Published in final edited form as:
Nature. 2007 November 8; 450(7167): 219–232.

Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2,*], **Michael F. Lin**[1,2,*], **Pouya Kheradpour**[2,*], **Jakob S. Pedersen**[3,4,*], **Leopold Parts**[5,6], **Joseph W. Carlson**[7], **Madeline A. Crosby**[8], **Matthew D. Rasmussen**[2], **Sushmita Roy**[9], **Ameya N. Deoras**[2], **J. Graham Ruby**[10,11], **Julius Brennecke**[12], **Harvard FlyBase curators**[†], **Berkeley Drosophila Genome Project**[†], **Emily Hodges**[12], **Angie S. Hinrichs**[4], **Anat Caspi**[13], **Benedict Paten**[4,5,14], **Seung-Won Park**[15], **Mira V. Han**[16], **Morgan L. Maeder**[17], **Benjamin J. Polansky**[17], **Bryanne E. Robson**[17], **Stein Aerts**[18,19], **Jacques van Helden**[20], **Bassem Hassan**[18,19], **Donald G. Gilbert**[21], **Deborah A. Eastman**[17], **Michael Rice**[22], **Michael Weir**[23], **Matthew W. Hahn**[16], **Yongkyu Park**[15], **Colin N. Dewey**[24], **Lior Pachter**[25,26], **W. James Kent**[4], **David Haussler**[4], **Eric C. Lai**[27], **David P. Bartel**[10,11], **Gregory J. Hannon**[12], **Thomas C. Kaufman**[21], **Michael B. Eisen**[28,29], **Andrew G. Clark**[30], **Douglas Smith**[31], **Susan E. Celniker**[7], **William M. Gelbart**[8,32], and **Manolis Kellis**[1,2]

1 *The Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02140, USA*

2 *Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, USA*

3 *The Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark*

4 *Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA*

5 *Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*

6 *Institute of Computer Science, University of Tartu, Estonia*

7 *BDGP, LBNL, 1 Cyclotron Road MS 64-0119, Berkeley, California 94720, USA*

8 *FlyBase, The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA*

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu).

*These authors contributed equally to this work.

†Lists of participants and affiliations appear at the end of the paper.

Author Contributions Organizing committee: Manolis Kellis, William Gelbart, Doug Smith, Andrew G. Clark, Michael E. Eisen, Thomas C. Kaufman; protein-coding gene prediction: Michael F. Lin, Ameya N. Deoras, Mira V. Han, Matthew W. Hahn, Donald G. Gilbert, Michael Weir, Michael Rice, Manolis Kellis; manual curation of protein-coding genes: Madeline A. Crosby, Harvard FlyBase curators, William M. Gelbart; validation of protein-coding genes: Joseph W. Carlson, Berkeley Drosophila Genome Project, Susan E. Celniker; non-coding RNA gene prediction: Jakob S. Pedersen, David Haussler, Yongkyu Park, Seung-Won Park, Manolis Kellis; microRNA gene prediction: Alexander Stark, Pouya Kheradpour, Leopold Parts, Manolis Kellis; microRNA cloning and sequencing: Julius Brennecke, Emily Hodges, Gregory J. Hannon; microRNA target prediction: Alexander Stark, J. Graham Ruby, Manolis Kellis, Eric C. Lai, David P. Bartel; motif identification: Alexander Stark, Pouya Kheradpour, Manolis Kellis; motif instance prediction: Alexander Stark, Pouya Kheradpour, Sushmita Roy, Morgan L. Maeder, Benjamin J. Polansky, Bryanne E. Robson, Deborah A. Eastman, Stein Aerts, Bassem Hassan, Jacques van Helden, Manolis Kellis; genome alignments: Angie S. Hinrichs, W. James Kent, Anat Caspi, Lior Pachter, Colin N. Dewey, Benedict Paten; phylogeny and branch length estimation: Matthew D. Rasmussen, Manolis Kellis; final manuscript preparation: Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob Pedersen, Manolis Kellis.

Harvard FlyBase curators Madeline A. Crosby1, Beverley B. Matthews1, Andrew J. Schroeder1, L. Sian Gramates1, Susan E. St Pierre1, Margaret Roark1, Kenneth L. Wiley Jr1, Rob J. Kulathinal1, Peili Zhang1, Kyl V. Myrick1, Jerry V. Antone1 & William M. Gelbart1

Berkeley Drosophila Genome Project Joseph W. Carlson2, Charles Yu2, Soo Park2, Kenneth H. Wan2 & Susan E. Celniker2

9 *Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87131, USA*

10 *Department of Biology, MIT, Cambridge, Massachusetts 02139, USA*

11 *Whitehead Institute, Cambridge, Massachusetts 02142, USA*

12 *Cold Spring Harbor Laboratory, Watson School of Biological Sciences, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA*

13 *University of California, San Francisco/University of California, Berkeley Joint Graduate Group in Bioengineering, Berkeley, California 97210, USA*

14 *EMBL Nucleotide Sequence Submissions, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*

15 *Department of Cell Biology and Molecular Medicine, G-629, MSB, 185 South Orange Avenue, UMDNJ-New Jersey Medical School, Newark, New Jersey 07103, USA*

16 *Department of Biology and School of Informatics, Indiana University, Indiana 47405, USA*

17 *Department of Biology, Connecticut College, New London, Connecticut 06320, USA*

18 *Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, 3000 Leuven, Belgium*

19 *Department of Human Genetics, K. U. Leuven School of Medicine, 3000 Leuven, Belgium*

20 *Department de Biologie Moleculaire, Universite Libre de Bruxelles, 1050 Brussels, Belgium*

21 *Department of Biology, Indiana University, Bloomington, Indiana 47405, USA*

22 *Department of Mathematics and Computer Science, Wesleyan University, Middletown, Connecticut 06459, USA*

23 *Biology Department, Wesleyan University Middletown, Connecticut 06459, USA*

24 *Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA*

25 *Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA*

26 *Department of Computer Science, University of California at Berkeley, Berkeley, California 94720, USA*

27 *Department of Developmental Biology, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA*

28 *Graduate Group in Biophysics, Department of Molecular and Cell Biology, and Center for Integrative Genomics, University of California, Berkeley, California 94720, USA*

29 *Lawrence Berkeley National Laboratory, Life Sciences Division, Berkeley, California 94720, USA*

30 *Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA*

31 *Agencourt Bioscience Corporation, 500 Cummings Center, Suite 2450, Beverly, Massachusetts 01915, USA*

32 *The Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA*

Abstract

Sequencing of multiple related species followed by comparative genomics analysis constitutes a powerful approach for the systematic understanding of any genome. Here, we use the genomes of 12 *Drosophila* species for the *de novo* discovery of functional elements in the fly. Each type of

functional element shows characteristic patterns of change, or ‘evolutionary signatures’, dictated by its precise selective constraints. Such signatures enable recognition of new protein-coding genes and exons, spurious and incorrect gene annotations, and numerous unusual gene structures, including abundant stop-codon readthrough. Similarly, we predict non-protein-coding RNA genes and structures, and new microRNA (miRNA) genes. We provide evidence of miRNA processing and functionality from both hairpin arms and both DNA strands. We identify several classes of pre- and post-transcriptional regulatory motifs, and predict individual motif instances with high confidence. We also study how discovery power scales with the divergence and number of species compared, and we provide general guidelines for comparative studies.

The sequencing of the human genome and the genomes of dozens of other metazoan species has intensified the need for systematic methods to extract biological information directly from DNA sequence. Comparative genomics has emerged as a powerful methodology for this endeavour[1,2]. Comparison of few (two–four) closely related genomes has proven successful for the discovery of protein-coding genes[3–5], RNA genes[6,7], miRNA genes[8–11] and catalogues of regulatory elements[3,4,12–14]. The resolution and discovery power of these studies should increase with the number of genomes[15–20], in principle enabling the systematic discovery of all conserved functional elements.

The fruitfly *Drosophila melanogaster* is an ideal system for developing and evaluating comparative genomics methodologies. Over the past century, *Drosophila* has been a pioneering model in which many of the basic principles governing animal development and population biology were established[21]. In the past decade, the genome sequence of *D. melanogaster* provided one of the first systematic views of a metazoan genome[22], and the ongoing effort by the FlyBase and Berkeley Drosophila Genome Project (BDGP) groups established a systematic high-quality genome annotation[23–25]. Moreover, the fruit-fly benefits from extensive experimental resources[26–28], which enable novel functional elements to be systematically tested and used in the evaluation of genetic screens[29,30].

The fly research community has sequenced, assembled and annotated the genomes of 12 *Drosophila* species[22,31,32] at a range of evolutionary distances from *D. melanogaster* (Fig. 1a, b). The analysis of these genomes was organized around two complementary aims. The first, described in an accompanying paper[32], was to understand the evolution of genes and chromosomes on the *Drosophila* phylogeny, and how it relates to speciation and adaptation. The second goal, described here, was to develop general comparative methodologies to discover and refine functional elements in *D. melanogaster* using the 12 genomes, and to investigate the scaling of discovery power and its implications for studies in vertebrates (Fig. 1c).

Here, we report genome-wide alignments of the 12 species (Supplementary Information 1), and the systematic discovery of euchromatic functional elements in the *D. melanogaster* genome. We predict and refine thousands of protein-coding exons, RNA genes and structures, miRNAs, pre- and post-transcriptional regulatory motifs and regulatory targets. We validate many of these elements using complementary DNA (cDNA) sequencing, human curation, small RNA sequencing, and correlation with experimentally supported transcription factor and miRNA targets. In addition, our analysis leads to several specific biological findings, listed below.

- We predict 123 novel polycistronic transcripts, 149 genes with apparent stop-codon readthrough and several candidate programmed frameshifts, with potential roles in regulation, localization and function of the corresponding protein products.

- We make available the first systematic prediction of general RNA genes and structures (non-coding RNAs (ncRNAs)) in *Drosophila*, including several structures probably involved in translational regulation and adenosine-to-inosine RNA editing (A-to-I editing).
- We present comparative and experimental evidence that some miRNA loci yield multiple functional products, from both hairpin arms or from both DNA strands, thereby increasing the versatility and complexity of miRNA-mediated regulation.
- We provide further comparative evidence for miRNA targeting in protein-coding exons.
- We report an initial network of pre- and post-transcriptional regulatory targets in *Drosophila* on the basis of individual high-confidence motif occurrences.

Comparative genomics and evolutionary signatures

Although multiple closely related genomes provide sufficient neutral divergence for recognition of functional regions in stretches of highly conserved nucleotides[16,17,33], measures of nucleotide conservation alone do not distinguish between different types of functional elements. Moreover, functional elements that tolerate abundant 'silent' mutations, such as protein-coding exons and many regulatory motifs, might not be detected when searching on the basis of strong nucleotide conservation.

Across many genomes spanning larger evolutionary distances, the information in the patterns of sequence change reveals evolutionary signatures (Fig. 2) that can be used for systematic genome annotation. Protein-coding regions show highly constrained codon substitution frequencies[34] and insertions and deletions that are heavily biased to be multiples of three[3] (Fig. 2a). RNA genes and structures tolerate substitutions that preserve base pairing[35,36] (Fig. 2b). MicroRNA hairpins show a characteristic conservation profile with high conservation in the stem and mutations in loop regions[10,11] (Fig. 2c). Finally, regulatory motifs are marked by high levels of genome-wide conservation[3,4,12–14], and post-transcriptional motifs show strand-biased conservation[12] (Fig. 2d, e).

We find that these signatures can be much more precise for genome annotation than the overall level of nucleotide conservation (for example, Fig. 3a).

Revisiting the protein-coding gene catalogue

The annotation of protein-coding genes remains difficult in metazoan genomes owing to short exons and complex gene structures with abundant alternative splicing. Comparative information has improved computational gene predictors[5], but their accuracy still falls far short of well-studied gene catalogues such as the FlyBase annotation, which combines computational gene prediction[37], high-throughput experimental data[38–42] and extensive manual curation[23]. Recognizing this, we set out not only to produce an independent computational annotation of protein-coding genes in the fly genome, but also to assess and refine its already high-quality annotations[43].

Our analyses of *D. melanogaster* coding genes are based on two independent evolutionary signatures unique to protein-coding regions (Fig. 2a): (1) reading frame conservation (RFC)[3], which observes the tendency of nucleotide insertions and deletions to preserve the codon reading frame; and (2) codon substitution frequencies (CSF, see Supplementary Methods 2a), which observes mutational biases towards synonymous codon substitutions and conservative amino acid changes, similar to the non-synonymous/synonymous substitution ratio K_A/K_S[34] and other methods[44–46].

Assessing and refining existing gene annotations

We first assessed the 13,733 euchromatic genes in FlyBase[47] release 4.3. Using the above measures, we defined tests that 'confirmed' genes supported by the evolutionary evidence, 'rejected' genes inconsistent with protein-coding selection, or 'abstained' for genes that were not aligned or with ambiguous comparative evidence (Supplementary Methods 2a). Of the 4,711 genes with descriptive names, we confirmed 97%, rejected 1% and abstained for 2%, whereas the same criteria applied to 15,000 random non-coding regions ≥300 nucleotides rejected 99% of candidates and confirmed virtually none (Table 1). Together, these results illustrate the high sensitivity and specificity of our criteria.

Applying the same criteria to the 9,022 genes lacking a descriptive name (genes designated only by a CG identifier, referred to hereafter as CGid-only genes), our tests accepted 87%, rejected 5% (414 genes) and abstained for 8%. This provides strong evidence that most CGid-only genes encode proteins, but also suggests that they may be less constrained[20,32] and/or may include incorrect annotations. Indeed, on manual review, 222 (54%) of the 414 rejected CGid-only genes were re-categorized as non-protein-coding or deleted (of which 55 were due to genomically primed clones), 73 (18%) were flagged as being of uncertain quality, and the remaining 119 (29%) were kept unchanged (Fig. 3b). Some of these are probably rapidly evolving protein-coding genes, but others may also prove to be non-protein-coding genes or spurious; in fact, none of these had any functional gene ontology (GO) annotation[48].

In addition, we proposed specific corrections and adjustments to hundreds of existing transcript models, including translation start site adjustments (Supplementary Fig. 2b), alternative splice boundaries (Supplementary Fig. 2b), recent nonsense mutations (Supplementary Fig. 2c) and alternative translational reading frames[43].

Identifying new genes and exons

To predict new protein-coding exons, we integrated our metrics into a probabilistic algorithm that determines an optimal segmentation of the genome into protein-coding and non-coding regions (Fig. 3a) on the basis of whole-genome sequence alignments of the 12 fly species (Supplementary Methods 2a). Our genome-wide search predicted 1,193 new protein-coding exons, mostly in euchromatic regions annotated as intergenic (43%), intronic (26%), or 5′/3′ untranslated region (UTR; 23%) in FlyBase annotation release 4.3.

We manually reviewed 928 of these predictions according to FlyBase standards[23] (Supplementary Methods 2a), leading to 142 new gene models (incorporating 192 predictions) and 438 revised gene models (incorporating 562 predictions) (Fig. 3b). In parallel, we tested 184 predictions (126 intergenic, 58 intronic) by directed cDNA sequencing using inverse polymerase chain reaction (inverse PCR) of circularized full-length clones[49–51] (Fig. 3c), which validated 120 targeted predictions (65%) and an additional 42 predictions not directly targeted but contained within the recovered transcripts. Predictions in intergenic regions yielded 88 full-length cDNAs, providing evidence for 50 new genes and modification of 39 gene models. Predictions within introns of existing annotations yielded 32 full-length cDNAs, of which only 18 (56%) represent new splice variants of the surrounding gene, whereas the remaining 14 revealed nested or interleaved gene structures. This provides additional evidence that such complex gene structures are not rare in *Drosophila*[23].

Overall, 83% of the 948 predicted exons that we assessed by manual curation or cDNA sequencing were incorporated into FlyBase, resulting in 150 new genes and modifications to hundreds of existing gene models. Finally, the 245 predictions that we did not assess were in non-coding regions of existing transcript models, or were already included in FlyBase independent of our study. In an independent analysis[52], we predicted 98 new genes on the

basis of inferred homology to predicted genes in the informant species[32], of which 63% matched the above predictions.

Discovering unusual features of protein-coding genes

Our analysis also predicted an abundance of unusual protein-coding genes that call for follow-up experimental investigation. First, we found open reading frames with clear protein-coding signatures and conserved start and stop sites on the transcribed strand of annotated UTRs, indicative of polycistronic transcripts[23,53,54]. These include 73% of 115 annotated dicistronic transcripts and 135 new candidate cistrons of 123 genes (Supplementary Fig. 2b).

Second, we predicted that 149 genes undergo stop codon readthrough, with protein-coding selection continuing past a deeply conserved stop codon (Fig. 3d), in some cases for hundreds of amino acids. It is unlikely that these genes are selenoproteins, as they appear to lack SECIS elements that direct selenocysteine recoding[55–58]. Other mechanisms may instead be at work, such as regulation of ribosomal release factors[59], A-to-I editing[39,60,61], alternative splicing, or other less-characterized mechanisms[62]. In fact, these genes are significantly enriched in neuronal proteins ($P = 10^{-4}$), which frequently undergo A-to-I editing[63].

Third, we found four genes in which CSF signatures abruptly shift from one reading frame to another in the absence of nearby intron–exon boundaries or insertions and deletions (Fig. 3e). These are suggestive of conserved 'programmed' frameshifts[64], which are thought to be rare in eukaryotes.

Overall, our results affected over 10% of protein-coding genes, and will be available in future releases of FlyBase. They also suggest that several types of unusual protein-coding gene structure may be more prevalent in the fly than previously appreciated.

RNA genes and structures

Several comparative approaches to RNA gene identification have been developed[6,7,65] that recognize their characteristic properties: compensatory double substitutions of paired nucleotides (for example, A•U↔C•G), structure-preserving single-nucleotide mutations involving G•U base pairs (G•U↔G•C and G•U↔A•U), and few nucleotide substitutions disrupting functional base pairs (Fig. 2b). To predict new structures, we applied EvoFold[7] in highly conserved segments of the 12 *Drosophila* species and focused on high-stringency candidates with strong support by compensatory changes (Supplementary Methods 4).

Our search led to 394 predictions, recovering 68 known RNA structures (primarily transfer RNA genes) in 0.02% of the genome (570-fold enrichment). The novel candidates consisted of 177 structures in intergenic regions (54%), 103 in introns (32%), 36 in 3′ UTRs (11%) and 10 in 5′ UTRs (3%). In addition, we predicted 200 structures in protein-coding regions (Supplementary Methods 3). Notably, 75% of 3′ UTR structures and 80% of 5′ UTR structures were predicted on the transcribed strand, suggesting that they are frequently part of the messenger RNA. In contrast, only 47% of intronic structures are on the transcribed strand, suggesting that they are largely independent of the surrounding genes.

Known and novel types of RNA genes

Of the 177 predicted intergenic structures, 30 were detected in a tiling-array expression study[42]. This fraction (17%) is significantly above that for all conserved intergenic regions (12%, $P = 0.007$), but lower than that of known intergenic ncRNAs (21%), suggesting that these candidates may be of lower abundance, temporally or spatially constrained, or might include false positives. Two predictions were expressed throughout development, one extending the annotation of a previously reported but uncharacterized ncRNA[66] and the other probably

representing a novel type of ncRNA. The predictions also included nine novel H/ACA-box small nucleolar RNA candidates in introns of ribosomal genes, known to frequently contain small nucleolar RNAs that guide post-transcriptional base modifications of ncRNAs[67].

Likely A-to-I editing structures

Many of the 48 intronic candidates on the transcribed strand and many of the 200 hairpins in coding sequence are probably involved in A-to-I editing or post-transcriptional regulation (Fig. 4a). Hairpins in coding sequence were associated with 11 of the 157 known editing sites (120-fold enrichment) and both intronic and coding-sequence hairpins showed a strong enrichment for ion-channel genes (6%, $P = 0.007$ and 10%, $P = 2\times10^{-12}$, respectively), known to be frequent editing targets. Editing is known to occur at multiple sites in the same gene[63], and we find an additional 10 hairpins in known editing targets, as well as 40 additional hairpins clustered in 18 genes not previously known to be edited (for example *huntingtin*[68], which harbours four predicted hairpins, more than any other gene). Intronic predictions also showed the highest abundance of compensatory substitutions: for example, *Resistant to dieldrin* (Fig. 2b) contained a 26-base-pair (bp) intronic hairpin flanked by exons known to be edited[69] with a striking 16 compensatory changes, *lodestar* showed one hairpin with 11 compensatory changes, and *Inverted repeat-binding protein* showed one hairpin with 10 compensatory substitutions (Fig. 4b).

Likely regulatory UTR structures

We predicted 38 structures in 3′ UTRs, a density twofold higher than the genomic average, whereas fewer than 10 such examples are currently known[70]. A considerable fraction of these lies in regulatory genes (14 out of 38; $P = 10^{-4}$), including several transcriptional regulators (for example, *cas*, *spen* and *Alh*), the tyrosine phosphatase *PTP-ER* and the translation initiation factor *eIF3-S8*. This suggests that many regulatory genes may themselves be regulated post-transcriptionally through these structures.

3′ UTR structures were also enriched for genes involved in mRNA localization (3 out of 38, $P = 2.7\times10^{-4}$), including *oo18 RNA-binding protein* (*orb*) and *staufen* (*stau*), both of which contain double-stranded RNA-binding domains, are involved in axis specification during oogenesis, and interact with the mRNA of maternal effect protein *oskar*. The hairpin in *orb* is known to be important for mRNA transport and localization[71], whereas the highly similar *stau* hairpin has not been previously described to our knowledge.

The ten structures found in 5′ UTRs probably contain binding sites for factors that regulate translation. For example, the fly homologue of yeast ribosomal protein *RPL24* contains a hairpin structure overlapping its start codon (Fig. 4c). This is interesting in light of high conservation upstream of the start codon in yeast ribosomal proteins[3,4], and findings that ribosomal proteins bind to their mRNAs and control translation in prokaryotes[72,73].

Conserved RNA structures in *roX2* recruit MSL

In an independent study[74], we searched for conserved regions in the non-coding *roX1* and *roX2* (*RNA on the X*) genes to gain insights into their function. Both RNAs are components of the MSL (Male-specific lethal) complex and are crucial for dosage compensation in male flies, inducing lysine 16 acetylation of histone H4, leading to upregulation of hundreds of genes on the X chromosome[75]. We identified several stem-loop structures with repeated sequence motifs (for example, GUUNUACG), and found that tandem repeats of one of these were sufficient to recruit MSL complexes to the X chromosome and to induce acetylation of lysine 16 of histone H4. Although this structure could not fully rescue roX-deficient males, our results suggest that it mediates MSL recruitment during *roX2*-dependent chromatin modification and

dosage compensation, illustrating the power of evolutionary evidence for directing experimental studies.

Prediction and characterization of miRNA genes

Focusing on specific classes of RNA genes markedly increases the accuracy of RNA gene prediction, reviewed in refs [35, 76] and illustrated here for *Drosophila* miRNA genes. The common biogenesis and function of miRNAs[77] lead to evolutionary and structural signatures (Fig. 2c) that can be used for their systematic *de novo* discovery[8–11]. Using such signatures in the 12 fly genomes (Supplementary Methods 4a, b), we predicted 101 miRNAs[78] (Supplementary Table 4d), which include 60 of the 74 verified Rfam miRNAs (81%), while spanning less than 0.006% of the fly genome (13,500-fold nucleotide enrichment).

Comparison of our predictions with high-throughput sequencing data of short RNA libraries from different stages and tissues of *D. melanogaster*[78,79] revealed that 84 of the 101 predictions (83%), including 24 of the 41 novel predictions (59%), were authentic miRNA genes (Fig. 5a and Supplementary Table 4d). An independent computational method[79] had 20 of its 45 novel predictions validated when used across six *Drosophila* species. Additional candidates may represent genuine miRNAs whose temporal or spatial expression pattern does not overlap with the surveyed libraries.

Several of the validated miRNAs were on the transcribed strand of introns or clustered with other miRNAs. For example, *mir-11* and *mir-998* (the vertebrate homologue of which, *mir-29*, has been implicated in cancer[80]) were both found in the last intron of *E2f*, and might be involved in cell-cycle regulation (Fig. 5b). Notably, two predictions overlapped exons of previously annotated protein-coding genes that were independently rejected above (Fig. 5c), providing an explanation for the previously observed transcripts of these annotations and highlighting the importance of specific signatures for genome annotation.

High-throughput sequencing data discovered an additional 50 miRNAs not found computationally[79,81], thereby illustrating the limitations of purely computational approaches. Some of these had precursor structures not seen previously for animal miRNAs, including unusually long hairpins[79] and hairpins corresponding to short introns (mirtrons)[81,82]. The remaining were often less broadly conserved or showed unusual conservation properties.

Signatures for mature miRNA annotation

The exact position of 5′ cleavage of mature miRNAs is important, because it dictates the core of the target recognition sequence[83–85]. This leads to unique structural and evolutionary signatures, including direct signals, present at the 5′ cleavage site, and indirect signals, stemming from the relationship of miRNAs with their target genes (Supplementary Methods 4a, c). Combined into a computational framework[78], these signatures predicted the exact start position in 47 of the 60 cloned Rfam miRNAs (78%), and were within 1 bp in 51 cases (85%). The method disagreed with the previous annotation in 9 of the 14 Rfam miRNAs that were not previously cloned, of which 6 were confirmed by sequencing reads[78,79], leading to marked changes in the inferred target spectrum (Fig. 5d). Prediction accuracy was significantly lower (41% exact, 61% within 1 nucleotide) for novel miRNAs, which, however, also showed less accurate processing *in vivo*[78,79].

New insights into miRNA function and biogenesis

We predicted targets for all conserved miRNAs identified by high-throughput sequencing[79] searching for conserved matches to the seed region (similar to ref. [86]) evaluated using the branch length score (Supplementary Methods 5a), a new scoring scheme described below. Whereas the resulting miRNA targeting network changed substantially[79], we found that the

novel and revised miRNAs shared many of their predicted targets with previously known miRNAs, resulting in a denser network with increased potential for combinatorial regulation[78,79].

For ten miRNA hairpins, the mature miRNA and the corresponding miRNA star sequence (miRNA*, the small RNA from the opposite arm of the hairpin) both appeared to be functional: both reached high computational scores and were frequently sequenced[78,79], often exceeding the abundance of many mature miRNAs (Supplementary Table 4e). The Hox miRNA *mir-10* showed a particularly striking example of a functional star sequence (Fig. 5e): both arms showed abundant reads, high scores and highly conserved Hox gene targets[78,79], suggesting a key role in Hox regulation.

In addition, for 20 miRNA loci, the anti-sense strand also folded into a high-scoring hairpin suggestive of a functional miRNA[78] (Supplementary Table 4f). Indeed, sequencing reads confirmed that four of these anti-sense hairpins are processed into small RNAs *in vivo*[79]. Thus, a single genomic miRNA locus may produce up to four miRNAs, each with distinct targets.

Regulatory motif discovery and characterization

Regulatory motifs recognized by proteins and RNAs to control gene expression have been difficult to identify due to their short length, their many weakly specified positions, and the varying distances at which they can act[87,88]. Recent studies have shown that comparative genomics of a small number of species can be used for motif discovery[3,4,12–14], on the basis of hundreds of conserved instances across the genome (Fig. 2d). Many related genomes should lead to increased discovery power, but also pose new challenges, arising from sequencing, assembly, or alignment artefacts, and from movement or loss of motif instances in individual species.

To account for the unique properties of regulatory motifs, we developed a phylogenetic framework to assess the conservation of each motif instance across many genomes[89]. Briefly, we searched for motif instances in each of the aligned genomes, and based on the set of species that contained them, we evaluated the total branch length over which the *D. melanogaster* motif instance appears to be conserved (Supplementary Methods 5a, b), which we call the branch length score (BLS). We used BLS for the discovery of novel motifs (this section) and for the prediction of individual functional motif instances (next section).

Predicted motifs recover known regulators

To discover motifs, we estimated the conservation level of candidate sequence patterns with a motif excess conservation (MEC) score compared to overall conservation levels in promoters, UTRs, introns, protein-coding exons and intergenic regions (Supplementary Methods 5a).

Our search in regions with roles in pre-transcriptional regulation resulted in 145 distinct motifs (Table 2), obtained by collapsing variants across 83 motifs discovered in promoters, 35 in enhancers, 20 in 5′ UTRs, 35 in core promoters, 30 in introns and 84 in the remaining intergenic regions. Motifs discovered in each region showed similar properties and large overlap: 66 (46%) were discovered independently in at least two regions and 40 (28%) in at least three, consistent with shared regulatory elements in these regions[90].

The 145 discovered motifs match 40 (46%) of the 87 known transcription factors in *Drosophila* (Supplementary Table 5c) compared to 8% expected at random ($P = 1 \times 10^{-20}$). Several of the non-discovered known motifs are involved in early anterior–posterior segmentation of the embryo, consistent with reports that they are largely non-conserved[91]; indeed, 74% of these did not exceed the conservation expected by chance in promoter regions.

Other non-discovered motifs often lacked characteristics expected for transcription factor motifs, suggesting that some may be spurious: 49% were unusually long (>10 nucleotides) compared to 23% of recovered ones, and showed only one or a few total instances genome-wide, suggestive of individual regulatory sites rather than motifs.

Tissue-specific and functional enrichment of novel motifs

The discovered motifs showed strong signals with respect to embryonic expression patterns (Fig. 6a). Overall, 75 (52%) were either enriched or depleted in genes expressed in at least one tissue, compared to 59% of known motifs and 3% of random controls. Motif depletion may represent either specific repressors for individual tissues, or activators excluded from these tissues. Motif depletion was found more generally in ubiquitously expressed genes (30% of discovered and 34% of known motifs compared with 1% expected at random), similar to findings for *in vivo* binding sites[92], and probably reflecting less complex regulation. We also found significant motif enrichment in groups of genetically interacting genes (collected by FlyBase) that often function in common developmental contexts or signalling pathways, genes of metabolic pathways (Kyoto Encyclopedia of Genes and Genomes, KEGG[93]), and genes with shared functions (GO).

In total, 68% of discovered and 70% of known motifs were enriched or depleted in one of the functional categories (14% random). Noteworthy examples include motif ME93 (GCAACA), which was more highly enriched in neuroblasts ($P = 4 \times 10^{-12}$) than either of the two well-known regulators of neuroblast development, *prospero* and *asense* ($P = 4 \times 10^{-5}$ and 2×10^{-7}, respectively). Similarly, motifs ME89 (CACRCAC), ME11 (MATTAAWNATGCR) and ME117 (MAAMNNCAA) were highly enriched in malpighian tubule ($P = 4 \times 10^{-7}$), trachea ($P = 4 \times 10^{-5}$) and surface glia (6×10^{-7}), respectively, in each case ranking above motifs for factors known to be important in these tissues (Supplementary Table 5c). These presumably correspond to as-yet-unknown regulators for these tissues.

Exclusion, clustering and positional constraints

A large number of motifs were depleted in coding sequence (57% of discovered versus 57% of known and 10% of random motifs, $P = 3 \times 10^{-18}$) and in 3′ UTRs (30% versus 22% and 0%, $P = 4 \times 10^{-11}$), suggesting specific exclusion similar to *in vivo* binding[92].

Many of the intergenic or intronic instances occurred in clusters, a property of motifs that has been used to identify enhancer elements[91,94–96]. We assessed increased conservation of motifs when found near other instances of the same motif (whether conserved or not, to correct for regional conservation biases), and found significant multiplicity for 19% of the discovered motifs (compared to 24% of known and 4% of random motifs).

In addition, 15 of the discovered motifs (10%) were significantly enriched near transcription start sites (compared to 14% of known and 1% of random motifs). Several were enriched at precise positions and preferred orientations (Fig. 6b), including close matches to several known core promoter motifs involved in transcription initiation[97]. For example, ME5 (STATAWAWR), which matches the TATA-box motif, displayed a sharp peak on the transcribed strand, 27 nucleotides upstream of the transcription start site. Similarly, ME120 (TCAGTT), corresponding to the known initiator motif (Inr) strongly peaked directly on the transcription start site, and ME54 (RCGYRCGY), which matches a known downstream promoter element (DPE), peaked 30 nucleotides downstream of the transcription start site.

Regulatory motifs involved in post-transcriptional regulation

We also used BLS/MEC to discover motifs involved in post-transcriptional regulation, and developed methods to distinguish motifs acting at the DNA level, motifs acting at the RNA

level and motifs stemming from protein-coding codon biases (Supplementary Methods 5a). Motifs acting post-transcriptionally at the RNA level generally showed highly asymmetric conservation[12], as functional instances can only occur on the transcribed strand. Indeed, 71 of 90 motifs (79%) discovered in 3′ UTRs showed strand-specific conservation (compared with only 3% of 5′ UTR motifs and 5% of intron motifs, suggesting that these act primarily in pre-transcriptional regulation).

Overall, 33 motifs discovered in 3′ UTRs were complementary to the 5′ end of Rfam miRNAs, recovering 72% of known miRNAs (68% of 5′ unique miRNA families). An additional 21 motifs matched to 5′ ends of novel miRNAs predicted above, of which 12 were validated experimentally[78,79], and 3 motifs matched uniquely to miRNA star sequences, all of which were abundantly expressed *in vivo* (Supplementary Table 4e).

We found 33 additional motifs in 3′ UTRs that were apparently not associated with miRNAs. MO40 (TGTANWTW) closely matches the Puf-family Pumilio motif[98]. MO32 (AATAAA) corresponds to the polyadenylation signal and displays both very strong conservation and a sharply defined distance preference with respect to the end of the annotated 3′ UTR ($P = 10^{-69}$). Finally, several motifs (for example, MO24 =TAATTTAT; MO94 =TTATTTT) are variants of known AU-rich elements, which are known to mediate mRNA instability and degradation[99].

MicroRNA targeting in protein-coding regions

Protein-coding regions can also harbour functional regulatory motifs, such as exonic splicing regulatory elements[100]. However, motif conservation is difficult to assess within protein-coding regions because of the overlapping selective pressures. Indeed, the most highly conserved nucleotide sequence patterns of length seven (7mers) in coding sequence showed strong reading-frame-biased conservation, suggesting that they reflect protein-coding constraints rather than regulatory roles at the DNA or RNA level (Fig. 6c).

MicroRNA motifs, which function at the RNA level, instead showed high conservation in all three reading frames, suggesting that they are specifically selected within coding regions for their RNA-level function. Indeed, previous studies have shown that miRNA motifs in coding regions are preferentially conserved in vertebrates[86], that they can lead to repression in experimental assays[101,102], and that they are avoided in genes co-expressed with the miRNA[103]. Frame-invariant conservation allows us to demonstrate the coding-region targeting of individual miRNAs, and also enables the *de novo* discovery of miRNA motifs in coding regions. Using frame-invariant conservation, we recovered 11 miRNA motifs within the top 20 coding-region motifs (Supplementary Table 5g), whereas using overall conservation required several hundred candidates to recover 11 miRNA motifs.

Moreover, 7mers complementary to different positions in the mature miRNA show a distinctive conservation pattern indicative of functional targeting in coding regions (Fig. 6d) and similar to that found in 3′ UTRs[12,83] (correlation coefficient 0.96). Finally, 6mers complementary to miRNA 5′ ends were depleted in coding exons of anti-target genes (Supplementary Fig. 5f), similar to findings for these genes' 3′ UTRs[103,104]. Overall, these results, together with findings in vertebrates[86,101–103], suggest that important miRNA targets have been overlooked by many target prediction methods[105] that have traditionally focused exclusively on 3′ UTR sequences.

Prediction of individual regulator binding sites

Previous methods for regulatory motif discovery[3,4,12–14] integrated conservation information over hundreds of motif instances across the genome, leading to an exceedingly

clear signal for motif discovery even if many of these instances are only marginally conserved. In contrast, the reliable identification of individual motif instances has been hampered by lack of neutral divergence and would require many related genomes[15–19]. In the absence of such data, previous studies have relied on motif clustering[91,94–96] or other sequence characteristics[106] to predict regulatory targets or regions.

With the availability of the 12 fly genomes, we inferred high-confidence instances of regulatory motifs by mapping the BLS of each motif instance to a confidence value (Supplementary Methods 5a). This value represents the probability that a motif instance is functional, on the basis of the conservation level of appropriate control motifs evaluated in the same type of region (promoter, 3′ UTR, coding, and so on). Because the number of conserved instances decreases much more rapidly for control motifs than for real motifs, the many genomes allowed us to reach high confidence values for many transcription factors and miRNAs, even at relatively modest BLS thresholds (Fig. 2e).

Conserved motif instances identify functional *in vivo* targets

We found that increasing confidence levels selected for functional instances for both transcription factor and miRNA motifs: the normalized fraction of transcription factor motif instances within promoter regions rose from 20% to 90%; that of miRNA motif instances within 3′ UTRs rose from 20% to 90%; and the fraction of miRNA motif instances on the transcribed strand of 3′ UTRs rose from 50% (uniform) to 100% (Fig. 7a); in each case selecting the regions and strands where the motifs are known to be functional.

We further assessed how predicted motif instances compared with *in vivo* targets in promoter regions, defined experimentally (without comparative information). We used a set of high-confidence direct CrebA targets[107] and three genome-wide chromatin immunoprecipitation (ChIP) data sets for Snail, Mef2 and Twist[92,108,109], and in each case found that the enrichment between conserved motif instances and known *in vivo* regions increased sharply for increasing confidence values (Fig. 7b).

We also found that a large fraction of motif instances in experimentally determined target regions was conserved (Fig. 7c): 76% of motif instances in direct CrebA targets and 90% of motif instances in experimentally supported miRNA targets[104,110] were recovered at 60% confidence. Although many of the miRNA targets stem from comparative predictions and are expected to be well conserved, their high recovery rate illustrates the increased sensitivity of the BLS measure compared to perfect conservation (Supplementary Fig. 7d). Similar results were found for motifs in known enhancers that were determined to be bound by ChIP (‘ChIP-bound’): 65% of Mef2 motifs, 65% of Snail motifs and 25% of Twist motifs were conserved (Fig. 7c).

ChIP-determined and conservation-determined targets show similar enrichment

To determine whether ChIP-bound motifs that lack conservation are biologically meaningful, we studied their enrichment in muscle gene promoters. We found that motifs that were both bound and evolutionarily conserved showed very strong correlation with muscle genes for all three factors: Mef2 showed eightfold enrichment, Twist showed sevenfold enrichment and Snail, a mesodermal repressor, showed threefold depletion for muscle genes. However, when only non-conserved sites were considered, the correlation dropped significantly to 1–2-fold for all three factors, suggesting that non-conserved ChIP-bound sites may be of decreased biological significance (Fig. 7d).

We also used the correlation with muscle genes to compare ChIP-on-chip and evolutionary conservation as two complementary methods for target identification (Fig. 7d). We found that

the enrichment of conservation-inferred targets was consistently higher than the enrichment of ChIP-inferred targets for each of the three factors. Finally, we assessed the functional significance of motif instances that were only found by the conservation approach, specifically excluding those in ChIP-bound regions, and found that these were also enriched in the same functional categories as ChIP-bound sites with comparable or higher functional correlations (Fig. 7d). This suggests that the additional conserved instances are indeed functional, probably reflecting the higher coverage of conservation-based approaches, which are not restricted to the experimental conditions surveyed, or that they may be bound *in vivo* yet missed by ChIP-on-chip technology[111,112].

In an independent study[113] we compared several strategies for the prediction of motif instances and *cis*-regulatory modules and found that using the 12 fly genomes led to substantial improvements. In another study, we reported the recovery of conserved motifs for several known regulators, including *Suppressor of Hairless*, in genes of the *Enhancer of split* complex[114].

A regulatory network of *D. melanogaster* at 60% confidence

Having established the accuracy of conserved motif instances, we present an initial regulatory network for *D. melanogaster* at 60% confidence (Supplementary Fig. 5i), containing 46,525 regulatory connections between 67 transcription factors and 8,287 genes, and 3,662 connections between 81 cloned miRNAs (clustered in 49 families with unique seed sequences) and 2,003 genes.

The distribution of predicted sites per target gene is highly nonuniform and indicative of varying levels of regulatory control. Genes with the highest number of sites appeared to be enriched in morphogenesis, organogenesis, neurogenesis and a variety of tissues, whereas ubiquitously expressed genes and maternal genes with housekeeping functions had the fewest sites[104]. Interestingly, transcription factors appeared to be more heavily targeted than other genes, both by transcription factors (10 sites versus 5.5 on average, $P = 10^{-15}$) and by miRNAs (2.3 versus 1.8 miRNAs, $P = 5 \times 10^{-5}$). Moreover, genes with many transcription factor sites also had many miRNA sites, and conversely, genes with few transcription factor sites also had few miRNA sites ($P = 10^{-4}$ and $P = 7 \times 10^{-3}$, respectively).

Several of the predicted regulatory connections have independent experimental support (Supplementary Table 5h), including direct regulation of *achaete* by Hairy[115], of *giant* by Bicoid[116], of *Enhancer of split* complex genes by Suppressor of Hairless[117], and of *bagpipe* by Tinman (known to cooperate in mesoderm induction and heart specification[118]). More generally, when tissue-specific expression data were available, we found that on average 46% of all targets were co-expressed with their factor in at least one tissue (Supplementary Fig. 5i), which is significantly higher than expected by chance ($P = 2 \times 10^{-3}$).

Scaling of comparative genomics power

Theoretical considerations and pilot studies on selected genomic regions showed that the discovery power of comparative methods scales with the number and phylogenetic distance of the species compared[16–20,46,119,120]. We extended these analyses by investigating the scaling of genome-wide discovery power using evolutionary signatures for each class of functional elements (Fig. 8), on the basis of the recovery of known elements using different subsets of informant species (at a fixed stringency).

We found that recovery consistently increased with the total number of informant species, and that multi-species comparisons outperformed pairwise comparisons within the same phylogenetic clade. When we examined subsets of informants with similar total branch length

(for example, several close species versus one distant species), multi-species comparisons sometimes performed better (protein-coding exons, ncRNAs), comparably (motifs), or worse (miRNAs) than pairwise comparisons. This complex relationship between total branch length and actual discovery power probably reflects imperfect genome assemblies/alignments, characteristics of each class of functional elements, and the specific methods we used. For example, ncRNA discovery probably benefits from observing more compensatory changes across more genomes, whereas miRNA discovery may be more sensitive to artefacts in low-coverage genomes, given the expected high conservation of miRNA arms.

As expected, longer elements were easier to discover than shorter elements. Long protein-coding exons (>300 nucleotides) were recovered at very high rates even with few species at close distances (leaving little room for improvement with additional species). In contrast, more informant species and larger distances were crucial for recovering short exons, miRNAs and regulatory motifs.

Notably, the optimal evolutionary distance for pairwise comparisons to *D. melanogaster* also seemed to depend on element length: for long protein-coding exons, the best pairwise informant was the closely related *D. erecta*, for exons of intermediate lengths *D. ananassae*, and for the shortest exons the distant *D. willistoni* (Supplementary Table 7a). Distant species were also optimal for other classes of short elements (ncRNAs, miRNAs and motifs, Fig. 8b–d). This suggests that a small number of species at close evolutionary distances may generally allow the discovery of long elements, possibly including clade-specific elements, whereas short clade-specific elements may not be reliably detectable without many genomes at close distances.

Finally, we investigated the effect of alignment choice on our results (Supplementary Fig. 8). We found high similarity between different alignment strategies for longer elements (>93% agreement for exons), whereas shorter elements showed larger discrepancies between alignments (81% and 59% agreement for miRNA and motif instances, respectively).

Although factors such as genome size, repeat density, pseudogene abundance and physiological differences might confound a simple analogy to the vertebrate phylogeny based on neutral branch length (Fig. 1c), our results suggest that comparisons spanning marsupials, birds and reptiles may prove surprisingly useful for biological signal discovery in the human genome.

Discussion

Our results demonstrate the potential of comparative genomics for the systematic characterization of functional elements in a complete genome. Even in a species as intensely studied as *D. melanogaster*, our methods predicted several thousand new functional elements, including protein-coding genes and exons, novel RNA genes and structures, miRNA genes, regulatory motifs, and regulator targets. Our novel predictions have overwhelming statistical support, often surpassing that of known functional elements, and are additionally supported by experimental evidence in hundreds of cases. The common underlying methodology in this study has been the recognition of specific evolutionary signatures associated with each class of functional elements, which can be much more informative for genome annotation than overall measures of nucleotide conservation. These signatures are general and are immediately relevant to the analysis of the human genome and more generally of any species.

In addition to the many new elements, we gained specific biological insights and formulated hypotheses that we hope will guide follow-up experiments. We found 149 genes with potential translational readthrough, showing protein-like evolution downstream of a highly conserved stop codon, and possibly encoding additional protein domains or peptides specific to certain developmental contexts. We also found several candidate programmed frameshifts, which

might be part of regulatory circuits (as for *ODC*/*Oda* [64]) or help expand the diversity of protein products generated from one mRNA, similar to their role in prokaryotes[121]. We also presented evidence of miRNA processing from both arms of a miRNA hairpin and from both DNA strands of a miRNA locus in some cases, potentially leading to as many as four functional miRNAs per locus. As miRNA/miRNA* pairs are expressed from a single precursor and thus co-regulated, whereas sense/anti-sense pairs are expressed from distinct promoters, the use of both arms or both strands provides compelling general building blocks for higher-level miRNA-mediated regulation.

The newly discovered elements did not dramatically increase the total number of annotated nucleotides. Known and predicted elements explain 42% of nucleotides in phastCons elements[33], compared to 35.5% for previous annotations (Supplementary Fig. 6), an 18% increase (mostly owing to conserved motif instances). The remaining phastCons elements and independent estimates based on transcriptional activity[42] would suggest that a much higher fraction of the genome may be functional (Supplementary Fig. 6). Although it is possible that these estimates are artificially high and that we are in fact converging on a complete annotation of the fly genome, they might instead indicate that much remains to be discovered, which may require the recognition of as-yet-unknown classes of functional elements with distinct evolutionary signatures.

Our results also allowed us to compare and contrast evolutionary and experimental methods for the recovery of functional elements, particularly for the identification of regulator targets. We found that comparative genomics resulted in many functionally meaningful sites for transcription factors Mef2, Twist and Snail outside ChIP-bound regions, probably representing targets from diverse conditions not surveyed experimentally. Similarly, ChIP resulted in many additional sites outside those recovered by comparative genomics: some of these may have been replaced by functionally equivalent non-orthologous sequence, rendering them apparently non-conserved in sequence alignments[122–124]; others may have species- or lineage-specific roles, thus lacking sufficient signal for their comparative detection; finally, some bound sites may be biochemically active yet selectively neutral[125]. It is worth noting, however, that ChIP-bound motifs that were not conserved showed decreased enrichment in muscle/mesoderm development where the factors are known to act, suggesting that potential lineage-specific roles may lie outside the regulators' conserved functions. To resolve these questions, comparative genomics studies would benefit greatly from experimental studies in several related species in parallel.

Overall, comparative genomics and species-specific experimental studies provide complementary approaches to biological signal discovery. Comparative studies help pinpoint evolutionarily selected functional elements across diverse conditions, whereas experimental studies reveal stage- and tissue-specific information, as well as species-specific sites. Ultimately, their integration is a necessary step towards a comprehensive understanding of animal genomes.

METHODS SUMMARY

The Methods are described in Supplementary Information, with more details found in the cited companion papers for each section. The sections of the Supplementary Methods are arranged in the same order as the manuscript to facilitate cross-referencing, with an index on the first page to aid navigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the National Human Genome Research Institute (NHGRI) for continued support. A.S. was supported in part by the Schering AG/Ernst Schering Foundation and in part by the Human Frontier Science Program Organization (HFSPO). P.K. was supported in part by a National Science Foundation Graduate Research Fellowship. J.S.P. thanks B. Raney and R. Baertsch, and the Danish Medical Research Council and the National Cancer Institute for support. J.B. thanks the Schering AG/Ernst Schering Foundation for a postdoctoral fellowship. L. Parts thanks J. Vilo. S.R. was supported by a HHMI-NIH/NIBIB Interfaces Training Grant and thanks T. Lane and M. Werner-Washburne. D.H., D.P.B., G.J.H. and T.C.K. are Investigators of the Howard Hughes Medical Institute, and B.P., J.G.R., E.H. and J.B. are affiliated with these investigators. J.W.C. and S.E.C. were supported by the NHGRI. M.K. was supported by start-up funds from the MIT Electrical Engineering and Computer Science Laboratory, the Broad Institute of MIT and Harvard, and the MIT Computer Science and Artificial Intelligence Laboratory, and by the Distinguished Alumnus (1964) Career Development Professorship.

References

1. Miller W, Makova KD, Nekrutenko A, Hardison RC. Comparative genomics. Annu Rev Genomics Hum Genet 2004;5:15–56. [PubMed: 15485342]
2. Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nature Rev Genet 2003;4:251–262. [PubMed: 12671656]
3. Kellis M, et al. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 2003;423:241–254. [PubMed: 12748633]
4. Cliften P, et al. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 2003;301:71–76. [PubMed: 12775844]
5. Brent MR. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res 2005;15:1777–1786. [PubMed: 16339376]
6. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 2005;102:2454–2459. [PubMed: 15665081]
7. Pedersen JS, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2006;2:e33. [PubMed: 16628248]
8. Lim LP, et al. The microRNAs of *Caenorhabditis elegans*. Genes Dev 2003;17:991–1008. [PubMed: 12672692]
9. Lim LP, et al. Vertebrate microRNA genes. Science 2003;299:1540. [PubMed: 12624257]
10. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of *Drosophila* microRNA genes. Genome Biol 2003;4:R42. [PubMed: 12844358]
11. Berezikov E, et al. Phylogenetic shadowing and computational identification of human microRNA genes. Cell 2005;120:21–24. [PubMed: 15652478]
12. Xie X, et al. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 2005;434:338–345. [PubMed: 15735639]
13. Ettwiller L, et al. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. Genome Biol 2005;6:R104. [PubMed: 16356267]
14. Chan CS, Elemento O, Tavazoie S. Revealing posttranscriptional regulatory elements through network-level conservation. PLoS Comput Biol 2005;1:e69. [PubMed: 16355253]
15. Boffelli D, et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 2003;299:1391–1394. [PubMed: 12610304]
16. Cooper GM, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15:901–913. [PubMed: 15965027]
17. Margulies EH, Blanchette M, Haussler D, Green ED. Identification and characterization of multi-species conserved sequences. Genome Res 2003;13:2507–2518. [PubMed: 14656959]
18. Thomas JW, et al. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 2003;424:788–793. [PubMed: 12917688]
19. Eddy SR. A model of the statistical power of comparative genome sequence analysis. PLoS Biol 2005;3:e10. [PubMed: 15660152]
20. Bergman CM, et al. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol 2002;3:RESEARCH0086. [PubMed: 12537575]

21. Rubin GM, Lewis EB. A brief history of *Drosophila*'s contributions to genome research. Science 2000;287:2216–2218. [PubMed: 10731135]
22. Adams MD, et al. The genome sequence of *Drosophila melanogaster*. Science 2000;287:2185–2195. [PubMed: 10731132]
23. Misra S, et al. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol 2002;3:RESEARCH0083. [PubMed: 12537572]
24. Celniker SE, Rubin GM. The *Drosophila melanogaster* genome. Annu Rev Genomics Hum Genet 2003;4:89–117. [PubMed: 14527298]
25. Ashburner M, Bergman C. M *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. Genome Res 2005;15:1661–1667. [PubMed: 16339363]
26. Matthews KA, Kaufman TC, Gelbart WM. Research resources for *Drosophila*: the expanding universe. Nature Rev Genet 2005;6:179–193. [PubMed: 15738962]
27. Venken KJ, He Y, Hoskins RA, Bellen HJ. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. Science 2006;314:1747–1751. [PubMed: 17138868]
28. Dietzl G, et al. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. Nature 2007;448:151–156. [PubMed: 17625558]
29. Spradling AC, et al. The Berkeley *Drosophila* Genome Project gene disruption project: Single P-element insertions mutating 25% of vital *Drosophila* genes. Genetics 1999;153:135–177. [PubMed: 10471706]
30. St Johnston D. The art and design of genetic screens: *Drosophila melanogaster*. Nature Rev Genet 2002;3:176–188. [PubMed: 11972155]
31. Richards S, et al. Comparative genome sequencing of *Drosophila* pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res 2005;15:1–18. [PubMed: 15632085]
32. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 200710.1038/nature06341this issue
33. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–1050. [PubMed: 16024819]
34. Nekrutenko A, Makova KD, Li WH. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res 2002;12:198–202. [PubMed: 11779845]
35. Eddy SR. Computational genomics of noncoding RNA genes. Cell 2002;109:137–140. [PubMed: 12007398]
36. Bompfuenewerer AF, et al. Evolutionary patterns of non-coding RNAs. Theor Biosci 2004;123:301–369.
37. Reese MG, et al. Genome annotation assessment in *Drosophila melanogaster*. Genome Res 2000;10:483–501. [PubMed: 10779488]
38. Rubin GM, et al. A *Drosophila* complementary DNA resource. Science 2000;287:2222–2224. [PubMed: 10731138]
39. Stapleton M, et al. A *Drosophila* full-length cDNA resource. Genome Biol 2002;3:RESEARCH0080. [PubMed: 12537569]
40. Hild M, et al. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. Genome Biol 2003;5:R3. [PubMed: 14709175]
41. Yandell M, et al. A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. Proc Natl Acad Sci USA 2005;102:1566–1571. [PubMed: 15668397]
42. Manak JR, et al. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. Nature Genet 2006;38:1151–1158. [PubMed: 16951679]
43. Lin MF, et al. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using twelve fly genomes. Genome Res. 10.1101/gr.6679507in the press
44. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 2000;15:496–503. [PubMed: 11114436]

45. Mignone F, Grillo G, Liuni S, Pesole G. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucleic Acids Res 2003;31:4639–4645. [PubMed: 12888525]
46. Zhang L, Pavlovic V, Cantor CR, Kasif S. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. Genome Res 2003;13(6A):1190–1202. [PubMed: 12743024]
47. Crosby MA, et al. FlyBase: genomes by the dozen. Nucleic Acids Res 2007;35(Database issue):D486–D491. [PubMed: 17099233]
48. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet 2000;25:25–29. [PubMed: 10802651]
49. Ochman H, Ajioka JW, Garza D, Hartl DL. Inverse polymerase chain reaction. Bio/Technology 1990;8:759–760. [PubMed: 1366903]
50. Hoskins RA, et al. Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). Nucleic Acids Res 2005;33:e185. [PubMed: 16326860]
51. Wan KH, et al. High-throughput plasmid cDNA library screening. Nature Protocols 2006;1:624–632.
52. Hahn MW, Han MV, Han SG. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet 2007;3:e197. [PubMed: 17997610]
53. Andrews J, et al. The stoned locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides. Genetics 1996;143:1699–1711. [PubMed: 8844157]
54. Brogna S, Ashburner M. The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. EMBO J 1997;16:2023–2031. [PubMed: 9155028]
55. Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the genetic code. Mol Cell Biol 2002;22:3565–3576. [PubMed: 11997494]
56. Kryukov GV, et al. Characterization of mammalian selenoproteomes. Science 2003;300:1439–1443. [PubMed: 12775843]
57. Copeland PR. Regulation of gene expression by stop codon recoding: selenocysteine. Gene 2003;312:17–25. [PubMed: 12909337]
58. Castellano S, et al. *In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. EMBO Rep 2001;2:697–702. [PubMed: 11493597]
59. von der Haar T, Tuite MF. Regulated translational bypass of stop codons in yeast. Trends Microbiol 2007;15:78–86. [PubMed: 17187982]
60. Luo GX, et al. A specific base transition occurs on replicating hepatitis delta virus RNA. J Virol 1990;64:1021–1027. [PubMed: 2304136]
61. Casey JL, Gerin JL. Hepatitis D virus RNA editing: specific modification of adenosine in the antigenomic RNA. J Virol 1995;69:7593–7600. [PubMed: 7494266]
62. Steneberg P, et al. Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the *Drosophila* trachea. Genes Dev 1998;12:956–967. [PubMed: 9531534]
63. Bass BL. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem 2002;71:817–846. [PubMed: 12045112]
64. Ivanov IP, et al. The *Drosophila* gene for antizyme requires ribosomal frameshifting for expression and contains an intronic gene for snRNP Sm D3 on the opposite strand. Mol Cell Biol 1998;18:1553–1561. [PubMed: 9488472]
65. Eddy SR. Non-coding RNA genes and the modern RNA world. Nature Rev Genet 2001;2:919–929. [PubMed: 11733745]
66. Yuan G, et al. RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. Nucleic Acids Res 2003;31:2495–2507. [PubMed: 12736298]
67. Lestrade L, Weber M. J snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res 2006;34(Database issue):D158–D162. [PubMed: 16381836]
68. Bier E. *Drosophila*, the golden bug, emerges as a tool for human genetics. Nature Rev Genet 2005;6:9–23. [PubMed: 15630418]
69. Hoopengardner B, Bhalla T, Staber C, Reenan R. Nervous system targets of RNA editing identified by comparative genomics. Science 2003;301:832–836. [PubMed: 12907802]

70. Mignone F, et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 2005;33(Database issue):D141–D146. [PubMed: 15608165]
71. Cohen RS, Zhang S, Dollar GL. The positional, structural, and sequence requirements of the *Drosophila* TLS RNA localization element. RNA 2005;11:1017–1029. [PubMed: 15987813]
72. Allemand F, et al. *Escherichia coli* ribosomal protein L20 binds as a single monomer to its own mRNA bearing two potential binding sites. Nucleic Acids Res 2007;35:3016–3031. [PubMed: 17439971]
73. Okumura T, Matsumoto A, Tanimura T, Murakami R. An endoderm-specific GATA factor gene, dGATAe, is required for the terminal differentiation of the *Drosophila* endoderm. Dev Biol 2005;278:576–586. [PubMed: 15680371]
74. Park SW, et al. An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. Genetics. in the press
75. Park Y, Kuroda MI. Epigenetic aspects of X-chromosome dosage compensation. Science 2001;293:1083–1085. [PubMed: 11498577]
76. Berezikov E, Cuppen E, Plasterk RH. Approaches to microRNA discovery. Nature Genet 2006;38 (Suppl 1):S2–S7. [PubMed: 16736019]
77. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–297. [PubMed: 14744438]
78. Stark A, et al. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. Genome Res. 10.1101/gr.6593807in the press
79. Ruby JG, et al. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. Genome Res. 10.1101/gr.6597907in the press
80. Pekarsky Y, et al. Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. Cancer Res 2006;66:11590–11593. [PubMed: 17178851]
81. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature 2007;448:83–86. [PubMed: 17589500]
82. Okamura K, et al. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. Cell 2007;130:89–100. [PubMed: 17599402]
83. Lewis BP, et al. Prediction of mammalian microRNA targets. Cell 2003;115:787–798. [PubMed: 14697198]
84. Stark A, Brennecke J, Russell RB, Cohen SM. Identification of *Drosophila* microRNA targets. PLoS Biol 2003;1:E60. [PubMed: 14691535]
85. Lai EC. Micro RNAs are complementary to 3′UTR sequence motifs that mediate negative post-transcriptional regulation. Nature Genet 2002;30:363–364. [PubMed: 11896390]
86. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005;120:15–20. [PubMed: 15652477]
87. Tompa M. Identifying functional elements by comparative DNA sequence analysis. Genome Res 2001;11:1143–1144. [PubMed: 11435394]
88. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics 2000;16:16–23. [PubMed: 10812473]
89. Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. Genome Res. 10.1101/gr.7090407in the press
90. Stathopoulos A, Levine M. Genomic regulatory networks and animal development. Dev Cell 2005;9:449–462. [PubMed: 16198288]
91. Schroeder MD, et al. Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biol 2004;2:e271. [PubMed: 15340490]
92. Zeitlinger J, et al. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. Genes Dev 2007;21:385–390. [PubMed: 17322397]
93. Kanehisa M, et al. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32 (Database issue):D277–D280. [PubMed: 14681412]

94. Berman BP, et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci USA 2002;99:757–762. [PubMed: 11805330]
95. Markstein M, et al. A regulatory code for neurogenic gene expression in the *Drosophila* embryo. Development 2004;131:2387–2394. [PubMed: 15128669]
96. Philippakis AA, et al. Expression-guided *in silico* evaluation of candidate cis regulatory codes for *Drosophila* muscle founder cells. PLoS Comput Biol 2006;2:e53. [PubMed: 16733548]
97. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. Annu Rev Biochem 2003;72:449–479. [PubMed: 12651739]
98. Gerber AP, et al. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. Proc Natl Acad Sci USA 2006;103:4487–4492. [PubMed: 16537387]
99. Zubiaga AM, Belasco JG, Greenberg ME. The nonamer UUAUUUAUU is the key AU-rich sequence motif that mediates mRNA degradation. Mol Cell Biol 1995;15:2219–2230. [PubMed: 7891716]
100. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science 2002;297:1007–1013. [PubMed: 12114529]
101. Kloosterman WP, Wienholds E, Ketting RF, Plasterk RH. Substrate requirements for let-7 function in the developing zebrafish embryo. Nucleic Acids Res 2004;32:6284–6291. [PubMed: 15585662]
102. Grimson A, et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007;27:91–105. [PubMed: 17612493]
103. Farh KK, et al. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science 2005;310:1817–1821. [PubMed: 16308420]
104. Stark A, et al. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′ UTR evolution. Cell 2005;123:1133–1146. [PubMed: 16337999]
105. Rajewsky N. microRNA target predictions in animals. Nature Genet 2006;38(suppl 1):S8–S13. [PubMed: 16736023]
106. Elnitski L, et al. Distinguishing regulatory DNA from neutral sites. Genome Res 2003;13:64–72. [PubMed: 12529307]
107. Abrams EW, Andrew DJ. CrebA regulates secretory activity in the *Drosophila* salivary gland and epidermis. Development 2005;132:2743–2758. [PubMed: 15901661]
108. Sandmann T, et al. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. Dev Cell 2006;10:797–807. [PubMed: 16740481]
109. Sandmann T, et al. A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. Genes Dev 2007;21:436–449. [PubMed: 17322403]
110. Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA 2006;12:192–197. [PubMed: 16373484]
111. Lee TI, et al. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 2006;125:301–313. [PubMed: 16630818]
112. Boyer LA, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 2005;122:947–956. [PubMed: 16153702]
113. Aerts S, van Helden J, Sand O, Hassan B. Fine-tuning enhancer models to predict transcriptional targets across multiple genomes. PLoS ONE 2007;2(11):e1115. [PubMed: 17973026]
114. Maeder M, Polansky B, Robson B, Eastman D. Phylogenetic footprinting analysis in the upstream regulatory regions of the *Drosophila* Enhancer of split genes. Genetics. in the press
115. Van Doren M, et al. Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. Genes Dev 1994;8:2729–2742. [PubMed: 7958929]
116. Kraut R, Levine M. Spatial regulation of the gap gene giant during *Drosophila* development. Development 1991;111:601–609. [PubMed: 1893877]
117. Bailey AM, Posakony JW. Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. Genes Dev 1995;9:2609–2622. [PubMed: 7590239]
118. Yin Z, Frasch M. Regulation and function of tinman during dorsal mesoderm induction and heart specification in *Drosophila*. Dev Genet 1998;22:187–200. [PubMed: 9621427]

119. Margulies EH, et al. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci USA 2005;102:4795–4800. [PubMed: 15778292]

120. Margulies EH, Chen CW, Green ED. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends Genet 2006;22:187–193. [PubMed: 16499991]

121. Farabaugh PJ. Programmed translational frameshifting. Annu Rev Genet 1996;30:507–528. [PubMed: 8982463]

122. Odom DT, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature Genet 2007;39:730–732. [PubMed: 17529977]

123. Ludwig MZ, Kreitman M. Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. Mol Biol Evol 1995;12:1002–1011. [PubMed: 8524036]

124. Ludwig MZ, et al. Functional evolution of a cis-regulatory module. PLoS Biol 2005;3:e93. [PubMed: 15757364]

125. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799–816. [PubMed: 17571346]

126. Kent WJ, et al. The human genome browser at UCSC. Genome Res 2002;12:996–1006. [PubMed: 12045153]

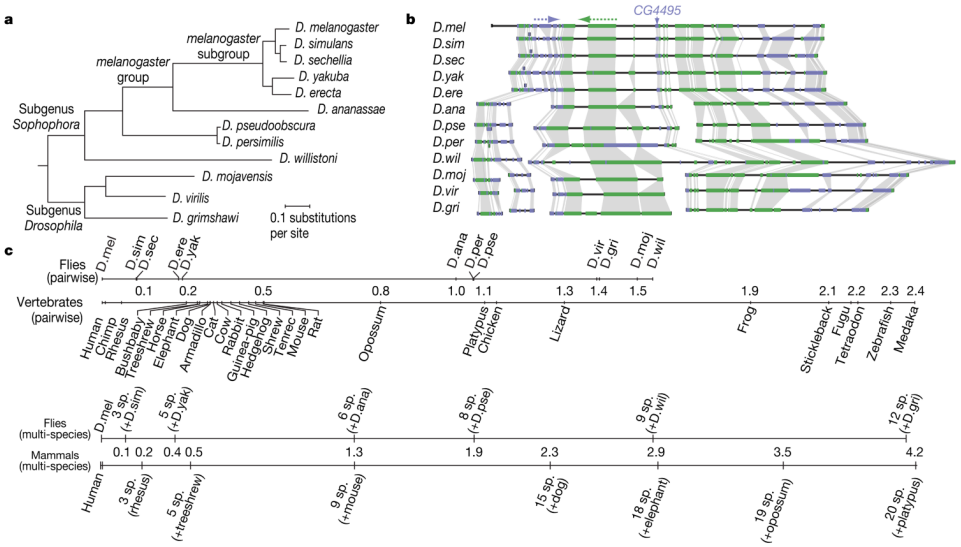

Figure 1. Phylogeny and alignment of 12 *Drosophila* species

a, Phylogenetic tree relating the 12 *Drosophila* species, estimated from fourfold degenerate sites (Supplementary Methods 1). The 12 species span a total branch length of 4.13 substitutions per neutral site. **b**, Gene order conservation for a 0.45-Mb region of chromosome 2L centred on *CG4495*, for which we predict a new exon (Fig. 3a), and spanning 35 genes. Colour represents the direction of transcription. Boxes represent full gene models. Individual exons and introns are not shown. **c**, Comparison of evolutionary distances spanned by fly and vertebrate trees. Pairwise and multi-species distances (in substitutions per fourfold degenerate site) are shown from *D. melanogaster* and from human as reference genomes. Note that species with longer branches (for example, mouse) show higher pairwise distances, not always reflecting the order of divergence. Multi-species distances include all species within a phylogenetic clade.

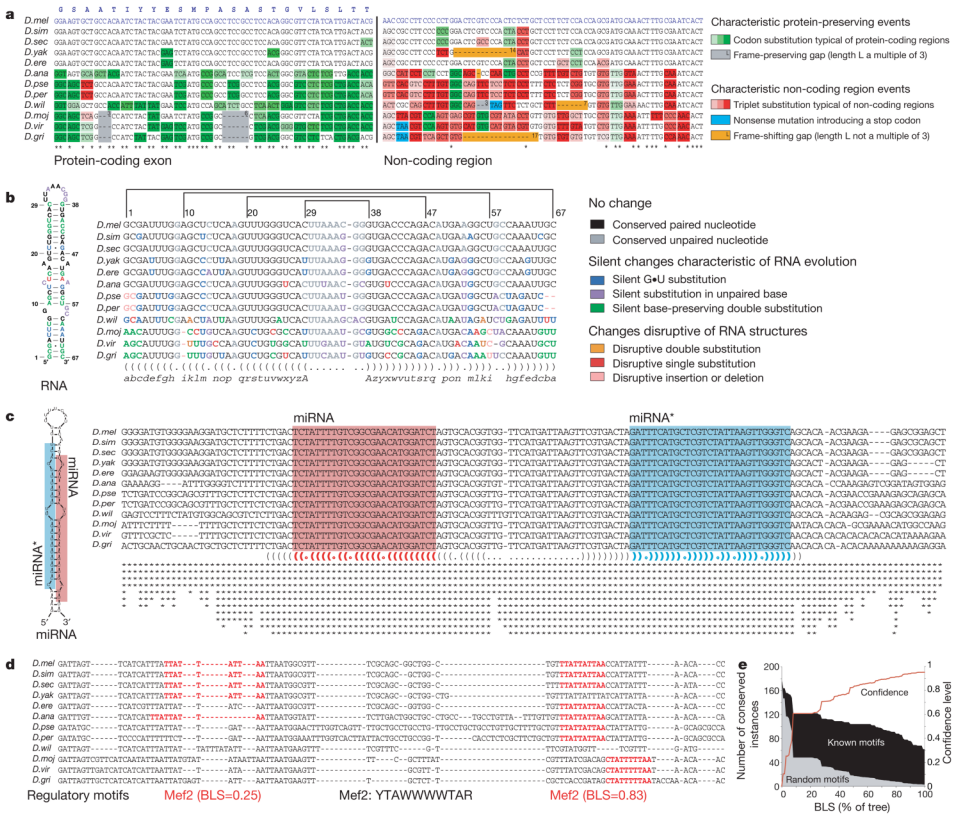

Figure 2. Distinct evolutionary signatures for diverse classes of functional elements

a, Protein-coding genes tolerate mutations that preserve the amino-acid translation, leading to abundant conservative codon substitutions (green). Insertions and deletions are largely constrained to be a multiple of three (grey). In contrast, non-coding regions show abundant non-conservative triplet substitutions (red), nonsense mutations (blue) and frame-shifting insertions and deletions (orange). **b**, RNA genes tolerate mutations that preserve the secondary structure (for example, single substitutions involving G•U base pairs and compensatory changes) and exclude structure-disrupting mutations. Matching parentheses and matching letters of the alphabet indicate paired bases. **c**, MicroRNA genes, in contrast, generally do not show changes in stem regions, but tolerate substitutions in loop regions and flanking unpaired regions, leading to a distinctive conservation profile. Asterisks denote the number of informant species matching the *melanogaster* sequence at each position. **d**, Regulatory motifs tolerate local movement and nucleotide substitutions consistent with their degeneracy patterns, and show increased conservation across the phylogenetic tree, measured as the branch length score (BLS; Supplementary Methods 5a). **e**, Increasing BLS thresholds select for instances of known motifs (black) at increasing confidence (red), as the number of conserved instances of control motifs (grey) drops significantly faster.

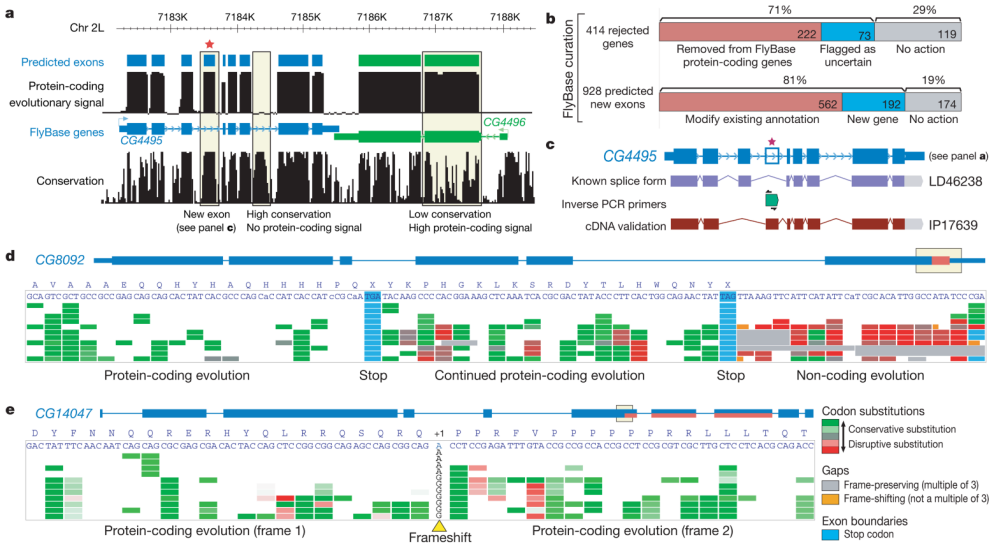

Figure 3. Revisiting the protein-coding gene catalogue and revealing unusual gene structures

a, Protein-coding evolutionary signatures correlate with annotated protein-coding exons more precisely than the overall conservation level (phastCons track[33]), for example excluding highly conserved yet non-coding elements. Asterisk denotes new predicted exon, which we validate with cDNA sequencing (see panel **c**). The height of the black tracks indicates protein-coding potential according to evolutionary signatures (top) and overall sequence conservation (bottom). Blue and green boxes indicate predicted coding exons (top) and the current FlyBase annotation (bottom). The region shown represents the central 6 kb of Fig. 1b, rendered by the UCSC genome browser[126]. **b**, Results of FlyBase curation of 414 genes rejected by evolutionary signatures (Table 1), and 928 predicted new exons. **c**, Experimental validation of predicted new exon from panel **a**. Inverse PCR with primers in the predicted exon (green) results in a full-length cDNA clone, confirming the predicted exon and revealing a new alternative splice form for *CG4495*. **d**, Protein-coding evolution continues downstream of a conserved stop codon in 149 genes, suggesting translational readthrough. **e**, Codon-based evolutionary signatures (CSF score) abruptly shift from one reading frame to another within a protein-coding exon, suggesting a conserved, ‘programmed’ frameshift.

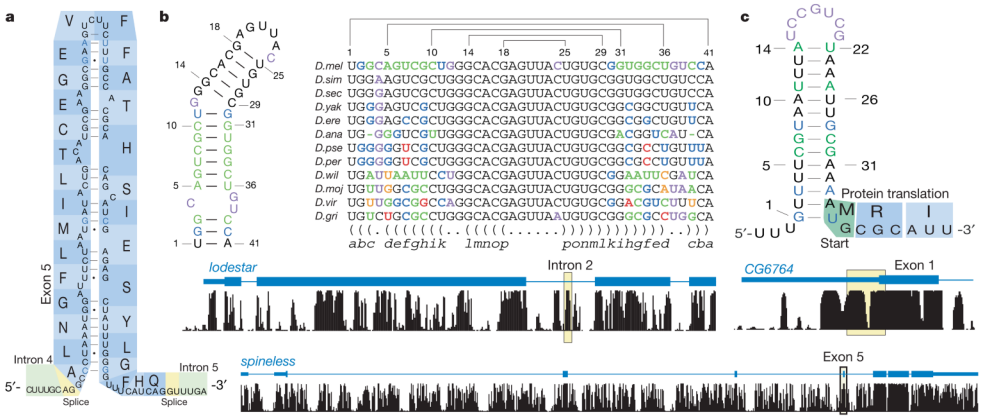

Figure 4. Novel RNA structures

a, New exonic RNA structure spanning 78 of 90 nucleotides of *spineless* exon 5. **b**, New intronic RNA structure in *lodestar* shows 11 compensatory substitutions and 10 silent G•U substitutions, providing strong evidence of structural selection (colours as in Fig. 2b). **c**, New 5′ UTR structure that overlaps the translation start site of *CG6764*, the fly orthologue of yeast ribosomal protein *RPL24*, suggesting a potential role in translational regulation. **a**–**c**, Structure shown corresponds to shaded region in the gene model.

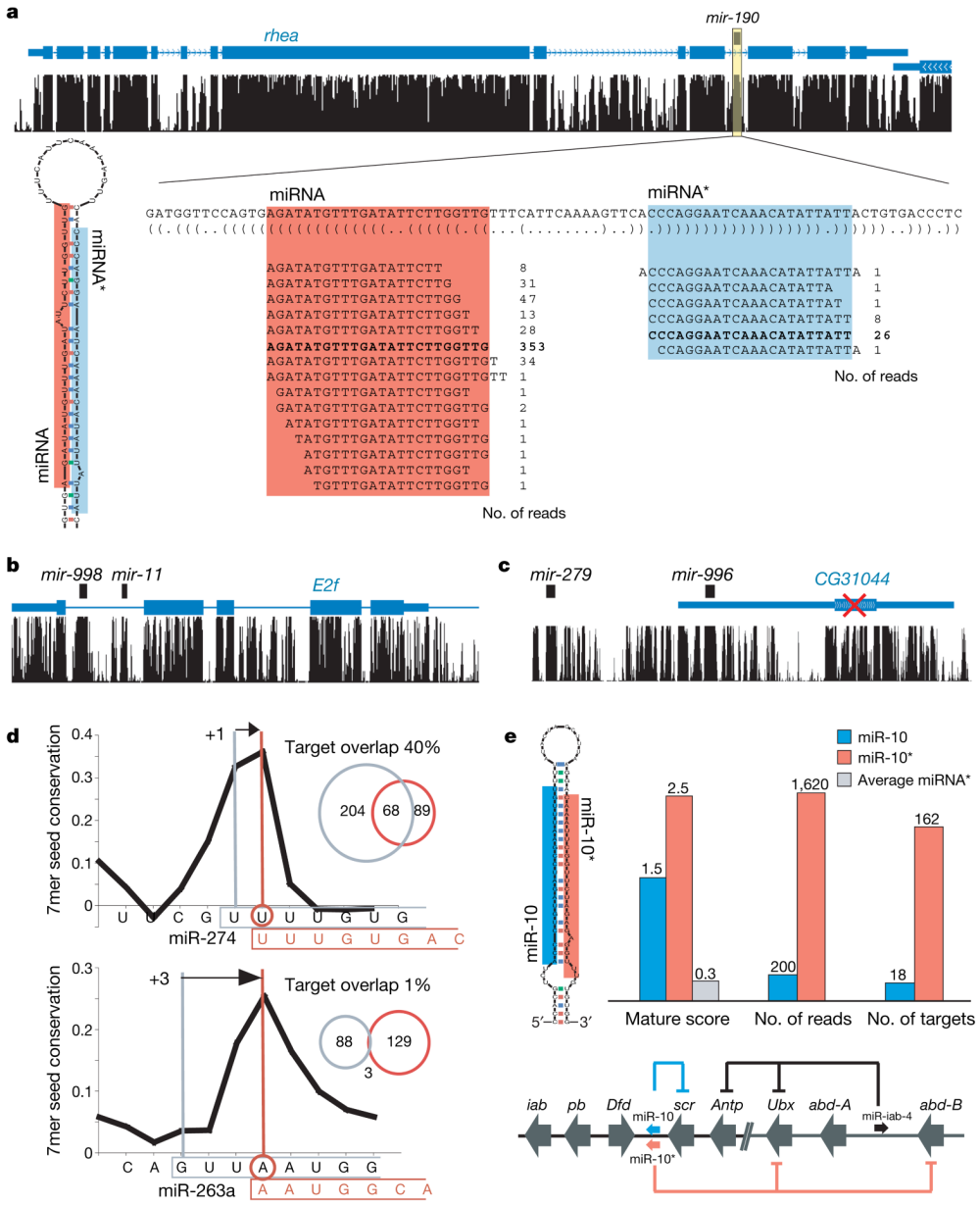

Figure 5. MicroRNA gene identification and functional implications

a, New predicted miRNA (*mir-190*) and its validation by sequencing reads. Total read counts for mature miRNA (red) and miRNA* (blue) show a characteristic pattern of processing indicative of miRNAs. Highlighted regions indicate most abundant processing products. **b**, Example of clustered known (*mir-11*) and new (*mir-998*) miRNAs in the intron of cell-cycle regulator *E2f*. **c**, Example of a new miRNA (*mir-996*) in the transcript of a spurious gene. *CG31044* was rejected by our protein-coding analysis, its transcript probably representing the precursor of *mir-996*, with no protein-coding function. **d**, Revisions to the 5′ end of miR-274 and miR-263a are proposed on the basis of evolutionary evidence (for example, 7mer seed conservation; black curve) and confirmed by sequencing reads. Changes at the 5′ end of more than one nucleotide results in marked changes to the predicted target spectra (venn diagrams). **e**, Evidence from evolutionary signals (mature score), sequencing reads and target predictions suggests that both miR-10 and miR-10* are functional, each targeting distinct Hox genes.

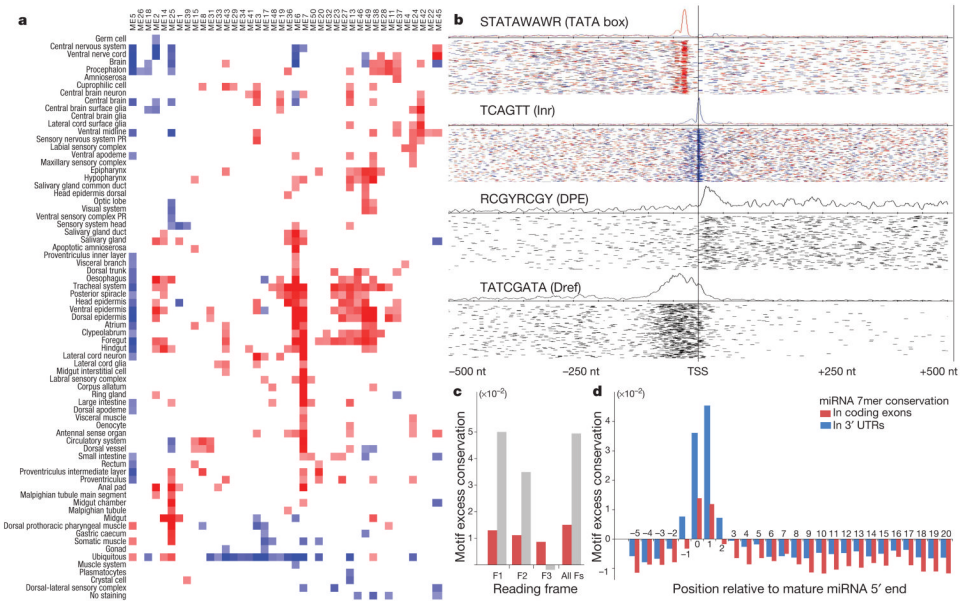

Figure 6. Regulatory motif discovery

a, Discovered motifs show enrichment (red) or depletion (blue) in genes expressed in a given tissue (log colour range from $P = 10^{-5}$ enrichment to $P = 10^{-5}$ depletion). Bi-clustering reveals groups of motifs with similar tissue enrichment and groups of tissues with similar motif content. Full matrix and randomized control is shown in Supplementary Fig. 6d. **b**, Positional bias of discovered motifs relative to transcription start sites (TSS). Peaks with highly specific distances from the transcription start site (for example, first three plots) are characteristic of core promoter elements, and broad peaks (for example, fourth plot) are characteristic of transcription factors. For non-palindromic motifs, colours indicate forward-strand (red) and reverse-strand (blue) instances. Curves denote the density of all instances and individual segments denote individual motif instances, summed across groups of 50 genes (each line). **c**, Coding regions show reading-frame-invariant conservation for miRNA motifs (red) and reading-frame-biased conservation for protein motifs (grey). MEC scores are evaluated for each of the three reading frame offsets (F1–F3) and also without frame correction (all Fs). Plots show average MEC for all miRNA motifs and 500 top-scoring protein-coding motifs (based on MEC without frame correction). **d**, Motif excess conservation (MEC) of 7mer complements at different offsets with respect to miRNA 5′ end, averaged across all Rfam miRNAs. MEC scores evaluated in protein-coding regions and 3′ UTRs show a highly similar profile (correlation coefficient 0.96), suggesting similar evolutionary constraints.

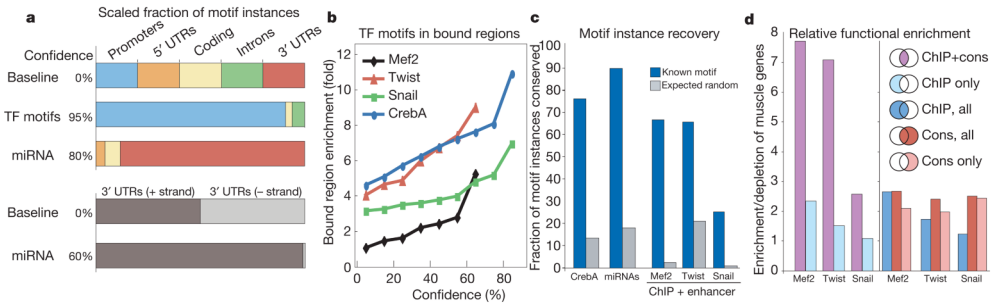

Figure 7. Identification of individual motif instances

a, Increasing confidence levels select for motif instances in regions they are known to be functional: conserved transcription factor (TF) motifs enrich for promoters; miRNA motifs for 3′UTRs, and specifically the transcribed strand. Regions are normalized for their overall length, measured by the number of motif instances without conservation (0% confidence baseline). **b**, Increasing confidence levels select for transcription factor motif instances with experimental support for each factor tested. **c**, The high fraction of experimentally supported motif instances that are recovered at 60% confidence for transcription factors and 80% confidence for miRNAs illustrates the high sensitivity of the BLS approach. **d**, Comparison of chromatin immunoprecipitation (ChIP) and conservation in their ability to identify functional motif instances. Motif instances that are both ChIP-bound and conserved (purple) show the strongest functional enrichment in muscle genes for Mef2 and Twist (depletion for Snail), whereas motif instances derived by ChIP alone (light blue) show substantially reduced enrichment levels. Comparing the enrichment of all instances recovered by ChIP (blue) and all instances recovered by conservation (red) suggests that the two approaches perform comparably. Even the sites recovered by conservation alone outside bound regions (pink) show enrichment levels comparable to ChIP, suggesting that they are also functional.

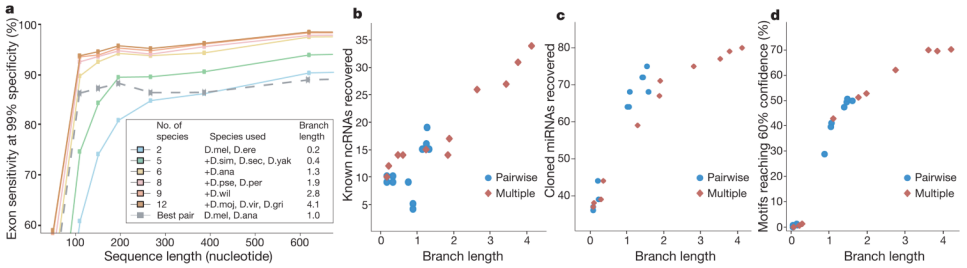

Figure 8. Scaling of discovery power with the number and distance of informant species

a, Discriminatory power of CSF protein-coding evolutionary metric for varying exon lengths and using different numbers of informant species. Sensitivity is shown for known exons at a fixed false-positive rate based on random non-coding regions. Mean length is shown for each exon length quantile. Multi-species comparisons increase discovery power, especially among short exons. **b**, Recovery of known ncRNAs (among the top 100 predictions) for pairwise (blue) and multi-species (red) comparisons. **c**, Recovery of cloned miRNAs (among the top 100 predictions). **d**, Recovery of transcription factor and miRNA motifs with instances at 60% confidence.

Table 1

Assessment of FlyBase euchromatic protein-coding gene annotations

Regions evaluated	Total	Confirm	Abstain	Reject*
Named genes	4,711	4,566 (96.9%)	105 (2.2%)	40 (0.8%)
CGid-only genes	9,022	7,879 (87.3%)	729 (8.1%)	414 (4.6%)
Non-coding regions†	15,564	3 (0.0%)	131 (0.8%)	15,430 (99.1%)

*A minority of rejected genes are false rejections; see Fig. 3b and text for details.

†Regions ≥300 nucleotides in length randomly chosen from the non-coding part of the genome (see Supplementary Methods 2a).

Table 2

Pre-transcriptional motifs

Name	Motif consensus	MEC	MCS	Region*	Known transcription factor†	Multiplicity score‡	ImaGO enrichment§	ImaGO score§
ME1	GTCACGTD	0.448	45.41	PIG	—	—	—	—
ME2	AWNTGGGTCA	0.393	26.97	PIG	Hr46	—	Oesophagus (13–16)	4.52
ME3	BCATAAATYA	0.369	36.02	PCEIG	Caudal	—	Ubiquitous (13–16)	−6.22
ME4	HAATTAYGCRH	0.365	32.71	PCE5IG	Engrailed	—	—	—
ME5	STATAWAWR	0.358	24.31	C	TATA	—	Ventral nerve cord (13–16)	−5.1
ME6	VATTWGCAT	0.356	44.06	PE5IG	—	3.73	Ubiquitous (11–12)	−7.15
ME7	BYAATTARH	0.338	15.45	PCE5IG	Engrailed	7.08	Ubiquitous (11–12)	−10.26
ME8	HRTCAATCA	0.338	42.32	PIG	—	—	Dorsal pharyngeal muscle PR (11–12)	−4.15
ME9	TGACANNNNNNTGACA	0.336	9	G	—	—	—	—
ME10	RCGTGNNNNGCAT	0.329	15.94	PIG	—	—	—	—
ME11	MATTAAWNATGCR	0.324	12.43	PIG	acj6	—	Tracheal PR (11–12)	4.11
ME12	TTAATGATG	0.32	20.31	PG	—	—	—	—
ME13	WTGACANBT	0.318	63.45	PE5IG	—	4.14	Ubiquitous (13–16)	−3.97
ME14	YGACMTTGA	0.313	27.06	PIG	—	—	Midgut (13–16)	4.32
ME15	AATTRNNNNCAATT	0.309	21.17	PG	—	—	—	—
ME16	TGACGTCAT	0.304	12.24	PC5IG	CrebA	—	—	—
ME17	MAATTNAATT	0.304	51.57	PE5IG	—	—	Ubiquitous (11–12)	−6.66
ME18	MRYTTCCGYY	0.304	39.04	PEIG	Dorsal	—	Ubiquitous (11–12)	−4.4
ME19	MATTRRCACNY	0.303	25.24	PIG	—	—	—	—
ME20	YTAATGAVS	0.298	44.5	PEIG	—	—	Foregut PR (11–12)	4.19
ME21	TAATTRANNTTNATG	0.294	8.67	G	—	—	—	—
ME22	WAATGCGCNT	0.291	18.17	G	—	—	—	—
ME23	MATTWRTCA	0.288	46.25	PEIG	—	—	Dorsal epidermis PR (11–12)	4.4
ME24	YAATTWNRYGC	0.287	30.91	PG	—	4.27	Ubiquitous (11–12)	−4.79
ME25	TTAYGTAA	0.283	13.06	5	Giant	—	Midgut (13–16)	5.32
ME26	YGCGTHAATTR	0.283	13.61	PEG	—	—	—	—
ME27	AATTRYGWCA	0.28	22.85	PEIG	—	—	Pericardial cell (13–16)	4.1
ME28	GCGCATGH	0.28	30.17	PCEG	—	—	Ventral nerve cord PR (11–12)	5.75
ME29	WAATCARCGC	0.275	13.82	G	—	—	—	—
ME30	AATTAANNNNNCATNA	0.271	16.44	G	Antennapedia	—	—	—
ME31	GCGTSAAA	0.271	29.95	PG	—	—	—	—
ME32	YGCGYRTCAWT	0.269	12.87	G	—	—	—	—
ME33	GCGTTGAYA	0.269	15.1	PG	—	—	—	—
ME34	AAATKKCATTA	0.266	14.04	PG	—	—	—	—
ME35	RACASCTGY	0.266	28.38	PCEG	Scute	—	Ventral sensory complex SA (11–12)	4.08
ME36	TGTCAATTG	0.265	12.65	PG	—	—	Tracheal system (13–16)	4.56
ME37	WAATKNNNNNCRCGY	0.261	23.34	PEG	—	—	—	—
ME38	CASGTAR	0.261	9.24	PEG	Single-minded	4.58	Ventral epidermis PR (11–12)	7.41
ME39	WCACGTGC	0.26	10.54	PCE5IG	Enhancer of split	—	—	—
ME40	CATTANNNWAATT	0.259	19.02	G	—	—	—	—

The top 40 of 145 are shown. MEC, motif excess conservation; MCS, motif conservation score. See Supplementary Table 5c for the full table.

*Region where the motif was found: P, promoter, C, core promoter; E, enhancers; 5, 5′ UTR; I, intron; G, intergenic genome.

[†] The known transcription factor motif matching the consensus sequence.

[‡] A multiplicity score is reported for motifs with many repeated occurrences.

[§] Tissue where motif is most strongly enriched or depleted, and corresponding score (positive, enrichment; negative, depletion). PR, primordium; SA, specific anlage.