in cultured silkworm cells mediated by phiC31 integrase. Mol Genet Genomics 275(1):1–8
191. Nimmo DD, Alphey L, Meredith JM et al (2006) High efficiency site-specific genetic engineering of the mosquito genome. Insect Mol Biol 15(2):129–136
192. Schetelig MF, Scolari F, Handler AM et al (2009) Site-specific recombination for the modification of transgenic strains of the Mediterranean fruit fly *Ceratitis capitata*. Proc Natl Acad Sci U S A 106(43):18171–18176
193. Labbe GM, Nimmo DD, Alphey L (2010) piggybac- and PhiC31-mediated genetic transformation of the Asian tiger mosquito, *Aedes albopictus* (Skuse). PLoS Negl Trop Dis 4(8):e788
194. Fu G, Lees RS, Nimmo D et al (2010) Female-specific flightless phenotype for mosquito control. Proc Natl Acad Sci U S A 107(10):4550–4554
195. O'Brochta DA, Pilitt KL, Harrell RA 2nd et al (2012) Gal4-based enhancer-trapping in the malaria mosquito *Anopheles stephensi*. G3 2(11):1305–1315
196. Adelman ZN, Jasinskiene N, Onal S et al (2007) nanos gene control DNA mediates developmentally regulated transposition in the yellow fever mosquito *Aedes aegypti*. Proc Natl Acad Sci U S A 104(24):9970–9975
197. Wise de Valdez MR, Nimmo D, Betz J et al (2011) Genetic elimination of dengue vector mosquitoes. Proc Natl Acad Sci U S A 108(12):4772–4775
198. Griffiths JR, Salanitri VR (1980) The strength of spider silk. J Mater Sci 15(2):491–496
199. Wen H, Lan X, Zhang Y et al (2010) Transgenic silkworms (*Bombyx mori)* produce recombinant spider dragline silk in cocoons. Mol Biol Rep 37(4):1815–1821
200. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206–216
201. Carroll SB (2005) Evolution at two levels: on genes and form. PLoS Biol 3(7):e245
202. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. Mol Ecol 15(5):1197–1211
203. Rubinstein M, de Souza FS (2013) Evolution of transcriptional enhancers and animal diversity. Philos Trans R Soc Lond B Biol Sci 368(1632):20130017
204. Sucena E, Stern DL (2000) Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby*. Proc Natl Acad Sci U S A 97(9):4530–4534

205. Carbone MA, Llopart A, deAngelis M et al (2005) Quantitative trait loci affecting the difference in pigmentation between *Drosophila yakuba* and *D. santomea*. Genetics 171(1):211–225
206. Pool JE, Aquadro CF (2007) The genetic basis of adaptive pigmentation variation in *Drosophila melanogaster*. Mol Ecol 16(14):2844–2851
207. Frankel N, Erezyilmaz DF, McGregor AP et al (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474(7353):598–603
208. Stone JR, Wray GA (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18(9):1764–1770
209. Deng Q, Ramskold D, Reinius B et al (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343(6167):193–196
210. Pennacchio LA, Bickmore W, Dean A et al (2013) Enhancers: five essential questions. Nat Rev Genet 14(4):288–295
211. D'Haeseleer P (2006) What are DNA sequence motifs? Nat Biotechnol 24(4):423–425
212. Zambelli F, Pesole G, Pavesi G (2013) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief Bioinform 14(2):225–237
213. MacIsaac KD, Fraenkel E (2006) Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol 2(4):e36
214. Haeussler M, Joly JS (2011) When needles look like hay: how to find tissue-specific enhancers in model organism genomes. Dev Biol 350(2):239–254
215. Van Loo P, Marynen P (2009) Computational methods for the detection of cis-regulatory modules. Brief Bioinform 10(5):509–524
216. Aerts S (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr Top Dev Biol 98:121–145
217. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5(1):19–21
218. Zentner GE, Scacheri PC (2012) The chromatin fingerprint of gene enhancer elements. J Biol Chem 287(37):30888–30896
219. Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15(4):272–286
220. Mohr SE, Hu Y, Kim K et al (2014) Resources for functional genomics studies in *Drosophila melanogaster*. Genetics 197(1):1–18
221. Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A 104(Suppl 1):8605–8612
222. Carroll SB, Grenier JK, Weatherbee SD (2005) From DNA to diversity: molecular genetics and the evolution of animal design, 2nd edn. Blackwell, Malden