Protein-Coding Structured RNAs
A Computational Survey of Conserved RNA Secondary Structures Overlapping Coding Regions in Drosophilids

Sven Findeiß[a], Jan Engelhardt[b,d], Sonja J. Prohaska[b,c], Peter F. Stadler[a,b,d,e,f,g,h]

[a]*Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[b]*Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[c]*Computational EvoDevo Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[d]*Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[e]*Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[f]*Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*
[g]*Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark*
[h]*Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

Functional RNA elements can be embedded also within exonic sequences coding for functional proteins. While not uncommon in viruses, only a few examples of this type have been described in some detail for eukaryotic genomes. Here we use `RNAz` and `RNAcode`, two comparative genomics methods that measure signatures of stabilizing selection acting on RNA secondary structure and peptide sequence, resp., to survey the fruit fly genomes. We estimate that there might be on the order of 1000 loci that are subject to dual selection pressure. The used genome-wide screens also expose the limitations of the currently available methods.

Key words: dual RNAs; RNA secondary structures; coding sequence; `RNAz`; `RNAcode`;

1. Introduction

The worlds of protein-coding genes and those of non-coding RNAs (ncRNAs) are often seen as clearly separated. A small number of examples from both prokaryotes and eukaryotes, however, demonstrates that this is not strictly true, see [1] for a recent review. The best studied case in animals is the Steroid receptor activator gene (SRA). Originally characterized as a non-coding RNA with a distinctive secondary structure [2], it was later found to have isoforms coding for the functional protein SRAP, see [3]. SRA is probably the most extreme example, as nearly the complete transcript is covered by both conserved RNA structure and protein-coding region. At the other extreme, however, structured RNA motifs, such as selenocystein insertion elements (SECIS), internal ribosome entry sites (IRES), or mRNA localization signals are not at all frequent in UTRs of protein-coding transcripts [4]. In some cases, in particular in viruses, such structured elements are found in coding regions. The software tool `RNAdecoder` [5], which implements a comparative method for finding and folding RNA secondary structures within protein-coding regions, provided statistical evidence for frequent superpositions of RNA structure and coding sequences [6]. Nevertheless, systematic genome-wide analyses of this phenomenon have not been published to date.

The overwhelming majority of annotated coding regions translates to proteins with more than 30 amino acids. The discovery of the very short, independently encoded *Tarsal-less* peptides [7, 8], however, suggests that more small ORFs of this type might be hidden in the genomic DNA. Examples such as plant ENOD40 [9] with its short ORFs embedded in a heavily structured RNA, furthermore, hint at a possible association of functional secondary structure with coding capacity in such atypical transcripts. These, however, are likely to have escaped standard gene annotation procedures.

We therefore start our survey of the drosophilid genomes with the independent prediction of evolution-

Email addresses: `sven@tbi.univie.ac.at` (Sven Findeiß), `jane@bioinf.uni-leipzig.de` (Jan Engelhardt), `sonja@bioinf.uni-leipzig.de` (Sonja J. Prohaska), `studla@bioinf.uni-leipzig.de` (Peter F. Stadler)

arily conserved RNA secondary structures and of open reading frames with evidence of stabilizing selection acting on the peptide sequence. To this end we employ `RNAz` [10, 11] and `RNAcode` [12], respectively.

2. Methods Summary

This study is based on the 15-way `Multiz` alignment of insect genomes, which contains the 12 sequenced drosophilids, mosquito, honeybee and beetle. All analysis refers to the genome of *Drosophila melanogaster*. Conserved secondary structure were assessed using `RNAz 2.0` [11] after slicing the alignment blocks into overlapping windows and removing alignment slices with too many gaps, too few sequences, or too extreme sequence divergence. Selection pressure on peptide sequence level was examined by `RNAcode` [12] with $p = 0.05$ as cut-off for individual high scoring segment (HSS). Adjacent HSS in the same reading frame separated by less than 51 nt were combined into a single hit (cHSS) with a p-value estimated as the product of its constituent HSS. `RNAdecoder` was used as described in [5]. The genomic MAF alignments were converted into col-format with codon positions determined from the `RNAcode` annotation. GO-term enrichtments were computed using `FuncAssociate` [13].

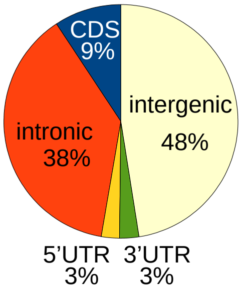
Due to space limitations detailed methods can be found online [1].

3. Results

3.1. *RNAz Screen*

`RNAz 2.0` detected 15912 loci (covering about 1.3 Mb) that show evidence for evolutionarily conserved secondary structure. Of these, 394 correspond or can be aligned with known structured ncRNAs. Taking into account that a sizable fraction of known ncRNAs are not included in the input alignments (notably tRNAs, due to variations in the number of tRNA genes in the various flies, and microRNAs, due the abundance of species-specific miRNAs), this amounts to an overall recall rate of about 69% (319/465), Figure 1, relative to the sequence alignments.

Compared to the previous `RNAz` screen in flies [14] we achieve a similar recall rate (69% vs. 65%) on the structured ncRNAs. On the other hand, the total number of predicted `RNAz` loci is reduced here by almost a factor of three (15912 vs. 42482). This discrepancy can be explained, on the one hand, by the different input alignments (here `Multiz` vs. `Pecan` in [14]), and on the other hand by using an improved version of `RNAz` [11], which is substantially more specific.



Summary

genomic DNA	129682844 nt
aligned DNA	129302568 nt
screened	71777669 nt
`RNAz` $p > 0.5$	
hits	15912
`RNAz` $p > 0.9$	
hits	6121

Recall

class	aligned	recovered	
tRNAs	179	148	83%
snoRNAs	130	52	40%
miRNAs	104	84	82%
other	52	35	67%

Figure 1: Summary of the `RNAz` screen using *D. melanogaster* as reference. Left panel: overlap of `RNAz` hits and protein-coding gene annotation. Right panel: Summary statistics of the `RNAz` screen and recall of the screen on known ncRNAs contained in alignment blocks that could be evaluated by `RNAz`.

`RNAz` hits located in 5' and 3' UTRs might correspond to structured regulatory elements. Unfortunately, we miss most of the Rfam annotated structured elements such as the bicoid 3'UTR regulatory element (bicoid_3) or the nanos 3' UTR translation control element (nos_TCE), since the corresponding loci do not pass the filtering steps of the input alignments. The small fraction of UTR elements might thus be an underestimate caused by the comparably poor sequence conservation in UTR regions, which in turn renders sequences alignments less reliable for these loci.

3.2. *RNAcode Screen*

`RNAcode` determined 131197 cHSS in the *D. melanogaster* genome. This in particular includes also coding sequences in transposable elements that are not part of the CDS annotation. After removing all cHSS that overlap annotated repeats, we are left with 95355 loci, of which 65788 (65%) overlapped on the same strand with annotated CDSs. This amounts to 94.7% of the sequence covered by `RNAcode` predictions. Most of the remaining 26702 cHSS are short and have a p-value larger than 0.001. At $p = 0.001$ only 5697 predictions outside annotated CDS are left, Figure 2. Due to the relaxed p-value cutoff we have to expect a large false discovery rate. On the other hand, the sensitivity is needed to include very short peptides such as *Tarsal-less* in the screen.

[1] `http://www.bioinf.uni-leipzig.de/publications/supplements/11-012`

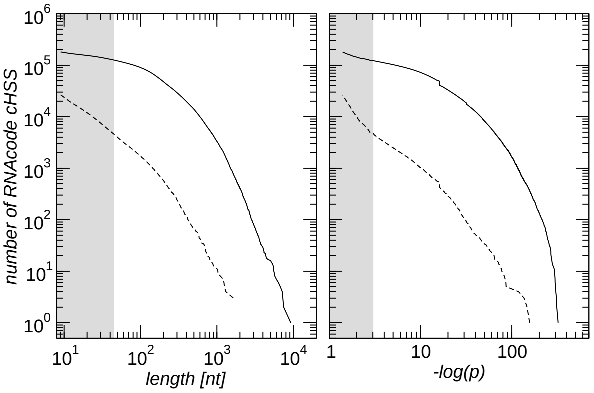


Figure 2: Summary of the RNAcode screen. Cumulative distribution of cHSS as function of lower bound on the length (l.h.s.) and upper bound on the *p*-value (r.h.s). The set of RNAcode cHSS (full line) is compared to the cHSS not previously annotated as coding sequence (dashed line). Gray background marks regions with very high false discovery rates.

Short peptides. In order to define a plausible set of candidates for novel small proteins that have been overlooked in existing annotation we choose a cut-off of $p < 0.001$ and a minimum length of 45 nt (i.e., 15 AA). Of the 2388 loci that do not overlap an annotated CDS or an annotated repeat, we find 352 RNAcode hits that map to exonic parts of FlyBase genes, and 439 hits overlap “FlyBase NonCoding transcripts”. A total of 654 hits overlap ESTs. After removing overlaps with FlyBase exons and “FlyBase NonCoding” elements we are left with 328 candidates of previously undescribed coding sequences. The majority of these candidates are additional exons of previously annotated proteins. Furthermore, we find clusters of RNAcode hits close to the ends of the chromosomes and in the heterochromatic DNA. These are likely related to un-annotated repetitive elements. Manual inspection nevertheless identifies several dozens of loci with novel small proteins.

Dicistronic transcripts and upstream ORFs. We find 487 cHSS that do not overlap any annotated CDSs within the annotated 5’ UTRs of 458 distinct genes. These predictions constitute candidates for dicistronic transcripts and so-called upstream ORFs (uORFs). In particular, RNAcode predicts cHSSs for 30 of the 31 dicistronic ORFs reported by [15], missing only *Kaz1-ORF A*.

Since RNAcode by construction does not produce complete gene models, we tested whether these cHSSs could be extended to an upstream start codon (ATG) and a downstream stop codon (TAG, TAA, or TGA) within 51 nt. Beyond the current FlyBase annotation, we identify 26 novel un-spliced protein candidates with lengths ranging from 31 to 110 amino acids. We compared these predictions to the results of [16], who identified 44 putative conserved peptide uORFs with evidence for stabilizing selection. Their data set consists of 19 spliced uORFs, all of which we recover by RNAcode cHSS, and 25 un-spliced uORFs. Of the latter, 10 are present in the current FlyBase annotation and were therefore excluded from our candidate set. All uORFs annotated by Flybase are also detected as cHSSs in our screen. Of the remaining 15 uORFs of [16], four are contained in our candidate set, leaving 11 cases without RNAcode signal. Conversely, we report 22 previously undescribed uORF candidates.

Table 1: Fraction of coding regions for which RNAdecoder predicts a structured region of minimum length *L* with average probability of 0.9 that the region is structured. To independently evaluate the predicted dual functional loci of our survey RNAdecoder was applied to RNAcode predictions that overlapped with RNAz hits at two significance levels. The background is estimated based on RNAcode predictions without RNAz overlap.

	Background	RNAz	
L		$p > 0.5$	$p > 0.9$
20	0.40	0.58	0.54
40	0.32	0.48	0.45
100	0.17	0.29	0.27

Read-through peptides. Read-through translation, i.e., skipping of stop codons, has been observed for several Drosophila genes. A similar effect is regularly observed for selenoproteins, where a stop-codon is re-interpreted as selenocystein under the influence of a conserved RNA secondary structure element in the 3’ UTR, the SECIS element [17]. Recently, SECIS-dependent read-through without selenocysteine incorporation has also been reported [18]. A computational study [19] suggests nearly 100 genes that are candidates for read-through translation. We therefore extracted all RNAcode hits that overlap annotated 3’ UTRs. In order to remove known isoforms, we considered only 3’ UTRs that do not overlap known CDS. Among the 229 candidates, 175 (76%) are in phase with the annotated stop codon. Furthermore, there are 72 novel RNAcode hits that overlap the stop codon, of which 54 (75%) are in phase. The strong enrichment of predictions in phase provides further evidence that read-through translation is a common phenomenon in fruit flies that leads to functional protein variants.

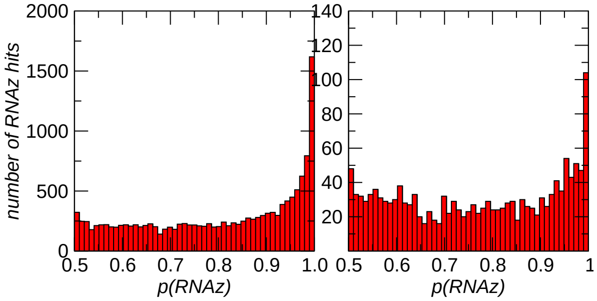


Figure 3: Comparison of RNAz prediction confidence values outside of CDS (l.h.s.) and overlapping CDS (r.h.s.). We obtain fewer high-confidence prediction in coding regions compared to other genomic regions.

3.3. *Dual Loci: Overlap of RNAz and RNAcode*

About 9% of RNAz loci, 1500 at $p_{RNAz} = 0.5$ and 449 at $p_{RNAz} = 0.9$, overlap annotated coding regions. In addition, some fifty previously unrecognized RNAcode predictions overlap with RNAz hits. Figure 3 compares the RNAz confidence values of predictions inside and outside of coding regions. The distribution is shifted towards lower confidence prediction in the coding regions, suggesting that RNAz has a moderately elevated false discovery rate on coding sequences. In the absence of a known positive set, however, this cannot be tested directly.

RNAdecoder [5, 6] is an SCFG-based tool specifically designed to detect conserved RNA secondary structure that is superimposed on coding sequences. The fraction of positive predictions is indeed elevated by 30% to 50% in coding RNAz hits compared to RNAcode predictions that are classified as unstructured by RNAz, Table 1. The false discovery rates of RNAdecoder, however, appears unacceptably large on the available genome-wide alignments, probably as a consequence of the limited alignment quality. It appears impossible, therefore to apply RNAdecoder independently to our data set.

The maternal effect protein *oskar* is currently the only known example of a “dual function RNA” in the fruit fly [20]. The transcript functions as a non-coding RNA during oogenesis, while the *oskar* protein is produced in the embryo, recently reviewed by [1]. Figure 4 depicts the genomic region corresponding to the *oskar* gene. Interestingly, two structured regions in the exons and one in the 3’ UTR have been predicted by RNAz. [21] showed that the 3’ UTR is sufficient to recover the full regulatory RNA function. However, to our knowledge the exact region and regulatory mechanism remains unknown. We hypothesize that the structured subregion within the 3’ UTR might cause the observed effect.

The absence of validated mRNAs in which functional RNA structure and coding sequence are superimposed implies that we also lack positive controls to estimate the sensitivity of our computational screen. Since coding sequences are typically rather well-conserved we can expect that the majority of such elements will be included in the input alignments. Assuming that both the sensitivity and the specificity of RNAz is comparable between CDS and other regions of the genome, we estimate that the fruit fly may contain on the order of 1000 structured coding regions. This ball park figure is based on a sensitivity of about 2/3 and a false discovery rate of about 50-60%, as estimated in several previous RNAz screens, see e.g. [14, 11].

In order to see if secondary structure structure superimposed on coding region is associated with particular gene functions, we performed a GO-term enrichment analysis. No significant associations were detectable, however.

3.4. *Effects on base pairing patterns*

The idea of investigating the conflicting selective constraints of RNA secondary structures superimposed on coding sequence data dates back at least to the 1970s, when Walter Fitch observed that optimization of RNA structure imposes limitations in nucleic acid sequences and thus should influence the amino acid sequence [22]. The different selective constraints on the three codon positions, furthermore, lead to different expected substitution patterns depending on the phase of the nucleotides with respect to the underlying coding sequence [23]. The latter effect is captured in the model underlying RNAdecoder. On the other hand, substantial differences in the nucleotide composition of the three codon position have been observed in fruit flies [24]. This bias in particular influences the stickiness, i.e., it makes the probability that two nucleotides can pair dependent upon the two codon positions.

In order to investigate the magnitude of these effects we compared the base pairing patterns of putative dual loci (with secondary structure superimposed on coding sequence in the same reading direction), dual loci in antisense orientation, structured RNA without conserved coding sequence, and protein-coding sequences without conserved secondary structure, Figure 5. Surprisingly, the stickiness is largest for anti-sense arrangements of CDS and RNA structure. The parallel arrangement, on the other hand, shows only a small increase in stickiness over coding sequence without constrained secondary structures.

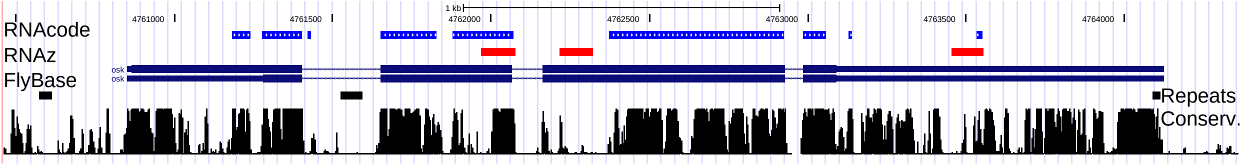


Figure 4: Genomic region (chr3R:4,760,853-4,764,123) encoding the oskar gene. Two alternative protein-coding transcripts are annotated in the FlyBase. Exons are indicated by filled boxes and the thin lines that connect them correspond to intronic sequences. Arrows indicate the reading direction. On top RNAcode cHSSs (light blue) and RNAz hits with $p > 0.9$ (red) are shown. RepeatMasker and conservation tracks are indicated at the bottom.

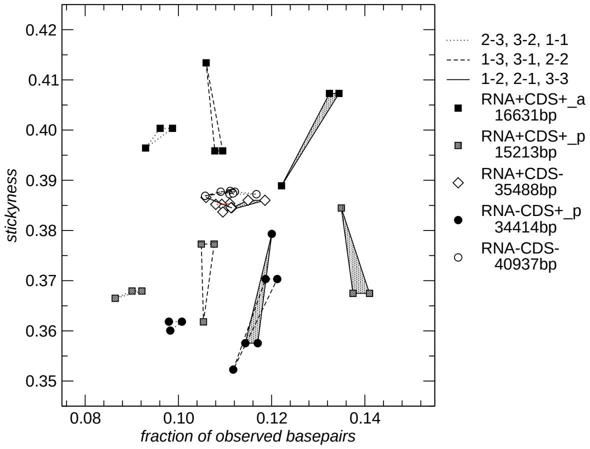


Figure 5: Dual loci in sense (RNA+CDS+_p) and anti-sense (RNA+CDS+_a) orientation, conserved coding sequence without evidence for conservation of RNA structure (RNA-CDS+), and conserved RNA in non-protein-coding locations (RNA+CDA-).

For all types of coding sequences, we find that $\begin{pmatrix}1 & 2 & 3\\ 2 & 1 & 3\end{pmatrix}$ is the most prevalent arrangement of base pairs, followed by $\begin{pmatrix}1 & 2 & 3\\ 3 & 2 & 1\end{pmatrix}$ and $\begin{pmatrix}1 & 2 & 3\\ 1 & 3 & 2\end{pmatrix}$. These differences in abundance may be large, reaching up to 50%. The observed pattern of abundances is surprising, since the prevalence of pairings does not match up with the pairings that are easiest to maintain by compensatory mutations [23]. Furthermore, there is little, if any, correlation between the observed number of base pairs in one of these three base pair arrangements and the stickiness. Hence it appears that stickiness is not a determining factor for the relative location of conserved coding sequence and conserved secondary structure. Although the data suggest that sequence bias in coding sequence and selective constraints on secondary structure influence each other, it remains unclear to what extent the observed biases are specific to flies and whether they are strong enough to be used as a component in prediction tools for dual loci.

4. Discussion

In this contribution we have explored to what extent existing computational methods are capable of surveying loci at which coding sequence and secondary structure is simultaneously under stabilizing selection. With the notable exception of RNAdecoder [5, 6] this topic has received very little attention, probably because of the lack of well-studied examples. Hence we have combined a survey of structured RNA elements and a search for conserved coding sequences with RNAz and RNAcode, respectively. Together, they suggest that loci with dual function are not uncommon in the fruit fly genomes, proposing several hundred candidate loci featuring conserved RNA structures within coding sequence. The manual curation of human transcript annotation in the GENCODE project [25], revealed that a large fraction of protein-coding genes also give rise to non-protein coding isoforms. At the genomic level, thus, superpositions of coding and non-coding functionalities may be common. Indeed, this is also true for the famous SRA1 locus [3], where coding and non-coding functionalities appear to be exerted at least in part by distinct transcripts.

The analysis of the computational surveys reported here exposes severe limitations in the current arsenal of computational methods. In a genome-wide setting, all *de novo* approaches suffer from substantial false discovery rates which at least in part arise from an incomplete understanding of the background model. In the case of superimposed selection pressures, which are the main focus of this contribution, this problem is aggravated: it cannot be expected that simultaneous selection on peptide sequence and on secondary structure leads to independent effects on sequence and structure variation. Although RNAdecoder attempts to include such effects, we observed a false discovery rate that is prohibitive for genome-wide applications. Of course, the present study cannot replace a careful benchmark of alternative comparative RNA gene prediction tools, in particular evofold [26] and SISSIz [27], for the task of predicting conserved RNA secondary structure superimposed on CDS. The absence of a positive set, however, precludes such a benchmark study at this point in time.

An important issue in this context is that all currently available tools are strongly dependent on the quality of the input alignments. As observed in previous studies, see e.g. [28], the sensitivity of computational screens is limited since a substantial fraction of known ncRNAs is not present in the input alignments. In the absence of a sizable set of known dual loci we cannot estimate to what extent this is limiting here. Poorly aligned regions, on the other hand, can also easily lead to false positive predictions: the artefactually increased sequence variation easily generates unusual substitution patterns that are then misclassified.

Comparative genomics approaches, including the methods employed in this contribution, cannot distinguish whether a genomic locus gives rise to different transcripts, one of which is coding, while the other one acts as a structured RNA. Beyond the technical difficulties, therefore, it remains uncertain whether there are many RNAs with dual functions or just many genomic loci with overlapping transcripts of different types.

Acknowledgments

This work was supported in part by grants from the Deutsche Forschungsgemeinschaft (grant No. STA 850/7-1 under the auspices of SPP-1258 "Small Regulatory RNAs in Prokaryotes").

References

[1] D. Ulveling, C. Francastel, F. Hubé. When one is better than two: RNA with dual functions. Biochimie 93 (2011) 633–644.

[2] R. B. Lanz, B. Razani, A. D. Goldberg, B. W. O'Malley. Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA). Proc Natl Acad Sci U S A 99 (2002) 16081–16086.

[3] E. Leygue. Steroid receptor RNA activator (SRA1): unusual bifaceted gene products with suspected relevance to breast cancer. Nuclear Receptor Signaling 5 (2007) e006.

[4] G. Grillo, A. Turi, F. Licciulli, F. Mignone, S. Liuni, S. Banfi, V. A. Gennarino, D. S. Horner, G. Pavesi, E. Picardi, G. Pesole. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 38 (2010) D75–D80.

[5] J. S. Pedersen, I. M. Meyer, R. Forsberg, P. Simmonds, J. Hein. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Res. 32 (2004) 4925–4936.

[6] I. M. Meyer, I. Miklós. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acids Res. 33 (2005) 6338–6348.

[7] M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, J. P. Couso. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 5 (2007) e106.

[8] T. Kondo, Y. Hashimoto, K. Kato, S. Inagaki, S. Hayashi, Y. Kageyama. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9 (2007) 660–665.

[9] A. P. Gultyaev, A. Roussis. Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. Nucleic Acids Res. 35 (2007) 3144–3152.

[10] S. Washietl, I. L. Hofacker, P. F. Stadler. Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA 102 (2005) 2454–2459.

[11] A. R. Gruber, S. Findeiß, S. Washietl, I. L. Hofacker, P. F. Stadler. RNAz 2.0: improved noncoding RNA detection. Pac. Symp. Biocomput. 15 (2010) 69–79.

[12] S. Washietl, S. Findeiß, S. Müller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, N. Goldman. RNAcode: robust prediction of protein coding regions in comparative genomics data. RNA 17 (2011) 578–594.

[13] G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, F. P. Roth. Next generation software for functional trend analysis. Bioinformatics 25 (2009) 3043–3044.

[14] D. R. Rose, J. Hackermüller, S. Washietl, S. Findeiß, K. Reiche, J. Hertel, P. F. Stadler, S. J. Prohaska. Computational RNomics of Drosophilids. BMC Genomics 8 (2007) 406.

[15] S. Misra, M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell, P. Hradecky, Y. Huang, J. S. Kaminker, G. H. Millburn, S. E. Prochnik, C. D. Smith, J. L. Tupy, E. J. Whitfied, L. Bayraktaroglu, B. P. Berman, B. R. Bettencourt, S. E. Celniker, A. D. de Grey, R. A. Drysdale, N. L. Harris, J. Richter, S. Russo, A. J. Schroeder, S. Q. Shu, M. Stapleton, C. Yamada, M. Ashburner, W. M. Gelbart, G. M. Rubin, S. E. Lewis. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. 3 (2002) R0083.

[16] C. A. Hayden, G. Bosco. Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila melanogaster* and other dipteran species. BMC Genomics 9 (2008) 61.

[17] Q. P. Gu, M. A. Beilstein, S. C. Vendeland, A. Lugade, W. Ream, P. D. Whanger. Conserved features of selenocysteine insertion sequence (SECIS) elements in selenoprotein W cDNAs from five species. Gene 193 (1997) 187–196.

[18] M. Hirosawa-Takamori, D. Ossipov, S. V. Novoselov, A. A. Turanov, Y. Zhang, V. N. Gladyshev, A. Krol, G. Vorbrüggen, H. Jäckle. A novel stem loop control element-dependent UGA read-through system without translational selenocysteine incorporation in *Drosophila*. FASEB J. 23 (2009) 107–113.

[19] M. Sato, H. Umeki, R. Saito, A. Kanai, M. Tomita. Computational analysis of stop codon readthrough in *D. melanogaster*. Bioinformatics 19 (2003) 1371–1380.

[20] A. Jenny, O. Hachet, P. Závorszky, A. Cyrklaff, M. D. Weston, D. S. Johnston, M. Erdélyi, A. Ephrussi. A translation-independent role of oskar RNA in early Drosophila oogenesis. Development 133 (2006) 2827–2833.

[21] A. Jenny, O. Hachet, P. Závorszky, A. Cyrklaff, M. D. J. Weston, D. S. Johnston, M. Erdélyi, A. Ephrussi. A translation-independent role of oskar RNA in early Drosophila oogenesis. Development 133 (2006) 2827–2833.

[22] W. M. Fitch. The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid derived messenger-RNA. J. Mol. Evol. 3 (1974) 279–291.

[23] J. Konecny, M. Schniger, I. L. Hofacker, M.-D. Weitze, G. L. Hofacker. Concurrent Neutral Evolution of mRNA Secondary Structures and Encoded Protein. J. Mol. Evol. 50 (2000) 238–242.

[24] A. Filipski, S. J. Prohaska, S. Kumar. Molecular Signatures of

Adaptive Evolution. In M. Page (Ed.), Evolutionary Genomics and Proteomics, chapter 11. Sinauer Associates, Inc., Sunderland (2008) 241–254.

[25] A. J. Coffey, F. Kokocinski, M. S. Calafato, C. E. Scott, P. Palta, E. Drury, C. J. Joyce, E. M. Leproust, J. Harrow, S. Hunt, A. E. Lehesjoki, D. J. Turner, T. J. Hubbard, A. Palotie. The GENCODE exome: sequencing the complete human exome. Eur. J. Hum. Gen. 19 (2011) 827–831.

[26] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, D. Haussler. Classification of Conserved RNA Secondary Structures in the Human Genome. PLoS Comput. Biol. 2 (2006) e33.

[27] T. Gesell, S. Washietl. Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics 9 (2008) 248.

[28] S. Washietl, J. S. Pedersen, J. O. Korbel, A. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, C. Stocsits, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigó, M. Snyder, M. B. Gerstein, A. Reymond, I. L. Hofacker, P. F. Stadler. Structured RNAs in the ENCODE Selected Regions of the Human Genome. Gen. Res. 17 (2007) 852–864.