Supplementary Material

Splicing database

Alternative splicing events (alternative donor/acceptor sites, skipped exon, intron retention) were classified as in Gupta *et al*., based on the pattern of exon boundaries described by alternative transcripts (note that the categories are not mutually exclusive) (1). Additionally, an intron was said to be associated with a putative polyadenylation event if it contained the 3' end of at least one FlyBase transcript. We identified possible orthologs of *D. melanogaster* splice sites in each of the other fruit flies by using pair-wise BLASTZ chain alignment nucleotides (2). If a splice site was contained in more than one chain, we selected the combination of chains that maximizes the number of splice sites aligned per gene. The quality of the alignment was assessed by computing the proportion of splice sites aligned per gene in each organism. Only 28% of splice sites were aligned in all 12 fruit flies; the others were not aligned because either the sequence data or the chain alignment was missing in one or more species. Out of 90,000 splice sites, 49% were present in all but one, 64% were present in all but two, 76% were present in all but three, 85% were present in all but four, and 95% were present in all but five fruit flies.

Splice site strength were computed based on scoring matrices from Mount *et al*. (the results did not change qualitatively when more recent splicing data were used to compute scoring matrices) (3). The profiles covered 5 nucleotides upstream and 7 nucleotides downstream of the donor site and 9 nucleotides upstream and 3 nucleotides downstream of the acceptor site,

respectively. The distributions of donor and acceptor splice site strengths were symmetric and bell-shaped, centering at 612.7 and 537.9 with standard deviations of 52.9 and 47.4, respectively. Within alternative donor and acceptor splice sites, the mean strengths were 589.4 and 527.3 with standard deviations of 59.8 and 52.1, respectively.

Secondary structure prediction

The algorithm for detection of conserved complementary regions (boxes) does not require a multiple sequence alignment. For each intron sequence, we identified all pairs of complementary words of continuous stretches of at least 9 nucleotides, one located within 150 nucleotides of the donor site and the other located within 150 nucleotides of the acceptor site. A small number of G·U base pairs were allowed. A constraint of having at least two G·C base pairs was introduced to reduce the number of AT-rich words. We next identified complementary words which: *i*) are found in orthologous introns of at least 7 of the 12 species; and *ii*) differ by at most 3 nucleotides between any two of these species. These words were extended to a common non-branching secondary structure, in which we allowed internal loops or bulges of at most 4 nucleotides and helices of at least 3 nucleotides, using thermodynamic parameters for RNA folding (4). For the purpose of graphic representation, orthologous sequences were aligned using MUSCLE software (5). The algorithm was implemented in C++ (the source code is available at www.bioinf.fbb.msu.ru/~dp/dros/).

Cross-validation against RNAalifold

In order to compare the performance of our method to that of RNAalifold (6), we selected all introns below 150 nts in length (30,000 introns), created multiple sequence alignments using MUSCLE software (5), and sent the MUSCLE output to RNAalifold. The RNAalifold output was then parsed for structures that fulfill the same constraints as required by our search: at least two G·C base pairs, at most 1 G·U base pair, conservation in at least 7 of the 12 Drosophilae, and a maximum difference of 3 nucleotides between any two of these species.

Tests of significance

Standard errors for proportions were computed by the formula $p(1 - p)/n$, where p is the population proportion and n is the sample size. Standard errors for average splice site strengths were computed by the formula $\sigma / \sqrt{n}$, where σ is the population standard deviation and n is the sample size. In order to identify strong cryptic splice sites in a given intron, we screened the window of 100 nucleotides within each annotated splice site and identified the best-scoring splice site consensus (3). The differences between strengths of the annotated splice site and that of the best-scoring consensus comprised a matched sample, to which the one-sample z-test was then applied. The χ^2 goodness of fit test for distributions of box positions was carried out using 6 equal bins, 25 nucleotides each.

P-value computation

We scored each pair of boxes by computing an individual p-value, which was defined as a probability of observing two complementary w-long words in n random sequences (w = 9, n varies from 7 to 12). Importantly, the probability p_1 of observing two complementary words in one sequence is context-dependent, e.g., CpG-rich sequences are more likely to contain a pair of complementary words than sequences with uniform nucleotide distribution (7). This probability was estimated from the first-order Markov model inferred from the local nucleotide context of the intron with an addition of small number of pseudo-counts that reflect the average nucleotide composition across the gene. The value of p_1 was computed by the formula

$$p_1 = (L - w)^2 \sum_{a_1,\ldots,a_w} f_{a_1} f_{a_2}^{a_1} f_{a_3}^{a_2} \ldots f_{a_w}^{a_{w-1}} \tilde{f}_{a_{w-1}}^{a_w} \ldots \tilde{f}_{a_2}^{a_3} \tilde{f}_{a_1}^{a_2} \tilde{f}^{a_1},$$

where L is the length of the sequence, f_a^b and $\tilde{f}_b^a$ are the Markov transition probabilities for the original sequence and its reverse complement, respectively, and f_a and $\tilde{f}^b$ are the corresponding marginal distributions. Next, given that one sequence contains complementary words, the conditional probability p_2 of observing two complementary words in another sequence depends on the degree of similarity between sequences: in an extreme case of 100% identity, it is not surprising at all to find a pair of complementary words in one sequence given that another sequence contains such a pair. Accordingly, the value of p_2 was estimated as $(\alpha^2 + (1-\alpha)^2/3)^{w-1}$, where α is the pair-wise fractional identity between species; the latter is derived from the local nucleotide context using the k-mer distance (8). The resulting p-

value was $p_1 p_2^{n-1}$. In order to account for multiple simultaneous tests, we used the correction of the form $p^* = 1 - (1 - p)^M$, where p^* is the corrected p-value and m is the number of independent tests.

Supplementary Figure Legends

Supplementary Table 1. See the legend for Table 1.

Supplementary Table 2. Genomic positions of the predicted intronic secondary structures. The columns are (left to right): rank number (the same as in Table 1); chromosome (FlyBase); strand; genomic position of box 1; genomic position of box 2; distance between boxes (d); absolute value of the equilibrium free energy (E, kcal/mol) and length (L) of the structure; the maximum number of nucleotides, by which the 9 nucleotide complementary boxes differ between species (ε); and nucleotide sequences of box 1 and box 2 .

1. Gupta, S., Zink, D., Korn, B., Vingron, M. and Haas, S.A. (2004) Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics*, **20**, 2579-2585.
2. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*, **31**, 51-54.
3. Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O. and Fields, C. (1992) Splicing signals in Drosophila: intron size, information content, and consensus sequences. *Nucleic Acids Res*, **20**, 4255-4262.
4. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911-940.
5. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
6. Hofacker, I.L. and Stadler, P.F. (2006) Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, **22**, 1172-1176.
7. Hefferon, T.W., Groman, J.D., Yurk, C.E. and Cutting, G.R. (2004) A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc Natl Acad Sci U S A*, **101**, 3504-3509.

8. Edgar, R.C. (2004) Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res*, **32**, 380-385.