Supplement for Lin, Carlson, Crosby *et al.*: Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using twelve fly genomes

Supplemental figures

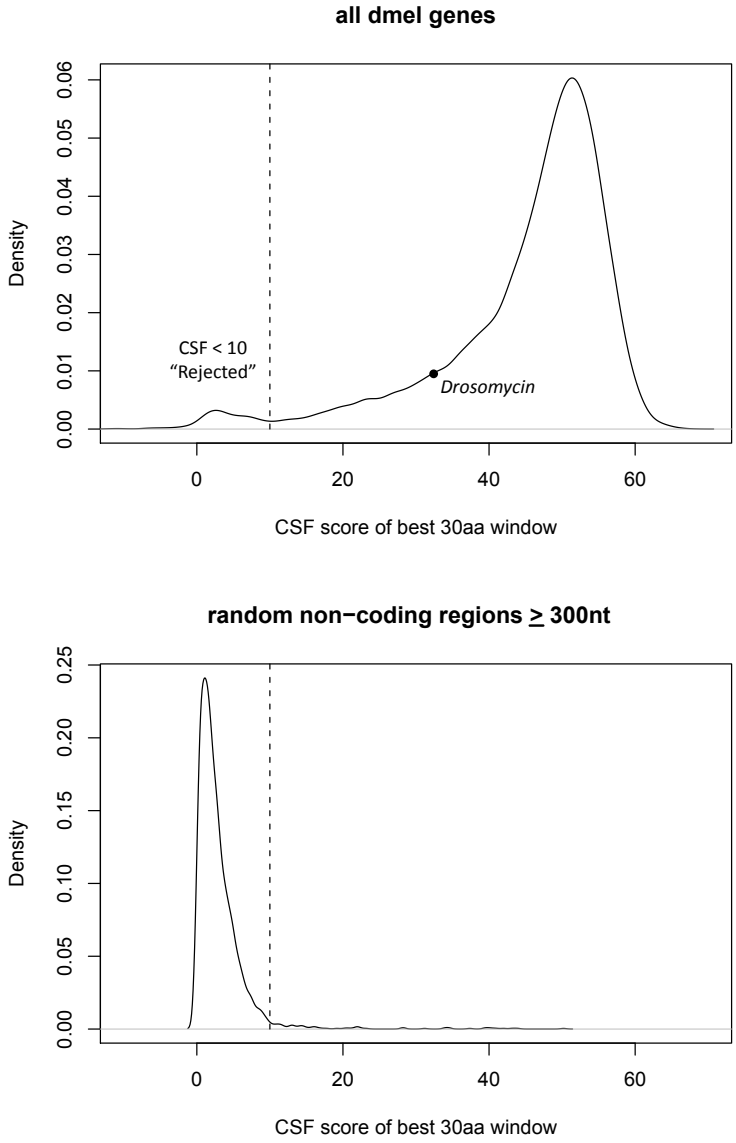


Figure S1. Selection of "rejected" genes. (top) The distribution of nonzero scores for the best-scoring 30aa window in each fly gene (see Supplemental Methods for details). Our test rejected genes with CSF < 10 on this distribution. Also shown is the score of *Drosomycin*, a known lineage-specific gene restricted to the *melanogaster* group. (bottom) The equivalent distribution for random non-coding regions $\geq$ 300nt (see Supplemental Methods for details about control regions). The "rejected" distribution closely resembles that of the random non-coding regions (although the peaks do not exactly align).

Figure S2. A candidate translational frameshift has a striking association with a highly conserved RNA structure. (This figure is attached to the end of this document due to its large size.)

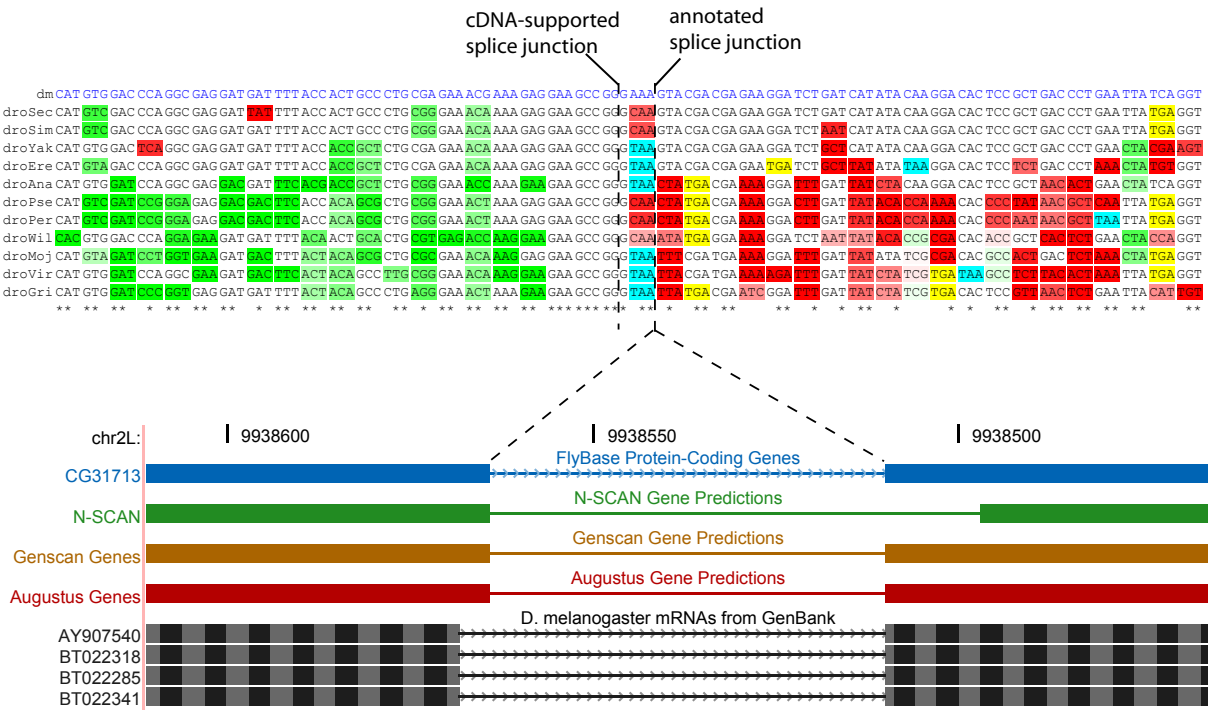


Figure S3. An example of a "suspicious" splice junction identified through evolutionary signatures. In the alignment of the transcript model, the reading frame of translation upon which selection appears to act abruptly changes at the annotated splice junction. In this particular case, there exist full-length cDNA sequences that indicate a donor site four bases upstream of the annotated site and support the putative corrected frame of the downstream exon. Closer examination reveals a possible non-canonical splicing mechanism (GA donor site) that likely prevented *de novo* gene predictors from identifying it. Alternatively, the GA donor site may be a strain-specific mutation. Annotations rendered by the UCSC Genome Browser [Kent et al., 2002].

Supplemental tables

CSF		RFC	
Sensitivity	Specificity	Sensitivity	Specificity
97.2%	95.1%	98.1%	91.2%
96.7%	96.2%	97.3%	94.5%
96.3%	97.1%	96.9%	95.6%
95.5%	98.2%	95.5%	97.1%
94.4%	99.1%	94.5%	97.5%
91.4%	99.9%	87.7%	98.4%

Table S1. Discriminatory power of CSF and RFC evolutionary metrics. Shown are the sensitivity and specificity (at various cutoffs) of each metric used to classify MULTIZ alignments of 5,567 exons of well-studied genes and 20,280 control regions randomly chosen from the non-coding part of the genome with the same length distribution. Percentages reflect the number of regions correctly classified. Note that we use the term "specificity" as it is defined in binary classification problems (the proportion of non-coding regions correctly classified as non-coding). Further information and comparison to other metrics can be found in a related paper (Lin, Deoras, Rasmussen and Kellis, submitted)

	Percent with hits, validated new genes	Percent with hits, known genes
D. ananassae	95%	90%
D. pseudoobscura	93%	77%
D. grimshawi	77%	82%
A. gambiae	37%	50%
A. mellifera	25%	45%
C. elegans	11%	28%
D. rerio	12%	31%
G. gallus	11%	31%
H. sapiens	11%	32%
S. cerevisiae	5%	20%

Table S2. TBLASTX homology searches suggest that newly discovered genes tend to be *Drosophila*- or insect-specific. Left-hand column shows the percentage of 57 non-redundant, full-length cDNA sequences representing newly discovered genes that have a TBLASTX hit (e-value $< 10^{-6}$) to the RepeatMasked genome assembly of each species. Right-hand column shows the corresponding percentage for 228 transcripts randomly chosen from FlyBase annotation release 4.3 with a comparable length distribution.

	Well-studied	Named	CGid-only	All genes
Exons	5,567	22,814	31,257	54,048
Exons missed	22.5%	24.3%	28.6%	26.8%
Exons poorly aligned[1]	6.0%	9.1%	12.8%	11.2%
Nucleotides	2.5 mbp	9.2 mbp	12.8 mbp	22.0 mbp
Nucleotides missed	11.3%	14.0%	21.0%	18.1%
Genes	893	4,711	9,022	13,733
Genes missed[2]	4.1%	10.0%	17.3%	14.8%

Table S3. Genome-wide sensitivity of our *de novo* exon prediction algorithm with respect to FlyBase annotation release 4.3. [1]Exons having no informant species outside of the *melanogaster* subgroup with at least 80% of bases aligned, in the Mercator/MAVID alignments to which our algorithm was applied. (The *melanogaster* subgroup spans neutral divergence comparable to primates.) [2]Genes in which all exons were missed. Note: this data is not directly related to the gene-level "confirmation" rates shown in Table 1, since our evaluations of complete gene models were carried out across their full length and using several genome alignment sets, nor to the raw discriminatory power of RFC and CSF shown in Table S1, which are based on the more sensitive MULTIZ alignments and do not involve predicting a segmentation.

Supplemental methods

Genome assemblies, alignments, annotations

We used "Comparative Analysis Freeze 1" assemblies of the twelve *Drosophila* genomes available from the following web site: `http://rana.lbl.gov/drosophila/assemblies.html`. We used two genome alignment sets derived from a synteny map generated by Mercator (C. Dewey, `http://www.biostat.wisc.edu/~cdewey/mercator/`). One was generated using the MAVID sequence aligner [Bray and Pachter, 2004] and the other generated by PECAN (B. Paten and E. Birney, `http://www.ebi.ac.uk/~bjp/pecan/`). Additionally, we used a third set of genome alignments generated by MULTIZ [Blanchette et al., 2004]. We obtained these alignments from the following web site: `http://rana.lbl.gov/drosophila/wiki/index.php/Alignment`.

We obtained FlyBase release 4.3 annotations from the following web site:
`ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r4.3_20060303/gff`.

CSF metric

The **Codon Substitution Frequencies (CSF)** metric is based on estimates of the frequencies at which all pairs of codons are substituted between genes in the target species and the informants. First, let us consider computing the score for a pairwise alignment only. Consider the alignment

of a putative ORF/exon as two sequences of codons A and B, where A_k is the target codon that aligns to the informant codon B_k at position k in the target codon sequence (position $3k$ in the in-frame target nucleotide sequence). CSF assigns a score to each codon position k where: (1) A_k and B_k are both un-gapped triplets, (2) A_k is not a stop codon, and (3) $A_k \neq B_k$. CSF then sums these scores to obtain an overall score for the sequence.

The score assigned to a codon substitution (a, b) is a log-likelihood ratio indicating how much more frequently that substitution occurs in coding regions than in non-coding regions. Each likelihood compared in this ratio is derived from a Codon Substitution Matrix (CSM), where

$$CSM_{a,b} = \mathbf{P}(\text{informant codon } b | \text{target codon } a, a \neq b)$$

The entries of the CSM are estimated for each target and informant by counting aligned codon pairs in training data, and then normalizing the rows to obtain the desired conditional probabilities. We train two CSMs, one for which the training data is alignments of known genes (CSM^C) and one for which the training data is alignments of random non-coding regions (CSM^N). The score that CSF assigns a codon substitution (a, b) is then $\log \frac{CSM^C_{a,b}}{CSM^N_{a,b}}$. For example, these scores for *D. melanogaster* and *D. ananassae* are visualized in Figure 1B.

With multiple informants, CSF uses an *ad hoc* strategy to combine evidence from the informants without double-counting multiple apparent substitutions among extant species that result from fewer evolutionary events in their ancestors. For each target codon position k, CSF assigns a score to codon substitutions between the target and each informant exactly as in the pairwise case, using the appropriate CSMs for each informant. CSF then takes the median of these scores to obtain a composite score for position k, and sums these composite scores to obtain an overall score for the sequence. Note that the median is usually taken on fewer than n pairwise scores, since the pairwise scores are only assigned to ungapped informant codons that differ from the target codon.

In this study, we used CSMs trained on all annotated *D. melanogaster* genes with each informant. We have also carried out cross-validated benchmarks with smaller training sets, with no appreciable difference in discriminatory power (Lin, Deoras, Rasmussen and Kellis, submitted).

Lastly, we note that CSF makes no attempt to explicitly "correct" for several well-known issues that frequently arise in modeling codon evolution, such as transitions/transversions, CpG hyper-

mutation, codon bias, site-specific rate variation, etc. The purpose of CSF is neither to realistically model evolution nor to obtain precise estimates of evolutionary rates, but rather to provide a computationally efficient metric that discriminates between coding and non-coding regions.

Evaluation and classification of existing gene annotations

For each euchromatic gene in FlyBase annotation release 4.3, we applied the RFC and CSF metrics to each of its transcript models. To score a transcript, we first generated an alignment by extracting each of its exons from whole-genome sequence alignments and then "splicing" them. We then used the best-scoring transcript model as a proxy for the gene, where the best-scoring transcript model is the one with the highest RFC score, or, in the event of a tie of the RFC score, the highest CSF score.

To define a test for whether the evolutionary evidence "confirms" each gene, we chose cutoffs on the RFC and CSF scores (computed in the MULTIZ alignments) based on random controls as follows. We extracted 15,564 regions $\geq$ 300nt in length from the genome sequence alignments, chosen uniformly at random from the portion of the genome not annotated as protein-coding. These alignments were preprocessed to remove columns containing in-frame stop codons in *D. melanogaster* (each control region is $\geq$ 300nt in length *after* removing stop codons, a detail previously omitted) and then scored by RFC and CSF. We considered a gene "confirmed" if its RFC score was greater than zero and its length-normalized CSF score (the CSF score divided by the length in nucleotides of the ORF) was greater than or equal to 0.03, cutoffs which exclude all but three of the 15,564 control regions (see Table 1). One of these three "false positive" regions coincided with a predicted new exon that was later validated by our cDNA sequencing experiments, and, following manual inspection, we consider the other two also likely to represent genuine coding sequence. Thus, our criteria for "confirmation" of a gene was very stringent, insofar as virtually no non-coding regions $\geq$ 300nt passed this test.

We next defined a much more relaxed test to identify gene annotations that not only fail to satisfy the above stringent criteria, but appear unlikely even to represent genuine protein-coding genes. We computed the CSF score over every overlapping 30aa window in every transcript model for each gene. Additionally, we computed these scores using the three different genome alignment sets and using three different subsets of the informant species, representing all twelve *Drosophila*

genomes, the subgenus *Sophophora*, and the *melanogaster* group. We took the highest scoring 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. The distribution of this score across all genes was clearly bimodal (Supplemental Figure 1). We chose a cutoff selecting the 454 genes forming the lower distribution as the "rejected" genes. Genes that were neither "confirmed" nor "rejected" by these tests form the "unclear" category (Table 1).

The scores and classification of each gene and the random control regions can be found in our online supplemental information (see below).

Predicting new exons

In order to define the precise boundaries of genomic regions showing RFC and CSF evolutionary signatures that are likely to represent new exons, we integrated our evolutionary metrics as features into a simple *de novo* exon predictor based on a semi-Markov conditional random field (SMCRF [Lafferty et al., 2001, Sarawagi and Cohen, 2005]), a probabilistic graphical model similar to a generalized hidden Markov model (GHMM). Unlike a GHMM, however, an SMCRF can *directly* incorporate any metric that provides a real-valued score for any segment of the genome, such as RFC and CSF. Our system is a straightforward application of standard SMCRF algorithms to parse the genome into coding and non-coding segments based on our metrics. In this sense, it may be considered more similar to simple interval segmentation algorithms that compute boundaries of high-scoring regions, than to full gene predictors such as GENSCAN [Burge and Karlin, 1997] or N-SCAN [Gross and Brent, 2006]. Initial applications of SMCRFs to create full gene predictors have recently been reported [Decaprio et al., 2007, Vinson et al., 2007, Bernal et al., 2007].

SMCRF structure. The graphical structure (state diagram) of our model follows the example of ExoniPhy [Siepel and Haussler, 2004], with some simplifications enabled by the more flexible nature of the SMCRF than the phylo-HMM used in that system. In particular, the model has only seven segment labels (states): one for each codon reading frame on each strand (+1, +2, +3, -1, -2, -3), and one for non-coding positions. Since each coding state labels a segment, not an individual nucleotide, the labels (+1, +2, +3, -1, -2, -3) specify the codon reading frame in which the segment should be read. For example, the label +1 means that the segment should be read as beginning on the first position of a complete codon on the positive strand. Each "coding" state is bidirectionally

connected to the non-coding state, the non-coding state is self-connected, and no other transitions are possible.

If there is no maximum segment length, then the SMCRF training and decoding algorithms have running time quadratic in the sequence length. Therefore, for practical reasons, non-coding "segments" are constrained to be one nucleotide in length, with non-coding regions modeled as sequences of 1nt non-coding segments. The maximum length of coding segments is *de facto* constrained by disallowing in-frame stop codons.

Feature functions. The features used by the SMCRF include:

1. the evolutionary metrics, which score coding segments.

2. indicator functions for start and stop codons, which score transitions between coding and non-coding segments. These are binary functions (later assigned a numerical weight by the SMCRF) indicating the presence of a start or stop codon in the *D. melanogaster* sequence. They also enforce "well-formedness" constraints: the start codon feature disallows (by returning a negative-infinity score) noncoding-to-coding transitions in the absence of a start codon or AG splice site, and the stop codon feature disallows coding-to-noncoding transitions in the absence of a stop codon or GT splice site. The stop codon feature also disallows coding segments with in-frame stop codons.

3. sequence-based discriminators for acceptor and donor sites, which score transitions between coding and non-coding segments on AG and GT splice sites (based on the *D. melanogaster* sequence). These discriminators, provided by the authors of a previous study [Yeo and Burge, 2004], consider 23 nucleotides surrounding acceptor sites and 9 nucleotides surrounding donor sites based on the principle of maximum entropy.

4. length distribution feature, which was set to a simple geometric distribution corresponding to the empirical mean lengths of annotated exons and non-coding regions. (We did not investigate other exon length distributions at the time of freezing our prediction set for this study, although this is possible in principle.)

Importantly, the SMCRF did not include any explicit coding sequence composition features (e.g. high-order Markov models), nor did it use any information about transcript sequence evidence or

homology to known proteins.

Training and decoding. The SMCRF training procedure determines optimal weights for a linear combination of the features. We trained our SMCRF using the standard maximum conditional likelihood algorithm [Lafferty et al., 2001, Sarawagi and Cohen, 2005] on a training set of 100 known genes. (The SMCRF training procedure must estimate only one parameter for each feature, in contrast to GHMM gene predictors which require thousands of generative parameters. The SMCRF thus requires less training data. In our case, much of the additional information that would be estimated in GHMM training is captured in the CSMs used by CSF.) We then used the SMCRF equivalent of the Viterbi algorithm to decode the whole *D. melanogaster* genome in the Mercator/MAVID alignments into coding exons and non-coding regions (see also Table S3).

All predicted exons that did not overlap any coding exon in FlyBase annotation release 4.3 (on the coding strand) were regarded as predicted new exons, except for a total of 217 predictions that were either within *Dscam* (a gene with exceptionally many exons and splice forms that are known but not represented in FlyBase), heterochromatic regions, or redundant or misassembled regions of the euchromatic genome assembly. For historical reasons, the new exon predictions were carried out only using the Mercator/MAVID alignments; the MULTIZ alignments of the 12 flies were not available until our experimental validation and manual curation efforts were already underway for a frozen prediction set.

Manual curation of new exons

FlyBase manual annotation procedures, including the various data sources examined by the curators, are described in detail in FlyBase documentation:
`http://flybase.bio.indiana.edu/static_pages/docs/refman/refman-G.html#G7`

Unusual protein-coding structures and adjustments to existing annotations

Translation start sites. We identified candidate alternate translation start sites (`ATG2`) downstream of the annotated site (`ATG1`) in all FlyBase transcripts. We selected the cases in which: (1) `ATG2` was conserved in more of the informant species than `ATG1`; (2) the CSF score of the region between `ATG1` and `ATG2` was negative; (3) The CSF score of the region downstream of `ATG2`, of the same length as the distance between `ATG1` and `ATG2`, was positive; (4) the proposed adjustment

shortened the protein by less than 50%. In transcripts for which there were multiple ATG2s that satisfied these criteria, we selected the one that was most highly conserved, or in case of a tie, the 5'-most.

Stop codon readthrough and recent nonsense mutations. We examined all FlyBase coding transcripts in which an ORF of at least 15aa extends immediately downstream of the annotated stop codon. If the transcript contained 3'UTR introns, then we examined the downstream ORF in the "spliced" alignment. Otherwise, we scored the downstream ORF in the genome, even if it extended beyond the annotated 3' end of the transcript (since these 3'-end annotations are not consistently based on biological evidence). We selected the transcripts that satisfied these criteria: (1) the RFC and CSF scores of the entire downstream ORF passed our confirmation criteria and (2) the CSF score of the first 20aa (or length of the ORF, whichever was shorter) was positive. These filters identified 257 transcripts. We then manually examined the genome sequence alignments of each of these transcripts to evaluate the conservation of the downstream ORF, leading to the 149 candidates for future investigation. A primary requirement for these candidates, in addition to good conservation of the downstream ORF, was conservation of the putatively read-through stop codon. Cases in which the stop codon aligned to sense codons in the informant species were recorded as likely recent nonsense mutations.

Polycistronic transcripts. To identify new candidate polycistronic transcripts, we searched for start-to-stop ORFs of at least 20aa within the annotated UTRs of existing fly transcripts, which passed our RFC and CSF confirmation criteria and did not overlap the annotated coding region of any other transcript, or any of the candidates for stop codon readthrough (since these frequently appeared as candidate 3' ORFs starting at their first start codon). These criteria led to the 135 reported cases. Additionally, among the high-scoring ORFs that *did* overlap the coding sequence of another transcript were 73% of the annotated dicistronic transcripts (see main text).

Translational frameshifts and recent frameshift mutations. To identify candidate translational frameshifts, we first used CSF to score all overlapping 20aa windows in the genome, in all frames and on both strands. We then identified adjacent windows that score highly in different reading frames (both windows score ≥ 10 and higher than other reading frames and the opposite strand). These filters identified 40 cases. We then manually examined the genome sequence alignments of these cases to identify the four in which there appears to be an abrupt frameshift that

is conserved across the informant species (i.e. with no nearby indels that might indicate a recent mutation). This manual inspection also led to the five cases that do appear to be recent mutations, in which the *D. melanogaster* sequence has a unique frame-shifting indel.

Adjustments to existing exon annotations. To identify possibly problematic transcript models, we used an approach similar to that which we used to identify candidate translational frameshifts. Instead of examining overlapping windows genome-wide, we scored windows within the ORFs of all FlyBase transcripts, and selected "frameshifts" that occur across intron boundaries in the spliced transcript models.

Online supplemental information

Additional data can be found on our web site `http://www.broad.mit.edu/~mlin/fly_genes/`

- New exons
  - Prediction coordinates and sequences
  - Manual curation and cDNA sequencing records for each prediction
  - Recovered full-length cDNA sequences
- Scores and classification of all annotated genes and random control regions
- Adjustemnts to existing annotations
  - List of translation start site adjustments
  - List of exon boundary adjustments
  - List of "suspicious" splice junctions
- Unusual gene structures
  - List of candidate translational readthrough genes
  - List of candidate polycistronic ORFs
  - List of candidate translational frameshifts

References

[Bernal et al., 2007] Bernal, A., Crammer, K., Hatzigeorgiou, A., and Pereira, F., 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Computational Biology*, **3**(3):e54.

[Blanchette et al., 2004] 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**(4):708–715.

[Bray and Pachter, 2004] Bray, N. and Pachter, L., 2004. Mavid: Constrained ancestral alignment of multiple sequences. *Genome Res.*, **14**(4):693–699.

[Burge and Karlin, 1997] Burge, C. and Karlin, S., 1997. Prediction of complete gene structures in human genomic dna. *J Mol Biol*, **268**(1):78–94.

[Decaprio et al., 2007] Decaprio, D., Vinson, J. P., Pearson, M. D., Montgomery, P., Doherty, M., and Galagan, J. E., 2007. Conrad: Gene prediction using conditional random fields. *Genome Res*, .

[Gross and Brent, 2006] Gross, S. and Brent, M., 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol*, **13**:379–393.

[Kent et al., 2002] 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**(6):996–1006.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

[Sarawagi and Cohen, 2005] Sarawagi, S. and Cohen, W., 2005. Semi-markov conditional random fields for information extraction. *Advances in Neural Information Processing Systems*, **17**:1185–1192.

[Siepel and Haussler, 2004] Siepel, A. and Haussler, D., 2004. Computational identification of evolutionarily conserved exons. *Proceedings of the Eighth Annual International Conference on Resaerch in Computational Molecular Biology (RECOMB '05)*, :177–186.

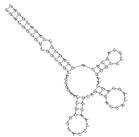
[Vinson et al., 2007] Vinson, J. P., DeCaprio, D., Pearson, M. D., Luoma, S., and Galagan, J. E., 2007. Comparative gene prediction using conditional random fields. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1441–1448. MIT Press, Cambridge, MA.

[Yeo and Burge, 2004] Yeo, G. and Burge, C. B., 2004. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *J Comput Biol*, **11**(2-3):377–394.

Region shown:
D. melanogaster chr3R:24,661,222-24,661,851
(within an exon of *Apc*)

Apc-RA

highly conserved RNA fold



apparent -1 frameshift (exact position unknown)

end of shifted ORF

selection reverts to original frame