Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes

Michael F. Lin,[1] Joseph W. Carlson,[2] Madeline A. Crosby,[3] Beverley B. Matthews,[3] Charles Yu,[2] Soo Park,[2] Kenneth H. Wan,[2] Andrew J. Schroeder,[3] L. Sian Gramates,[3] Susan E. St. Pierre,[3] Margaret Roark,[3] Kenneth L. Wiley Jr.,[4] Rob J. Kulathinal,[3] Peili Zhang,[3] Kyl V. Myrick,[4] Jerry V. Antone,[4] Susan E. Celniker,[2] William M. Gelbart,[3,4] and Manolis Kellis[1,5,6]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA; [2]Berkeley Drosophila Genome Project, Department of Genome Biology, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [3]FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA; [4]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA; [5]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA

The availability of sequenced genomes from 12 *Drosophila* species has enabled the use of comparative genomics for the systematic discovery of functional elements conserved within this genus. We have developed quantitative metrics for the evolutionary signatures specific to protein-coding regions and applied them genome-wide, resulting in 1193 candidate new protein-coding exons in the *D. melanogaster* genome. We have reviewed these predictions by manual curation and validated a subset by directed cDNA screening and sequencing, revealing both new genes and new alternative splice forms of known genes. We also used these evolutionary signatures to evaluate existing gene annotations, resulting in the validation of 87% of genes lacking descriptive names and identifying 414 poorly conserved genes that are likely to be spurious predictions, noncoding, or species-specific genes. Furthermore, our methods suggest a variety of refinements to hundreds of existing gene models, such as modifications to translation start codons and exon splice boundaries. Finally, we performed directed genome-wide searches for unusual protein-coding structures, discovering 149 possible examples of stop codon readthrough, 125 new candidate ORFs of polycistronic mRNAs, and several candidate translational frameshifts. These results affect >10% of annotated fly genes and demonstrate the power of comparative genomics to enhance our understanding of genome organization, even in a model organism as intensively studied as *Drosophila melanogaster*.

[Supplemental material is available online at www.genome.org. Additional supplemental materials are available online at http://compbio.mit.edu/fly/genes/. Full-length cDNA sequence data from this study have been submitted to GenBank under accession nos. BT029554–BT029635, BT029637–BT029727, BT029940–BT029957, BT030133–BT030144, BT030416–BT030421, and BT030448–BT030452. RT-PCR amplicon and primer sequence data have been submitted to GenBank under accession nos. ES439769–ES439782.]

The compilation of a complete and accurate catalog of all protein-coding genes is a critical step in fully understanding the functional elements in any genome. In *Drosophila melanogaster*, a century of classical genetics, large-scale EST and cDNA sequencing (Rubin et al. 2000; Stapleton et al. 2002b; http://www.fruitfly.org/EST), and manual curation (Adams et al. 2000; Misra et al. 2002) have led to a gene catalog of very high quality, containing 13,733 euchromatic protein-coding genes (as of FlyBase annotation Release 4.3, the benchmark release for the initial comparative analysis of the 12 sequenced species; see Methods). While all FlyBase genes are assigned a unique numerical identifier (CGid), their level of supporting evidence varies widely. We distinguish the following classes: 4711 genes have a phenotype or molecular function reported in the literature and have been assigned a descriptive name ("named genes"); of these, 893 have at least 50 literature citations ("well-studied genes"). The remaining 9022 genes lack a descriptive name ("CGid-only genes"); of these, 4373 have been assigned a putative molecular function on the basis of homology with known protein domains or genes in other species ("GO-annotated genes"), while the remaining 4649 gene annotations are essentially uncharacterized ("uncharacterized genes"). Most of the CGid-only genes are supported by cDNA sequence data or protein sequence similarity, but a small number are based primarily on de novo predictions.

It is unclear how close to completion the current gene set may be, or what fraction of the current annotations may be inaccurate. On one hand, numerous genes and alternative splice forms may be still missing from the current annotation, and indeed a pilot study suggests an additional 700 genes may lie amidst 10,000 existing de novo and microarray-based predictions (Yandell et al. 2005). On the other hand, existing gene models might be incomplete or contain inaccuracies, and some, especially those based solely on de novo predictions, may be completely spurious. Even some genes supported by cDNA and EST

[6]Corresponding author.
E-mail manoli@mit.edu; fax (617) 262-6121.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6679507. Freely available online through the *Genome Research* Open Access option.

evidence or mutation phenotypes could in fact represent RNA-coding genes without any protein-coding function.

Comparative genomic analysis is a powerful approach to the discovery of protein-coding genes. Comparative data have been used to significantly revise the established annotations of the yeast *Saccharomyces cerevisiae* genome (Cliften et al. 2003; Kellis et al. 2003), but the greater complexity of gene structures and other genomic features in large eukaryotic genomes presents many additional challenges. Initial efforts in vertebrates (Thomas et al. 2003) as well as flies (Bergman et al. 2002; Richards et al. 2005) suggest that comparative genomics can similarly lead to substantial improvements in the gene annotations of these species, and the incorporation of comparative data into de novo gene prediction systems has led to great improvements in their accuracy (Brent 2005). However, current de novo gene predictors still cannot be solely relied upon for complete annotation of complex eukaryotic genomes. Moreover, they are of limited use for revisiting existing annotations, since disagreements between predicted gene structures and gene annotations can be due to errors in the predictions at least as often as errors in the annotations. Thus, new methods are necessary in order to discover new genes and exons with high predictive value and to revisit existing annotations using comparative data in complex eukaryotes.

In this study, we use whole-genome alignments of 12 *Drosophila* genomes to systematically review the protein-coding gene annotations of *D. melanogaster*. By studying the conservation properties of known genes, we identify recurrent patterns of evolutionary change that are hallmarks of purifying selection operating upon protein-coding sequences. We use these evolutionary signatures to examine the entire genome and identify conserved protein-coding regions with high accuracy. These signatures confirm the protein-coding function of the vast majority of hypothetical genes and identify more than a thousand new exons. In contrast, these signatures strongly reject several hundred genes, most of which are likely to be spurious predictions or noncoding genes. We also used these signatures to refine the annotation and boundaries of existing genes, including translation initiation sites, splice sites, and functional reading frame of translation. Finally, our methods identify candidates for a variety of exceptional gene structures such as translational readthrough, dicistronic genes, and conserved reading frameshifts in the middle of protein-coding exons. We evaluated many of these proposed changes through manual curation and directed sequencing efforts. Overall, we used comparative data to propose revisions for >10% of *D. melanogaster* protein-coding gene models. While many extensions and future directions remain, this work is a substantial step toward achieving the best possible gene annotations for *D. melanogaster*. It also serves as a model for similar efforts to improve the annotation of other important target genomes, including the human.

Evolutionary signatures for protein-coding gene identification

Protein-coding DNA sequences evolve under distinctive evolutionary constraints since selective forces at the nucleotide level reflect constraints operating on the encoded protein. Thus, mutations to the DNA that preserve properties of the amino acid translation (e.g., synonymous substitutions) tend to be tolerated, while mutations that disrupt the translation (e.g., frame-shifting insertions or deletions or nonsense mutations) tend to be excluded by natural selection. In DNA sequence alignments of closely related species, these constraints manifest themselves as "evolutionary signatures", recurrent patterns of evolutionary change that we can use to uniquely identify protein-coding sequences (Fig. 1).

We applied two independent quantitative metrics that use evidence from multiple informant sequences to distinguish regions under protein-coding selection. The first metric observes reading frame conservation (RFC) and quantifies the strong tendency of insertions or deletions (indels) within coding regions to preserve the reading frame of translation. We have previously applied RFC in yeast species (Kellis et al. 2003). The second metric observes codon substitution frequencies (CSF) and identifies the distinctive biases in the frequency of codon substitutions in protein-coding regions, constrained by the selective preference for synonymous substitutions and amino acid substitutions preserving biochemical properties (Fig. 1). The CSF metric is similar in theme to the well-known K_a/K_s ratio and d_N/d_S rate (Yang and Bielawski 2000; Nekrutenko et al. 2002), but it is more suitable for genome-wide gene identification strategies with many informant genomes.

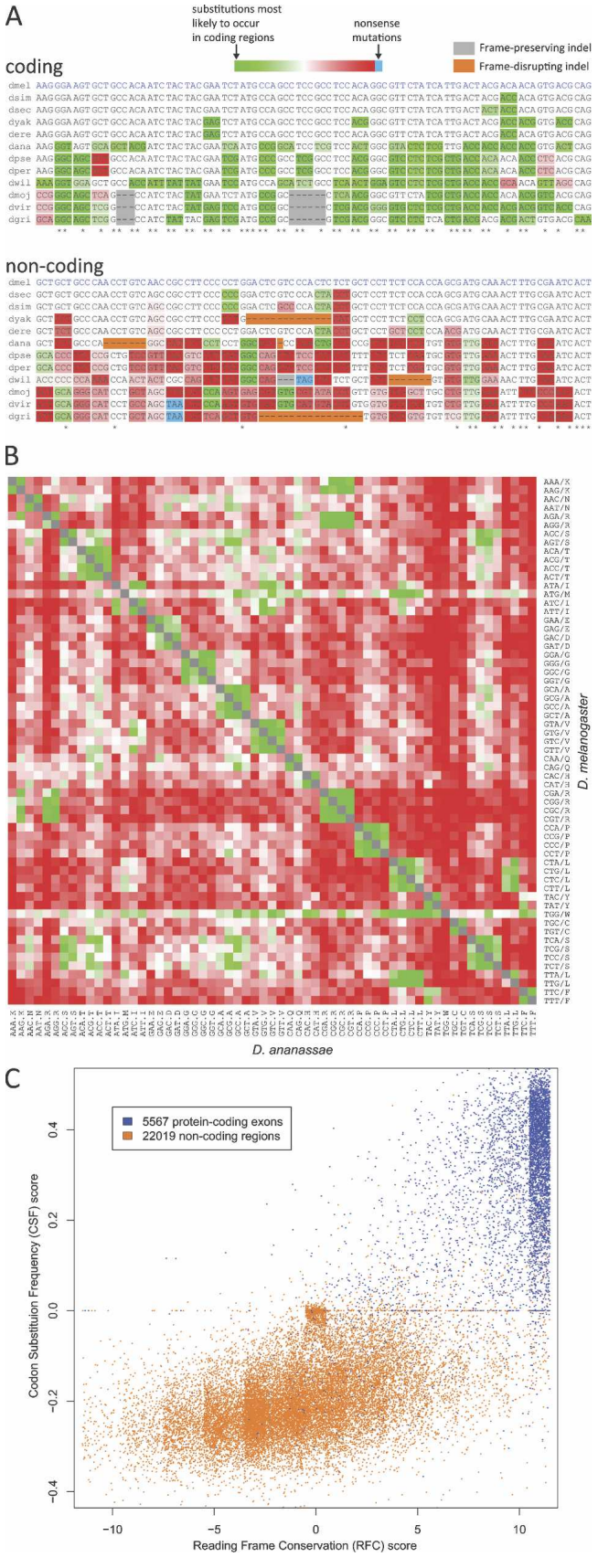
In contrast to methodologies that focus primarily on high sequence conservation to identify candidate genes, the RFC and CSF metrics focus on distinctive patterns of divergence in protein-coding genes, specific to their unique selective pressures. Therefore, functional RNA-level or DNA-level elements (such as RNA genes and structures, developmental enhancers, or other regulatory regions), which often exhibit high nucleotide conservation (Fig. 2), are very unlikely to show high RFC or CSF scores, enabling these metrics to distinguish coding and noncoding regions with higher accuracy. For example, when used to discriminate between exons of well-studied genes and random noncoding regions with the same length distribution, the CSF metric alone accepts 94% of coding exons while rejecting >99% of the control regions (Supplemental Table 1). This discriminatory power allowed us to systematically review the *D. melanogaster* genome annotation for protein-coding genes. We present detailed benchmarks of these and several other metrics elsewhere (M.F. Lin, A. Deoras, M. Rasmussen, and M. Kellis, in prep.).

Results

Benchmarking the RFC and CSF evolutionary signatures

Our first goal was to evaluate how well our approach worked on test data sets of well-annotated genes. For this purpose, we used the classes of "named" and "well-studied" genes defined earlier. We scored every gene model covered by whole-genome sequence alignments according to the RFC and CSF metrics. By studying the score distributions for known genes and noncoding control regions, we chose RFC and CSF cutoffs above which a given gene annotation is nearly certain to represent protein-coding sequence, and used these as a test to determine whether the comparative evidence confirms that a candidate gene is indeed protein-coding (although this test does not verify that the annotated gene structure is correct in every detail).

We first scored the 893 well-studied genes. Our test accepts 882 (99%) of these gene models. Only 11 of these genes did not pass our thresholds. Two of these (*y* and *bw*) are well-conserved genes that failed due to previously known strain-specific disrupting mutations in the sequenced strain of *D. melanogaster*. The remainder may represent fast-evolving genes or genes recently evolved from previously noncoding regions. We also applied the same test to the remaining 3818 named genes with <50 citations



and found that it accepts 97% (3684). Overall, the comparative evidence confirms that 4566 of 4711 "named" genes (97%) show the evolutionary signatures of protein-coding genes. We also evaluated 15,564 noncoding regions ≥300 nt in length, randomly chosen throughout the genome, and found that virtually none passed the same thresholds (Table 1; Supplemental Methods). Together, these results illustrate the high sensitivity and specificity of our approach.

Evolutionary confirmation of uncharacterized genes

We then turned our attention to the 9022 CGid-only genes in the Release 4.3 annotation set, which lack a descriptive gene name (including 4373 GO-annotated genes and 4649 uncharacterized genes). The evidence for these gene models varies widely and may include de novo gene model prediction, long open reading frames (ORFs), cDNA sequences, mRNA expression evidence, or homology with genes in other species. Since our evolutionary signatures are specific to protein-coding function, they can provide a powerful additional line of evidence indicating that these genes encode proteins, based on their alignments across *Drosophila* genomes.

Our test accepts 7879 of the 9022 CGid-only genes (87%), confirming that the vast majority of these annotations show the evolutionary signatures of protein-coding genes, and are therefore very likely to encode proteins. (Again, passing our test does not imply that all details of these gene structures are correctly annotated; we also note that it is possible that ancestral genes that have been very recently deactivated in *D. melanogaster*, have not yet acquired many disrupting mutations, and are still annotated as genes may pass our test.) The fraction of accepted CGid-only genes was only slightly higher for the "GO-annotated" subset than for uncharacterized (89% vs. 86%). It is not surprising that the proportion of accepted models for CGid-only genes (87%) is lower than for the named genes (97%): Some uncharacterized genes may be erroneous or spurious annotations (we consider this possibility further below), while others are likely to be under less stringent selective pressure than most named genes, many of which are conserved across very large evolutionary distances (Bergman et al. 2002).

New genes and exons

We next used evolutionary signatures to identify conserved protein-coding sequences missing from the current annotation. This requires not only a way to evaluate the protein-coding potential of a given region, but also a method to discover new coding

Figure 1. Evolutionary signatures for protein-coding gene identification. (*A*) Within coding regions, triplet substitutions are biased toward conservative codon substitutions (Codon Substitution Frequencies, CSF). Additionally, indels in coding regions are strongly biased to be a multiple of three in length (reading frame conservation; RFC). (*B*) The color of each codon substitution between the *D. melanogaster* sequence and an informant sequence corresponds to a log-odds score of observing that substitution in a coding region versus a noncoding region. (*C*) Quantitative metrics of RFC and CSF distinguish coding and noncoding regions. Shown in blue are 5567 coding exons of well-studied genes and in orange are 22,019 regions chosen uniformly at random from the noncoding part of the genome, with the same length distribution as the exons. The CSF score is length-normalized and the discrete RFC score is dithered by adding random noise uniformly from (−0.5,0.5) for the purposes of visualization.

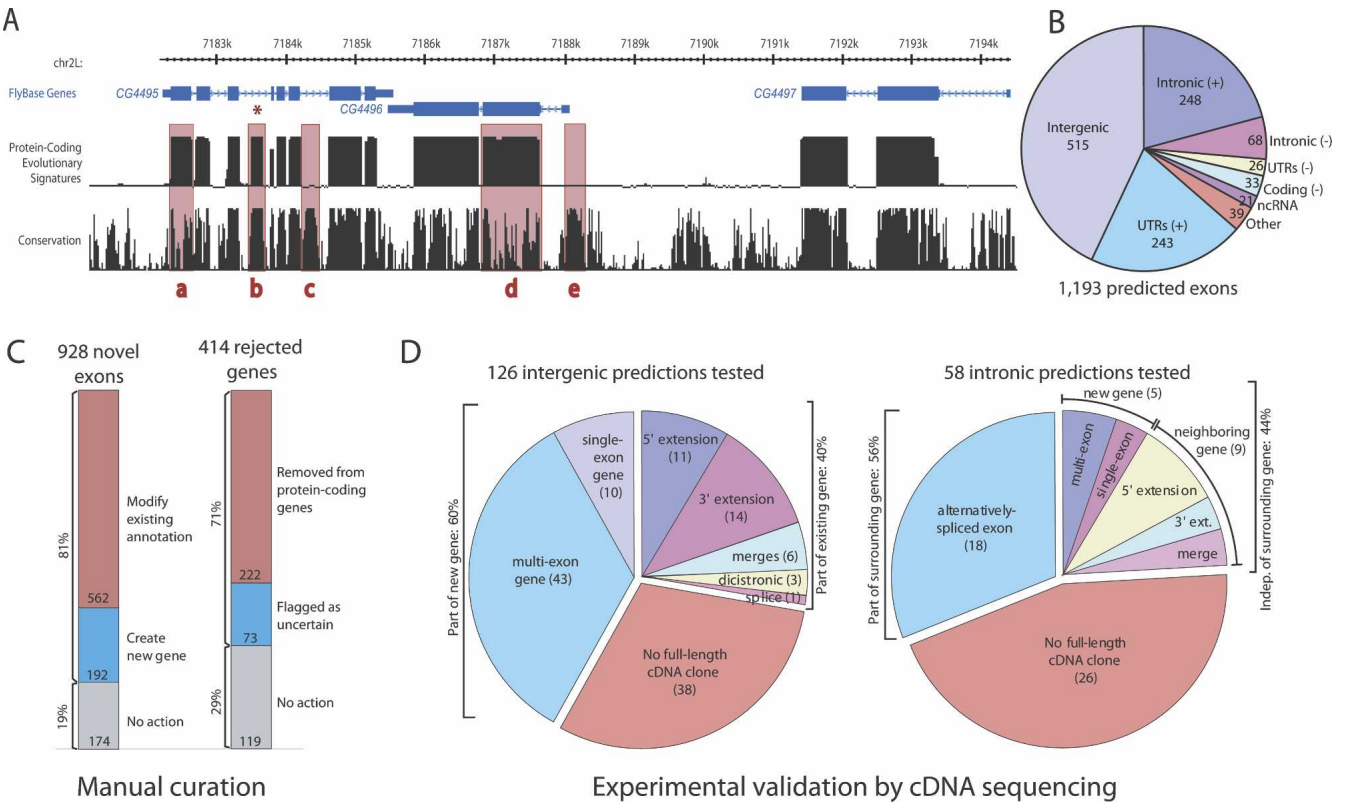


Figure 2. New protein-coding exons predicted by evolutionary signatures, examined by manual curation, and validated by cDNA sequencing. (*A*) The "Evolutionary Signatures" track shows the posterior probability of a protein-coding state in a probabilistic model integrating the RFC and CSF metrics. The "Conservation" track shows the analogous quantity from a model measuring nucleotide conservation only (Siepel et al. 2005). Note the high protein-coding scores of known exons despite lower nucleotide conservation (a,d), the low protein-coding scores of conserved noncoding regions (c,e), and the prediction of a novel exon within an intron of *CG4495* (b), subsequently validated (see Fig. 3). Rendered by the UCSC Genome Browser (Kent et al. 2002). (*B*) Distribution of 1193 new exon predictions throughout the genome. (*C*) Newly predicted exons were examined by manual curation, 81% leading to new and modified FlyBase gene annotations. Additionally, curation of genes rejected by evolutionary signatures led to the recognition of hundreds of spurious annotations. (*D*) A sample of predicted new exons was tested by cDNA sequencing with inverse PCR. Surprisingly, 44% of the validated predictions in "intronic" regions revealed a transcript independent of the surrounding gene, and 40% of the validated predictions in "intergenic" regions were part of existing genes. See Fig. 3 for examples.

intervals and to define their precise boundaries, in the absence of any existing annotation. To this end, we integrated our metrics of protein-coding evolutionary signatures into a probabilistic algorithm that determines an optimal segmentation of the genome into protein-coding and noncoding regions, based on syntenically anchored genome-wide sequence alignments of the 12 *Drosophila* species (see Methods). Our algorithm predicted 1193 new protein-coding exons not overlapping any coding exons in FlyBase Release 4.3. The large majority of these (68%) are in euchromatic intergenic and intronic regions: 515 (43%) are outside any annotated transcripts, and 316 (26%) are within an intron of an existing gene (248 transcribed from the same strand and 68 from the complementary strand). An additional 269 predicted exons (23%) overlap annotated untranslated regions of existing transcripts (243 from the same strand as the overlapping transcript and 26 from the complementary strand). The remaining predicted exons include 21 that overlap existing noncoding annotations, 33 that overlap protein-coding exons on the opposite strand, and 39 that overlap multiple Release 4.3 genes or are located in heterochromatin and cannot be easily categorized as intronic or intergenic. We manually examined most of these predictions, and also validated a subset through directed cDNA sequencing.

Manual curation incorporates most predicted exons into gene annotations

Of the 1193 predicted new exons, 928 were manually reviewed by FlyBase annotators and assessed relative to existing annotations, other gene predictions, cDNA/EST data, and protein sequence similarity evidence according to FlyBase Gene Model Annotation Guidelines (see Supplemental Methods). We excluded from manual review 265 predictions overlapping existing untranslated regions (UTRs), existing noncoding genes, or annotations independently created by FlyBase subsequent to Release 4.3 (our benchmark for this study).

Of the 928 assessed exons, 562 (61%) were incorporated into existing genes, leading to the revision of 438 gene models. The new exons most often led to the creation of alternative transcripts and, less frequently, to the modification of the intron/exon structure of an existing transcript isoform. Many of these changes (58%) were supported by additional evidence such as previously unincorporated BDGP cDNA sequences and/or sequence similarity to known proteins. Some revisions were complex, including 65 merges of two or more Release 4.3 gene models, 10 splits of Release 4.3 gene models, and four new dicistronic transcript models.

An additional 192 (21%) curated exons were incorporated in

Table 1. Categorization of existing gene annotations according to comparative evidence

	Total	Confirmed	Unclear	Rejected[a]
Named genes	4711	4566 (96.9%)	105 (2.2%)	40 (0.8%)
Well-studied genes	893	882 (98.8%)	8 (0.9%)	3 (0.3%)
Other named genes	3,818	3684 (96.5%)	97 (2.5%)	37 (1.0%)
CGid-only genes	9022	7879 (87.3%)	729 (8.1%)	414 (4.6%)
GO-annotated	4373	3897 (89.1%)	278 (6.4%)	198 (4.5%)
Uncharacterized	4649	3982 (85.7%)	451 (9.7%)	216 (4.6%)
All genes	13,733	12,445 (90.6%)	834 (6.1%)	454 (3.4%)
Noncoding regions	15,564	3 (0.0%)	131 (0.8%)	15,430 (99.1%)

Each annotated gene in FlyBase Release 4.3 is categorized as "confirmed" if it shows the evolutionary signatures of protein-coding genes, "unclear" if the gene is not alignable or the comparative evidence is otherwise ambiguous, and "rejected" if the gene is alignable to putatively orthologous sequence but appears unlikely to represent a genuine protein-coding gene. "Well-studied" genes are referenced by at least 50 publications in the FlyBase-indexed literature. "Named" genes have been assigned a descriptive symbol by investigators. All remaining genes are "CGid-only." "Noncoding regions" are ≥300 nt regions chosen randomly from the portion of the genome not annotated as protein-coding (see Supplemental Methods).
[a]A minority of rejected genes are falsely rejected; see text for explanation.

142 newly created gene models. Of these, 39% were supported by EST/cDNA and/or protein sequence similarity. Twenty-four of the new gene models (12%) lie within an intron of another gene on the same strand.

The remaining 174 curated exons (19%) were not incorporated into any gene models. Most of these are either small exon predictions, with a median length of 21 amino acids, or encode low-complexity sequence. Typically, these were unsupported by experimental data that would indicate inclusion in a gene model. These 174 exon predictions should be viewed as unresolved with regard to their validity, since future data may provide such experimental support.

Directed cDNA sequencing confirms predicted exons, reveals new genes and splice forms

In parallel to our manual curation efforts, we tested a subset of predictions by directed cDNA sequencing. To identify the most appropriate candidates for sequencing, we filtered the 1193 novel protein-coding exon predictions. We eliminated exon predictions of several types: those that map to certain known genes with incomplete annotations in FlyBase Release 4.3 (including heterochromatic genes and *Dscam*), to BDGP cDNA clones not yet represented in current FlyBase annotations, or to 5′ or 3′ UTRs. Additionally, we excluded any predicted exons that were deemed to be experimentally problematic because of small size or genomically repeated sequences (see Methods). Of the 434 remaining candidates for experimental validation, a sampling of 184, uniformly distributed throughout the genome, was selected. These included 126 within intergenic regions and 58 within introns of existing genes. We tested each of these 184 predictions by attempting to isolate and sequence a full-length cDNA transcript clone using self-ligation of inverse PCR products (Hoskins et al. 2005; Wan et al. 2006).

Of the 126 tested predictions within intergenic regions, we obtained a full-length cDNA for 88 exons (70%). The resulting cDNAs provide evidence for 50 new genes, including 10 single-exon genes and 40 multi-exon genes (which incorporate 43 predicted exons, and additional flanking exons that were not predicted by our algorithm). In addition, these cDNAs provided evidence for the modification of 39 existing Release 4.3 annotations: 11 new 5′ extensions or splice variants, 13 new 3′ extensions or splice variants (14 exons), two dicistronic transcripts (three exons), six transcripts merging multiple Release 4.3 gene models, and one internal splice variant.

Of the 58 tested predictions within introns of existing annotations, we obtained a full-length cDNA for 32 (55%). Only 18 of these represent new internal splice variants of the surrounding gene while the remaining 14 appeared independent of the surrounding gene. These 14 include eight alternative splice forms of previously annotated genes (five 5′ exons and two 3′ exons), two new single-exon genes, two new multi-exon genes, and two gene merges. Most surprising were data supporting an apparent example of overlapping coding sequence on opposite strands (Fig. 3D).

Overall, the cDNA data validated 120 of the 184 targeted predictions (65%). The recovered cDNA sequences also indirectly validated 42 predicted new exons that were not purposely targeted, as they were contained within the transcripts recovered from the 120 targeted predictions, leading to a total of 162 cDNA-validated predictions. The recovered cDNAs also captured additional translated and untranslated exons that were not predicted by our algorithm (see examples in Fig. 3). Finally, we note that the remaining 64 targeted predictions for which we did not obtain a high-quality, full-length cDNA sequence are not necessarily false predictions, since we only screened libraries derived from certain tissues and developmental stages (Hoskins et al. 2005).

Using TBLASTX, we searched other genomes for homologs of the new genes we recovered through cDNA sequencing. We found that many appear to be specific to the *Drosophila* or insect lineages (Supplemental Table 2). For example, 37% have a significant hit in the mosquito (*Anopheles gambiae*) or honeybee (*Apis mellifera*) genome assemblies, compared to 50% of randomly selected genes of comparable length; similarly, only 12% have significant hits to worm, yeast, or vertebrates, compared to 32% of random genes. Because gene annotation often relies on homology with known genes in other species, this might explain in part why these genes have not previously been identified.

An alternative strategy identifies relatively few additional exons

The completeness of our exon predictions is constrained by the coverage and quality of whole-genome alignments, the discriminatory power of our evolutionary metrics applied to the 12 genomes, and limitations of the probabilistic algorithm we used to integrate them. In fact, our exon prediction algorithm failed to identify 24% of exons in named genes (of which 37% were not well aligned; see Supplemental Table 3 for details). In order to

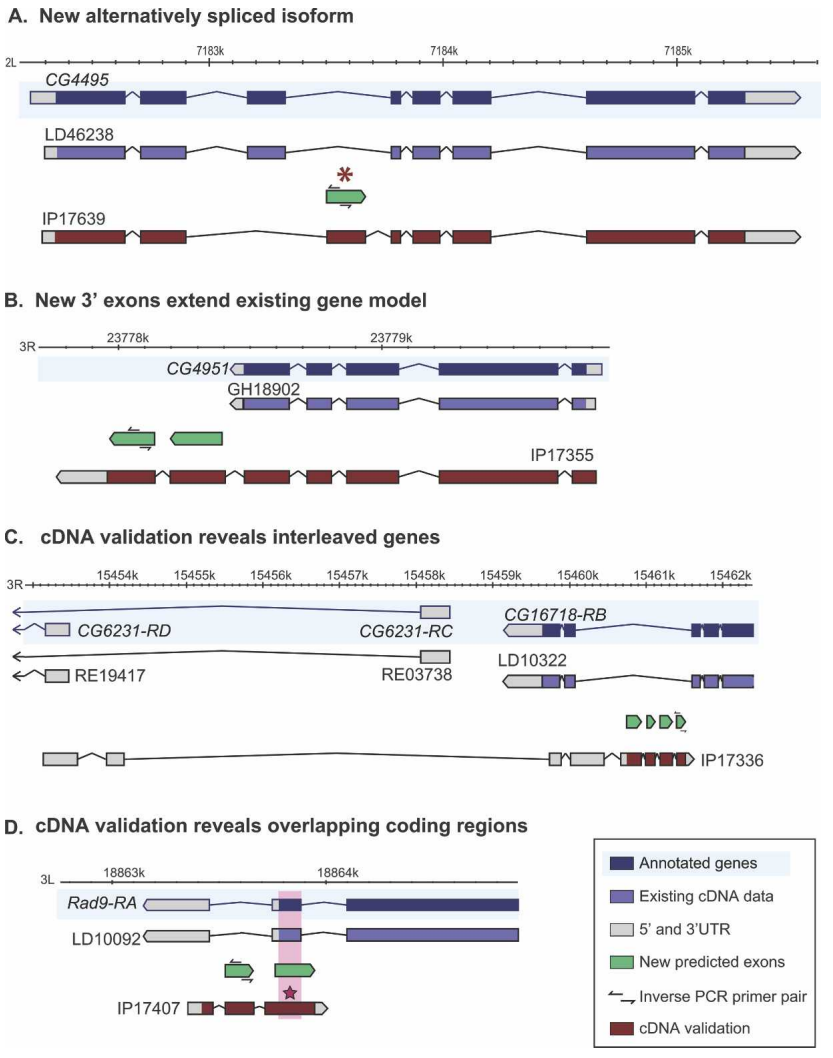


Figure 3. Full-length cDNA sequences recovered from exon predictions through inverse PCR. (*A*) Alternatively spliced transcripts—Exon Shuffling. The clone, IP17639, validates prediction congochr2L7183503 and provides evidence for an alternative transcript of the gene *CG4495*. Analysis of the embryonic microarray data (Manak et al. 2006) shows this exon is not used in embryogenesis, suggesting stage-specific splicing. Interestingly, the two alternative exons encode 20 identical amino acids at the N-terminal side of the exon. (*B*) 3′ CDS extension. The clone, IP17355, validates two predicted exons, congochr3R23777966 and congochr3R23778197, and provides evidence for an alternative transcript encoding an additional 126 aa at the C-terminal end of the gene, *CG4951*. In addition, the clone contains 185 bp of 3′ UTR. (*C*) New spliced interleaved gene. The clone, IP17336, validates four predicted exons, congochr3R15461397, congochr3R15461180, congochr3R15461031, and congochr3R15460742, and provides evidence for four additional exons. (*D*) Novel spliced overlapping gene. The clone, IP17407, validates prediction congochr3L18835687 and extends the CDS by 22 aa at the N terminus and 79 aa at the C terminus. The third coding exon overlaps the coding sequence of the gene on the opposite strand, *Rad9*, such that 45 aa on each strand are encoded in the region of overlap.

evaluate how much additional new protein-coding sequence may have escaped our notice, we undertook an alternative strategy that uses predictions from a variety of gene identification systems, representing several basic algorithmic approaches, and including both de novo and evidence-based strategies. These included AUGUSTUS (Stanke and Waack 2003), CONTRAST (S. Gross, C. Do, M. Sirota, and S. Batzoglou, Stanford University, http://contra.stanford.edu/contrast/), GENSCAN (Burge and Karlin 1997), NCBI GNOMON, geneid (Parra et al. 2000), Genie (Reese et al. 2000), and SNAP (Korf 2004).

We selected 193 "consensus" exons that are predicted by at least five of these algorithms, do not overlap annotated exons, transposable elements, or our predictions, and are at least 100 nt in length. After manual curation, 98 (51%) were incorporated into a gene model: 15 were incorporated into gene models that included exons identified by our algorithm, 63 were incorporated into existing gene models, and 20 were annotated as new or reinstated gene models. To test the validity of this approach, eight of the affected gene models were selected for evaluation by RT-PCR. Seven of the eight newly annotated "consensus" exons were validated. In several cases, additional newly annotated exons based on evolutionary signatures were also validated. Overall, 852 new exons were annotated by manual curation using both analyses, of which 88% were predicted by our algorithm based on evolutionary signatures.

Conclusion: New exons and genes

In summary, we integrated our metrics of protein-coding evolutionary signatures into a probabilistic algorithm that predicted 1193 new exons. Of these, 948 were subjected to manual curation or targeted experimentation, and 787 (83%) were supported by sufficient data to incorporate them into new or revised gene models, resulting in 150 new gene models, 70 gene merges, 10 gene splits of existing annotations, and four pairs of new dicistronic gene models. Some of the 161 predictions that were not supported following manual curation and targeted cDNA sequencing are likely to be validated in the future, e.g., as distant 5′ exons of annotated genes (Manak et al. 2006), when additional data become available. The 245 remaining predictions that were not assessed by either manual curation or experiments, most of which overlap annotated UTRs or noncoding gene models, await analysis.

Although the subsets of the predicted exons that we subjected to curation and sequencing were not selected entirely at random, neither were they selected in a way that would strongly bias them toward the highest-quality predictions. We conclude that our approach was able to identify new exons with very high predictive value, even when all existing gene annotations were excluded. Moreover, the results of an alternative strategy based on a variety of de novo and evidence-based predictions suggest that relatively few protein-coding exons remain unidentified in the euchromatin—at least that can be found at a reasonable false discovery rate using existing computational methods.

Many poorly conserved gene annotations are dubious

We next asked whether a subset of the CGid-only genes which failed to be confirmed as protein-coding is in fact spurious. Our previous analysis confirmed 97% of "named" genes, but only 88% of CGid-only genes; we reasoned that the remaining 1119 (12%) may be fast-evolving, recently gained, improperly aligned, or simply spurious. Here, we revisit this set to identify potential spurious annotations that do not correspond to protein-coding genes.

While our previous analysis evaluated each candidate gene over its entire length, here, we searched for any evidence of protein-coding selection. We allowed for fast-evolving domains or partially incorrect annotations by evaluating overlapping windows of 30 amino acids for evidence of protein-coding evolution. We also allowed for lineage-specific genes by searching for evolutionary evidence in groups of species at three different phylogenetic distances from *D. melanogaster*. Moreover, we tested three different genome alignment sets, to allow for potential misalignments (see Methods). Finally, we note that, if a gene is recently gained and its orthologous region is simply absent in the informant genomes, our methods make no statement about its veracity. Instead, we only evaluated regions that do align to putatively orthologous sequences in other species.

We found that 414 CGid-only genes (4.6% of 9022) are rejected even by these very lenient criteria. By comparison, only three of 893 well-studied genes (0.3%) are rejected and only 40 of all 4711 named genes (0.8%). If all rejected well-studied genes are false rejections, we would expect <30 of the 414 rejected CGid-only genes to be false rejections (95% confidence, binomial distribution). Based on named genes, we would expect that <91 of the 414 rejections are false rejections, and that at least 323 of the 414 rejected genes (78%) are indeed spurious. On one hand, this may be an overestimate, as the named and well-studied genes may be biased toward deeply conserved functions with vertebrate orthologs (Bergman et al. 2002). On the other hand, this may be an underestimate if not all rejected named genes are false rejections; some could in fact be incorrect annotations. In particular, we note that a gene can be named on the basis of a mutant phenotype, which does not necessarily imply that it is protein-coding.

Several statistics suggest that most of the genes rejected by our test are likely to be spurious predictions. As a group, they closely resemble random noncoding regions (Supplemental Fig. 1). The majority consist of relatively short, single-exon ORFs, many of which are likely to occur by chance across the whole genome. Their median coding sequence length is 381 nt, considerably shorter than the median length of all genes (1179 nt), and 63% are single-exon.

We manually examined each of the 414 CGid-only genes that were rejected by our test and all evidence supporting them, and we concluded that 222 (54%) can be immediately deleted from the annotations or recategorized as nonprotein-coding genes. These include 55 genes previously annotated as supported by cDNA sequences, which in fact turned out to be due to genomically primed clones. An additional 73 of the rejected genes (18%) had unclear or conflicting evidence and have been flagged as being of uncertain quality in the annotation comments, although they were not immediately deleted. Finally, the remaining 119 (29%) are adequately supported by existing evidence and were kept unchanged in the current database. A subset of these is likely to be rapidly evolving genes, while others may prove to be RNA-coding genes with no protein function.

We also manually examined the 40 named genes that were rejected by our test, and found that six of these should also be deleted or changed to nonprotein-coding annotations. The remaining 34 contain several genes known to be rapidly evolving, including seven male accessory gland peptides or other male-specific genes.

Last, we found that transcript evidence for at least some of the rejected genes may be explained by nonprotein-coding function. In particular, there is strong evidence that the transcripts for *CG33311* and *CG31044* are in fact precursor RNAs of microRNA genes rather than protein-coding mRNAs (Stark et al. 2007). In both cases, newly discovered microRNA genes lie within these transcripts and cluster with neighboring miRNAs of the same family. More generally, we note that some forms of the evidence supporting CGid-only genes, such as transcript cDNA sequence or genomic sequence conservation, do not directly imply translation to protein and could result from noncoding genes.

We conclude that most of the genes rejected by our test in fact do not represent genuine protein-coding genes, and the existence of many of these annotations is due to genomically primed cDNAs, erroneous de novo gene predictions, and sometimes functional RNA genes. A minority is likely to represent fast-evolving or species-specific genes that are not under purifying selection over the evolutionary distances we examined. Overall, our tests based on evolutionary signatures confirmed 7879 of 9022 CGid-only genes (87%) as clearly under protein-coding selection and rejected 414 (4.6%), most of which are likely to be spurious annotations (Table 1). We abstained from making a decision based on comparative evidence for the remaining 729 CGid-only genes (8.1%), which either could not be aligned or were supported by evolutionary signatures weakly or only over a fraction of their length. These results can help guide directed experimentation to resolve the function of all genes and transcripts, and also help focus curation efforts on a relatively small number of problem cases.

Refining existing gene annotations

The deep comparative evidence available within alignments of the 12 *Drosophila* genomes enables more fine-grained analysis than the evaluation of complete genes. We also used our metrics of protein-coding evolutionary signatures to propose a variety of detailed adjustments to existing gene annotations, affecting translation initiation sites, splice boundaries, and reading frame of translation, and to reveal likely species- or strain-specific disruptive mutations.

Translation start sites

Systematic annotation of fly genes has typically designated the longest ORF of each transcript model as the inferred protein translation, starting at its earliest in-frame ATG. However, translation may actually start at a downstream ATG. While the current understanding of the sequence and structural signals that direct translation initiation is still incomplete, the evolutionary signatures of protein-coding selection can often clearly distinguish the preferential site of translation initiation. Our analysis revealed 413 transcripts of 359 genes for which the translation start sites appear to be downstream from the presently annotated AUG, and allowed us to propose corresponding refinements to the annotations. In each case, the previously annotated start AUG is

not well-conserved, while the newly proposed AUG is conserved, and the intervening sequence in other species shows an abundance of nonconservative codon substitutions and frequent in-frame stop codons and frame-shifting indels. In many cases, the contrast between the conservation of the regions immediately upstream and downstream of our proposed translation start sites is striking (Fig. 4A). While we cannot rule out that some of these cases could represent species- or lineage-specific N-terminal protein extensions, a majority of our proposed downstream translation start sites are also supported by an independent analysis of the information content in their sequence contexts, to the exclusion of the annotated upstream site (M. Weir and M. Rice, in prep.).

Reading frame of translation

In addition to locating protein-coding regions, the comparative information reveals the reading frame of translation under purifying selection, since the signature of codon substitution frequencies is specific to the reading frame. This has allowed us to distinguish between overlapping ORFs, and reveal the one under selection when multiple ORFs of comparable length are all open (Fig. 4B). Such overlapping ORFs are sometimes found in short single-exon genes, where the systematic annotation has typically selected the longest, while it may in fact be a shorter ORF that is translated. We found five cases (*CG15281*, *CG13244*, *CG7738*, *CG18358*, and *CG12656*) where a shorter ORF is clearly under selection, to the exclusion of the annotated ORF. While this is a small number of cases, we note that this change leads to a completely different protein translation.

Adjustments to existing exons

We searched for potentially erroneous splice sites by identifying gene models in which a splice junction appears to coincide with a shift of the reading frame under selection (Supplemental Fig. 3). We found such events in 210 transcripts of 174 genes. While alternative splicing can use exons in different reading frames, we can at least say in these cases that selection appears to strongly favor one translation of the exon over the alternatives. We conclude that the alternatives should, therefore, be considered suspect, at least in the absence of transcript sequence data clearly indicating their use.

We also identified many existing exons that appear to be incompletely annotated, as the evolutionary signatures of protein-coding selection extend beyond their present splice bound-

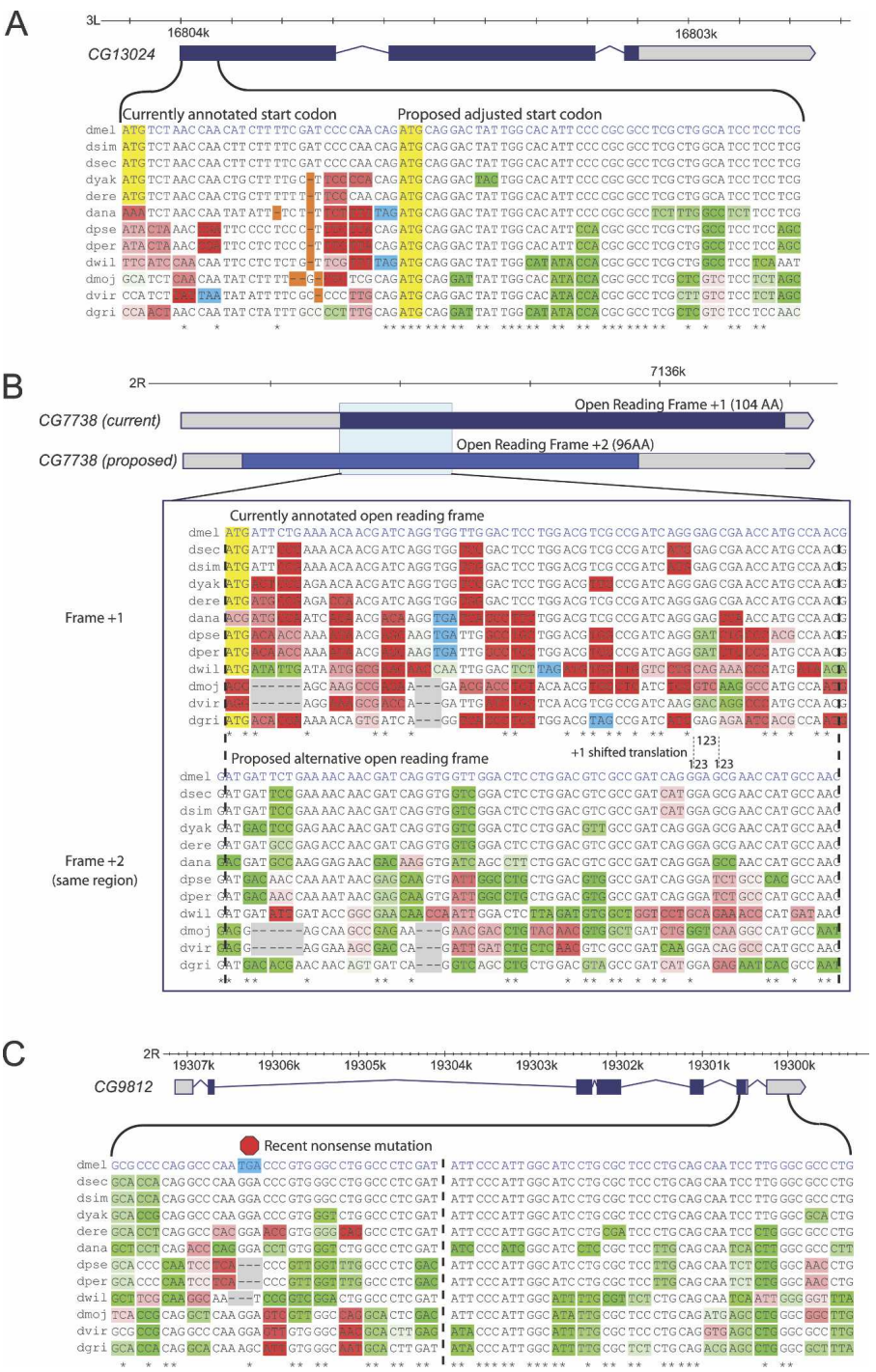


Figure 4. Examples of adjustments to existing annotations based on evolutionary signatures. (*A*) Translation start adjustment. The annotated coding sequence begins at the indicated ATG, but the informant species show frameshifts, nonsense mutations, and nonconservative substitutions in the immediately downstream region. Strikingly, however, coding signatures begin at a slightly downstream ATG. (*B*) Incorrect reading frame annotated. The transcript model contains two overlapping reading frames, the slightly longer of which is annotated as the coding sequence; but the evolutionary signatures clearly show that the other is the frame under selection. (*C*) Nonsense mutation in (the sequenced strain of) *D. melanogaster*.

aries, including 912 by at least 30 nt and 600 by at least 45 nt (see Supplemental materials). This may indicate either an alternative splice site or a simple mistaken annotation. When we considered the position of the likely corrected (or alternative) splice site, we found that the "extensions" of at least 30 nt are enriched for lengths divisible by three ($P < 2.2 \times 10^{-16}$, χ^2 test), suggesting that many may be alternatively spliced.

Recent nonsense and frameshift mutations

Finally, we used comparative information to identify several recent disrupting mutations in the sequenced strain of *D. melanogaster* (Celniker et al. 2002), which may have accumulated such characters during many years in laboratory culture. We identified two genes (*CG9812* and *CG33282*) in which an in-frame stop codon is aligned to a sense codon and followed by additional well-conserved protein-coding sequence in all aligned informant species. These appear to be recent nonsense mutations (Fig. 4C). An additional case, *CG14638*, may be a pseudogene.

We also identified locations in the *D. melanogaster* genome where protein-coding evolutionary selection abruptly shifts from one reading frame to another. In five cases, these coincide with a short frame-shifting indel, specific to the sequence of *D. melanogaster*, and absent from all of the other genomes. One of these (within *sdk*) was due to a previously known erroneous genomic sequence on chromosome arm 3L in *D. melanogaster*, while another (within *CG33294*, currently known as *CR33294*) may be a pseudogene. The remaining three cases (within *Ugt86Dd*, *Dscam*, and *CG34143*) are apparently recent frameshift mutations.

Identifying unusual protein-coding structures

The power of evolutionary signatures to distinguish regions under protein-coding selection has allowed us to recognize a variety of unusual phenomena that have not been amenable to systematic discovery, including stop codon readthrough, polycistronic transcripts, and translational frameshifts. We present here the results of this computational analysis, reflecting the best inference from the comparative data available to us. However, the underlying biological mechanisms remain unclear in most cases, and follow-up investigation will be required to explain these observed phenomena.

Stop codon readthrough

Just as evolutionary signatures can often distinguish the preferential site of translation initiation, they can also accurately identify the true site of translation termination. For the vast majority of genes, the comparative data show that protein-coding selection degrades exactly at the stop codon or shortly upstream. For 149 genes, however, evolutionary signatures strongly suggest that translation continues well past a deeply conserved, in-frame stop codon (Fig. 5A), indicating that these "extensions" of the corresponding proteins, which range in length from 15 to hundreds of amino acids, are under selection for their protein-coding function.

Translational readthrough of stop codons can occur through several mechanisms, among which our approach does not distinguish. However, it does not appear that many of these genes represent new selenoproteins, because many (37%) of the putatively readthrough stop codons are not UGA and we were unable to identify convincing examples of the related SECIS elements according to previously established criteria (Kryukov et al. 1999; Castellano et al. 2001). We found the set of putative readthrough genes to be enriched for nervous system expression patterns, according to in situ hybridization data (Tomancak et al. 2002; hypergeometric $P < 4.2 \times 10^{-5}$). For this reason, we speculate a possible role for A → I RNA editing by ADAR, which is most active in the nervous system (Bass 2002) and is known to mediate stop codon readthrough in a viral gene (Luo et al. 1990; Casey and Gerin 1995) and in a *D. melanogaster* neuropeptide receptor (Fig. 1 of Stapleton et al. 2002a). Still, other mechanisms may be responsible, and it is also possible that precisely positioned alternative splicing could lead to the observed signatures without direct readthrough. Overall, our results suggest that translational readthrough is not a rare phenomenon in *Drosophila* and provide candidate genes for further investigation.

Polycistronic messenger RNAs

Polycistronic messenger RNAs are single processed transcripts containing several nonoverlapping ORFs, each of which is individually translated (Andrews et al. 1996; Brogna and Ashburner 1997; Misra et al. 2002). We searched for complete (start-to-stop) ORFs that show clear signs of protein-coding selection and are fully contained within the untranslated region of an existing transcript model (Fig. 5B). This strategy rediscovered 85 of 115 annotated euchromatic dicistronic transcripts (73%) and predicts an additional 135 putative ORFs in 123 genes. We note that many of the ORFs of the previously annotated dicistronic transcripts are also found in single ORF mRNAs. This may also be the case for the genes we have identified. Our results provide a much richer set of candidate genes for further investigation, potentially doubling the number of genes with an annotated dicistronic transcript in the *D. melanogaster* genome.

"Programmed" translational frameshifts

Programmed translational frameshifts are common in viral genomes (Farabaugh 1996), and there is one known example in *D. melanogaster* (Ivanov et al. 1998). We found four locations in fly transcripts where protein-coding selection abruptly shifts from one reading frame to another, that are not readily explained as an incorrect transcript model or a species- or lineage-specific mutation. In these cases, the comparative evidence appears to indicate that a conserved translational frameshift occurs (Fig. 5C). One such case has a striking association with a highly conserved RNA structure (Supplemental Fig. 2), which we speculate might be involved in regulating its usage (Giedroc et al. 2000). We cannot, however, rule out mechanisms other than translational frameshifting, including alternative splicing, and further experimental study is required.

Discussion

A revised fly gene catalog

The availability of whole-genome alignments of the 12 *Drosophila* genomes allowed us to measure evolutionary signatures unique to protein-coding regions. In conjunction with manual curation and large-scale sequencing experimentation, these signatures enabled us to systematically revisit the fly genome annotation, with proposed changes affecting >10% of all genes. (1) We identified 1193 new exons with high predictive value, most of which were integrated into FlyBase gene annotations and many of which were validated by cDNA sequencing experiments, revealing many surprising new gene models and alternative

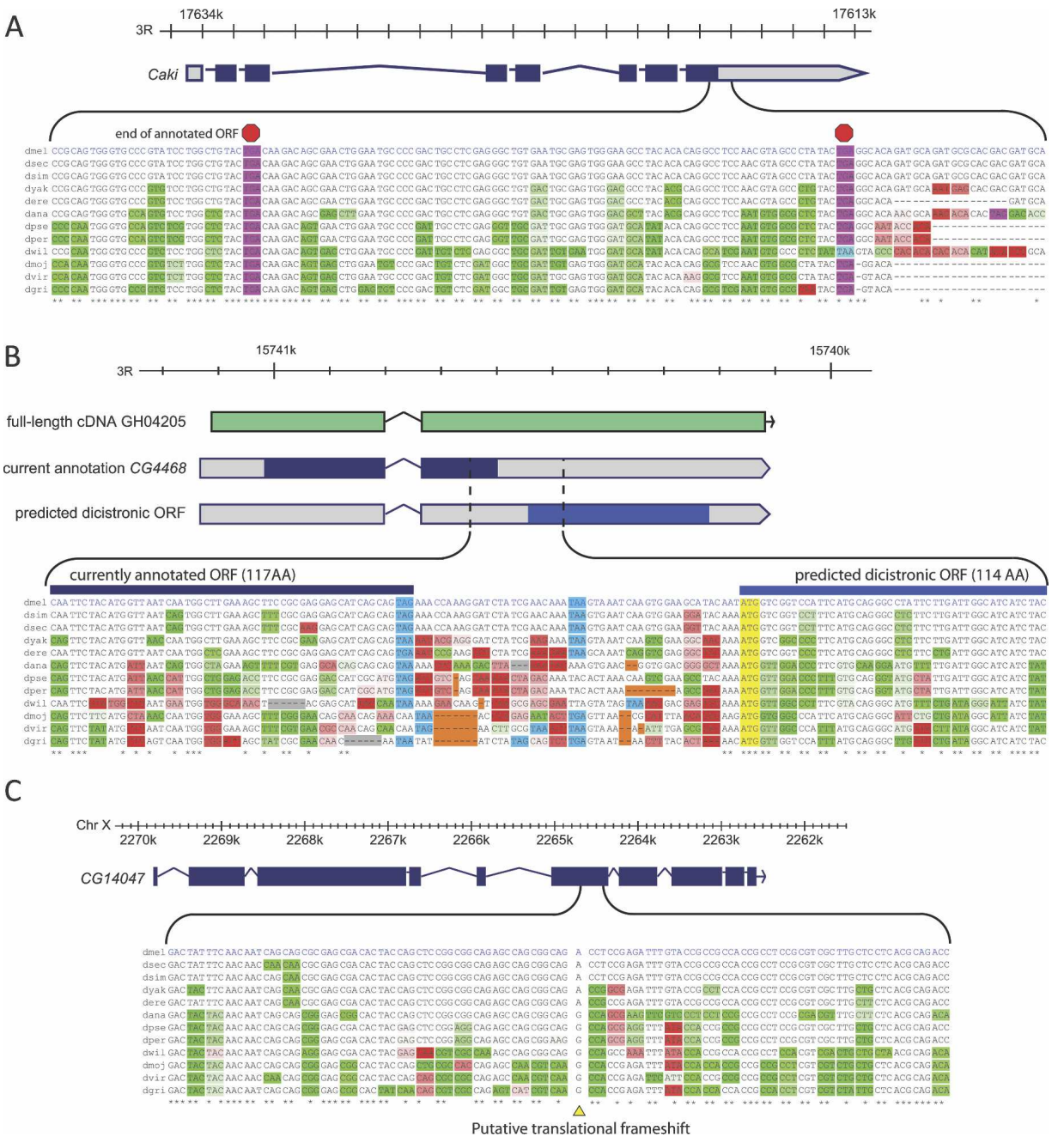


Figure 5. Unusual protein-coding structures identified by evolutionary signatures. (*A*) A well-conserved 30-aa ORF immediately following the stop codon in the gene *Caki* suggests translational readthrough. Note the perfect conservation of the putative readthrough stop codon, the "wobble" of the downstream stop codon, and the precipitous loss of conservation following the downstream stop codon, typical of a true translation stop. (*B*) A well-conserved ORF within the annotated 3′ UTR of *CG4468* suggests a dicistronic transcript structure. Note the region of poor conservation that extends precisely from the upstream stop codon to the downstream start codon, suggesting separate translation of the two ORFs. (*C*) An abrupt change in the reading frame upon which selection appears to act within an exon of *CG14047* is suggestive of a "programmed" translational frameshift (see also Supplemental Fig. 2).

splice forms. (2) In addition to discovering new genes, we used evolutionary signatures to revisit existing gene annotations. This led to confirmation that 87% of CGid-named annotations show evolutionary signatures of protein-coding genes and, conversely, to the identification of 3%–4% of CGid-only annotations that are likely to be spurious predictions or noncoding genes. (3) At a finer-grain level, evolutionary signatures allowed us to propose detailed refinements to hundreds of existing annotations, adjusting the translation start codon, correcting splice boundaries, resolving the functional reading frame in short single-exon transcripts, and identifying strain-specific disrupting mutations. (4) Lastly, the power of evolutionary signatures enabled us to recognize unusual gene structures, which challenge the current assumptions of gene annotation efforts: We found abundant evi-

dence of stop codon readthrough, polycistronic transcripts, and several candidates for conserved translational frameshifts.

Challenges for computational prediction of complete gene models

The comparative metrics we used in this study allowed us to distinguish individual protein-coding regions with high predictive value. To tie these exons into complete gene models, we relied on manual curation and large-scale cDNA sequencing experiments directed by our predictions. This allowed us to avoid simplifying assumptions about gene structures typically imposed by de novo gene structure predictors (Brent 2005).

Our results revealed important insights relevant to full gene model prediction. We obtained full-length cDNA clones for 162 of our predicted new exons, many of which fell into surprising gene models, reinforcing the difficulty of de novo gene model prediction. For example, when new exons were discovered within introns of existing genes on the same strand, the simplest expectation would be that they form alternatively spliced transcripts of the surrounding gene. In contrast to this expectation, however, only 56% were alternative transcripts, and the remaining 44% linked to other genes or formed independent transcription units. Such nested and interdigitated genes, as well as mutually exclusive exons within single genes, are refractory to most de novo gene structure predictors.

A further challenge to computational gene structure prediction is presented by exceptional biological phenomena, such as stop codon readthrough, polycistronic transcripts, and translational frameshifts. These are generally assumed to be rare and eukaryotic gene predictors are not built to recognize them. However, 115 dicistronic genes are currently annotated in FlyBase, and our results suggest that the true number may be substantially larger. Similarly, while only one functional translational frameshift has been described in *Drosophila* (Ivanov et al. 1998), our results revealed several new candidates. Most intriguing are the 149 genes we identified as potential targets of stop codon readthrough, which suggest that this phenomenon might be dramatically more common than currently understood (with only three known selenoproteins [Martin-Romero et al. 2001] and a few other cases [Xue and Cooley 1993; Bergstrom et al. 1995; Steneberg et al. 1998]). Although these phenomena may still be considered rare among ~14,000 genes, they represent some of the most intriguing examples of biological versatility, and a complete catalog of protein-coding genes cannot ignore them.

The next major advances in de novo gene prediction methods are likely to come from continued advances in our understanding of the sequence signals governing transcription, splicing, and translation regulation, as well as the advent of more flexible algorithmic frameworks that are well-suited to take advantage of such unconventional signals (Lafferty et al. 2001; Gross et al. 2006; Bernal et al. 2007). Still, the complex and noncanonical gene structures described above present challenges that appear difficult to overcome even for this next generation of eukaryotic gene structure predictors.

Applying the evolutionary signature approach to other target genomes

We believe that the work described in this report clearly demonstrates the power and practicality of our approach to improve the gene annotations of important genomes by complementing existing methodologies, including de novo gene structure prediction, large-scale cDNA sequencing, and manual curation. Our methods are directly applicable to other genome annotation projects that have this infrastructure in place, including the human (Harrow et al. 2006).

More generally, the preexisting, high-quality annotations for *D. melanogaster* allowed us to demonstrate the high sensitivity and specificity of the RFC and CSF tests based on evolutionary signatures. Since these signatures are universal consequences of natural selection and the genetic code, our results suggest that they can provide a strong foundation for the identification of protein-coding genes within any group of closely related species, even when cDNA library sequences are not immediately available or when no genomes with high-quality annotations exist in closely related taxa. Furthermore, it may also be possible to define specific evolutionary signatures—beyond mere sequence conservation—for other classes of functional elements, which suggests a general approach for the identification of functional elements in any genome. The derivation of reliable gene models for protein-coding genes remains a challenge, especially given the abundance of complex gene structures in metazoan genomes. It is also inherently difficult for comparative genomic methods to identify very fast-evolving, species-specific genes, which are centrally important to the study of evolution, speciation, and immunity. Thus, the complete genome annotation of any species will continue to be most effectively pursued through the concerted efforts of computational predictions, manual curation, and large-scale cDNA sequencing.

Methods

Genome alignments

We used several different sets of multiple sequence alignments of the 12 *Drosophila* genomes in this study. Two were based on a synteny map generated by Mercator (C. Dewey [University of Wisconsin, Madison] and L. Pachter [University of California at Berkeley]), with sequence alignments generated by MAVID (Bray and Pachter 2004) and Pecan (B. Paten and E. Birney, European Bioinformatics Institute, Cambridge, UK). Additionally, we used a MULTIZ (Blanchette et al. 2004) alignment of the 12 *Drosophila* genomes and three other insect genomes, excluding the non-*Drosophila* species. We used the Mercator/MAVID synteny-anchored alignments for predicting new exons and all three alignment sets for evaluating existing gene models (taking the highest-scoring version of the gene from the three alignments in order to have some robustness against alignment errors; see Supplemental Methods for details).

Reading frame conservation (RFC)

The RFC score was computed as we have previously described (Kellis et al. 2004). Briefly, given a region of the genome, a pairwise score between *D. melanogaster* in each informant is computed as the percentage of *D. melanogaster* nucleotides in the same reading frame in the informant (taking the largest such percentage out of the three possible reading frames). Each informant then votes +1, −1, or 0 based on an informant-specific cutoff on the pairwise RFC score: +1 if the score is above, −1 if the score is below, or 0 if there was no sequence aligned. These votes are then summed to obtain an overall score for the region.

Codon substitution frequencies (CSF)

CSF assigns a score to each pairwise codon substitution between *D. melanogaster* and an informant equal to the log-likelihood ra-

tio of observing that substitution in coding sequence versus noncoding sequence, conditioned on observing a substitution of the *D. melanogaster* codon. These log-likelihood ratio scores, shown in Figure 1B, were computed from codon distance matrices estimated by counting the frequencies of codon substitutions in alignments of annotated genes and noncoding regions for the appropriate pair of species, similar to the BLOSUM amino acid distance matrices estimated from protein alignments (Henikoff and Henikoff 1992). To obtain a score for a given genomic region, the scores of all codon substitutions in its alignment were summed; no score was assigned to gapped or perfectly conserved codons. With multiple informant sequences, the median of the scores of all codon substitutions in each codon column was used as the score of that column, and the score of each column was summed to score the region (see Supplemental Methods for complete details about CSF).

Thorough benchmarks of the RFC and CSF metrics, as well as various other discriminative metrics for protein-coding gene identification, with different alignments and different combinations of informant species, are presented elsewhere (M.F. Lin, A. Deoras, M. Rasmussen, and M. Kellis, in prep.).

"Confirming" genes

We obtained alignments for every transcript model in FlyBase annotation Release 4.3 by extracting them from the genome alignments (see Supplemental Methods). We scored each transcript by the CSF and RFC metrics and normalized the scores by length. Additionally, we scored thousands of disjoint intervals of at least 300 nt, selected uniformly at random from the noncoding part of the euchromatic genome. To define the test for "confirmation," we chose simple cutoffs on the metrics that exclude >99.9% of the control regions (see Table 1 and Supplemental Methods for specific cutoffs used).

"Rejecting" genes

Our test for identifying "rejected" genes was performed by computing the CSF score over every overlapping 30-aa window in every transcript model for each gene. Additionally, we computed these scores using all three genome alignment sets and using three different subsets of the informant species, representing all 12 *Drosophila* genomes, the subgenus *Sophophora*, and the *melanogaster* group. We took the highest scoring 30-amino-acid window in each gene, out of all its transcripts, all of the alignments, and all of the phylogenetic clades, as the score for that gene. We observed the distribution of this score to be bimodal, chose a cutoff to isolate the lower distribution, and found it to closely resemble our random controls (Supplemental Fig. 1).

Predicting new exons

We integrated our evolutionary metrics as features into a semi-Markov conditional random field (SMCRF), a probabilistic model similar to a generalized hidden Markov model but with more flexibility to directly incorporate discriminative metrics such as RFC and CSF (Lafferty et al. 2001; Sarawagi and Cohen 2005). The SMCRF uses the evolutionary metrics to predict only individual exons, not complete gene structures, and therefore may be considered more similar to interval segmentation algorithms that define the boundaries of high-scoring regions than to full gene predictors. The other features used by the SMCRF include sequence-based splice site discriminators (Yeo and Burge 2004), start/stop codon indicator functions, and a length distribution feature; however, it did not contain any explicit coding sequence composition features (e.g., high-order Markov models), nor did it use any information about transcript sequence evidence or homology with known proteins. The SMCRF had seven segment labels or "states": one for each codon position (reading frame) on each strand and one noncoding. The model was trained by maximum conditional likelihood using a training set of known genes, and the Viterbi algorithm was used to generate exon predictions for the whole genome in the Mercator/MAVID alignments (see Supplemental Methods for further details).

Selection of exon candidates for cDNA isolation

We used self-ligation of inverse PCR products (Hoskins et al. 2005; Wan et al. 2006) to screen four cDNA libraries to obtain clones that contained the predicted conserved exons using a modified primer design strategy. Primers were designed for optimal PCR conditions eliminating the requirement for 5′ bias in placement; 172 predicted exons failed the primer design step of our cDNA screening strategy because they were either too small or not unique in the genome. Of the 434 remaining candidates for validation (after exclusion of predictions with existing EST evidence and other filters; see main text), we selected 184 for validation by maximizing the genomic separation between tested predictions. After cloning of the PCR product four sequencing reads were produced: one from each cDNA end and one from each PCR primer. The composite sequence was used to evaluate whether the clone matched the targeted exon. Clones that matched were selected for complete sequencing.

RT-PCR

We extracted total RNA from the *D. melanogaster* sequenced strain at four time points (0–12-h embryos, 12–24-h embryos, first instar larvae, and adults), using the Micro-to-Midi Total RNA Purification System (Invitrogen). Processed mRNA was isolated using Oligotex mini mRNA Kit (Qiagen) and RT-PCR was performed using the OneStep RT-PCR kit (Qiagen) or Invitrogen Superscript II as reverse transcriptase. Gene-specific primers were designed using primer3 (http://primer3.sourceforge.net/). Primers were 20–24 bp in length and were designed to cross intron/exon boundaries. PCR products were directly sequenced and aligned to the genome using est2genome (http://emboss.sourceforge.net). Primer and amplicon sequences were deposited into GenBank under accession nos. ES439769–ES439782.

Refinements to existing annotations and unusual gene structures

We performed directed computational searches for these phenomena using RFC and CSF, and used the resulting lists to guide manual inspection and/or downstream computational analyses, leading to the choice of final cutoffs and data sets. For example, to identify likely recent nonsense mutations, we identified high-scoring regions in FlyBase transcripts downstream from *D. melanogaster* stop codons that align to sense codons in the other species; to identify possible stop codon readthrough genes, we identified similar cases where the stop codon is conserved across the informant species. See Supplemental Methods for further details and the cutoffs used.

Acknowledgments

We are indebted to the community effort for sequencing, assembly, and alignment of the 12 *Drosophila* genome sequences without which this project would not have been possible, and for the early release and collaborative data sharing. We thank Andy Clark, Tim Sackton, and Tony Greenberg for helpful discussions on lineage-specific genes; Gene Yeo and Jade Vinson for sharing code for a splice site discriminator; and Alex Stark, Pouya Kher-

apdour, Matt Rasmussen, Ameya Deoras, Josh Grochow, Erez Lieberman, and Aviva Presser for invaluable discussions.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Andrews, J., Smith, M., Merakovsky, J., Coulson, M., Hannan, F., and Kelly, L.E. 1996. The stoned locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics* **143:** 1699–1711.

Bass, B.L. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71:** 817–846.

Bergman, C.M., Pfeiffer, B.D., Rincon-Limas, D.E., Hoskins, R.A., Gnirke, A., Mungall, C.J., Wang, A.M., Kronmiller, B., Pacleb, J., Park, S., et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0086.

Bergstrom, D.E., Merli, C.A., Cygan, J.A., Shelby, R., and Blackman, R.K. 1995. Regulatory autonomy and molecular characterization of the *Drosophila* out at first gene. *Genetics* **139:** 1331–1346.

Bernal, A.E., Crammer, K., Hatzigeorgiou, A., and Pereira, F.C.N. 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput. Biol.* 3: e54. doi: 10.1371/journal.pcbi.0030054.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14:** 693–699.

Brent, M.R. 2005. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.* **15:** 1777–1786.

Brogna, S. and Ashburner, M. 1997. The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: Multigenic transcription in higher organisms. *EMBO J.* **16:** 2023–2031.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Casey, J.L. and Gerin, J.L. 1995. Hepatitis D virus RNA editing: Specific modification of adenosine in the antigenomic RNA. *J. Virol.* **69:** 7593–7600.

Castellano, S., Morozova, N., Morey, M., Berry, M.J., Serras, F., Corominas, M., and Guigo, R. 2001. In silico identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2:** 697–702.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., et al. 2002. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0079.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Farabaugh, P.J. 1996. Programmed translational frameshifting. *Microbiol. Rev.* **60:** 103–134.

Giedroc, D.P., Theimer, C.A., and Nixon, P.L. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298:** 167–185.

Gross, S.S., Russakovsky, O., Do, C.B., and Batzoglou, S. 2006. Training conditional random fields for maximum labelwise accuracy. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. http://books.nips.cc/papers/files/nips19/NIPS2006_0891.pdf.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* **7:** S4. doi: 10.1186/gb-2006-7-s1-s4.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Hoskins, R.A., Stapleton, M., George, R.A., Yu, C., Wan, K.H., Carlson, J.W., and Celniker, S.E. 2005. Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). *Nucleic Acids Res.* **33:** e185. doi: 10.1093/nar/gni184.

Ivanov, I.P., Simin, K., Letsou, A., Atkins, J.F., and Gesteland, R.F. 1998. 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.* **18:** 1553–1561.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E.S. 2004. Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.* **11:** 319–355.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59. doi: 10.1186/1471-2105-5-59.

Kryukov, G.V., Kryukov, V.M., and Gladyshev, V.N. 1999. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J. Biol. Chem.* **274:** 33888–33897.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289. Morgan Kaufmann, San Francisco.

Luo, G.X., Chao, M., Hsieh, S.Y., Sureau, C., Nishikura, K., and Taylor, J. 1990. A specific base transition occurs on replicating hepatitis delta virus RNA. *J. Virol.* **64:** 1021–1027.

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., et al. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38:** 1151–1158.

Martin-Romero, F.J., Kryukov, G.V., Lobanov, A.V., Carlson, B.A., Lee, B.J., Gladyshev, V.N., and Hatfield, D.L. 2001. Selenium metabolism in *Drosophila*: Selenoproteins, selenoproteins mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276:** 29798–29804.

Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0083.

Nekrutenko, A., Makova, K.D., and Li, W.H. 2002. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Res.* **12:** 198–202.

Parra, G., Blanco, E., and Guigo, R. 2000. GeneID in *Drosophila*. *Genome Res.* **10:** 511–515.

Reese, M.G., Kulp, D., Tammana, H., and Haussler, D. 2000. Genie—Gene finding in *Drosophila melanogaster*. *Genome Res.* **10:** 529–538.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.

Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. 2000. A *Drosophila* complementary DNA resource. *Science* **287:** 2222–2224.

Sarawagi, S. and Cohen, W. 2005. Semi-Markov conditional random fields for information extraction. *Adv. Neural Inf. Process. Syst.* **17:** 1185–1192.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

Stanke, M. and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19:** II215–II225.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., et al. 2002a. A *Drosophila* full-length cDNA resource. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0080.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., et al. 2002b. The *Drosophila* gene collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res.* **12:** 1294–1300.

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G.J., and Kellis, M. 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.* (this issue). doi: 10.1101/gr.6593807.

Steneberg, P., Englund, C., Kronhamn, J., Weaver, T.A., and Samakovlis, C. 1998. Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the *Drosophila* trachea. *Genes & Dev.* **12:** 956–967.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel,

A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., et al. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0088.
Wan, K.H., Yu, C., George, R.A., Carlson, J.W., Hoskins, R.A., Svirskas, R., Stapleton, M., and Celniker, S.E. 2006. High-throughput plasmid cDNA library screening. *Nat. Protoc.* **1:** 624–632.
Xue, F. and Cooley, L. 1993. kelch encodes a component of intercellular bridges in *Drosophila* egg chambers. *Cell* **72:** 681–693.
Yandell, M., Bailey, A.M., Misra, S., Shu, S., Wiel, C., Evans-Holm, M., Celniker, S.E., and Rubin, G.M. 2005. A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci.* **102:** 1566–1571.
Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15:** 496–503.
Yeo, G. and Burge, C.B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11:** 377–394.

Received May 7, 2007; accepted in revised form September 21, 2007.