Navigating the unexplored seascape of pre-miRNA candidates in single-genome approaches

Nuno D. Mendes Steffen Heyne Ana T. Freitas
Marie-France Sagot Rolf Backofen

A Supplementary materials

A.1 Finding the most appropriate vectorial representation

In this work we have tested several different vectorial representations summarising key features of the primary/secondary structure of a given stem-loop. The eight representations we considered differ on the amount of information they represent and thus on their ability to distinguish the structural characteristics of different stem-loops.

The first representation is called TripletS. This representation consists of a vector of normalised counts. To build this representation, a sliding window of length 3 is passed through the structure. At each step, a count position in the vector is incremented. The appropriate position in the vector is mapped considering whether each nucleotide in the window is paired or unpaired in the MFE structure and which base is present on the midpoint. In the end, the counts on each position of the vector are divided by the length of the structure. The vector has thus 32 positions.

The second representation is called TripletB and is built in a way similar to that of TripletS, except that it distinguishes whether the paired nucleotide is in the $5'$ or $3'$ stem arm. In this case the vector has 108 positions.

The third representation is called TripletL and it extends TripletB

by distinguishing nucleotides at the terminal loop from other unpaired nucleotides. This mapping yields a vector with 256 positions.

Three additional representations called QuintupletS, QuintupletB, and QuintupletL are calculated in a way similar to those previously discussed except that they scan the structural information of five consecutive positions yielding vectors with 128, 972, and 4096 positions respectively.

Finally, three representations termed StructureS, StructureB, and StructureL are also similar to the first three representations but the identity of the nucleotide at the midpoint is not considered. These representations, therefore, only include structural information and give rise to vectors with 8, 27, and 64 positions, respectively.

These vectorial representations allow us to capture different types of information about the sequence/structure of our candidates and position them across a hyperplane on a multidimensional space. In order to use the Euclidian distance consistently as a measure of similarity between these vectorial representations it is preferable to represent our candidates using a set of independent and scaled dimensions. A straightforward way to guarantee these conditions is to perform a Principal Components Analysis (PCA) as described in the Methods section.

To determine which of these representations better reflects the results of conventional structural clustering we take the structural clusters obtained using `LocARNA` for 100 samples of 1000 randomly chosen stem-loops from each of the datasets (*D. melanogaster* and *A. gambiae*). As described before, the optimal partition into clusters is done by performing a node evaluation rule for various significance levels (k-levels) where low values for k produce clusters of highly similar structures and increasing values of k allow for increasingly heterogeneous clusters.

After calculating the centroid of each `LocARNA` cluster on the principal components space we can then calculate the rate of correct assignments as the proportion of cluster members that are closer to their cluster centroid rather than the centroid of another cluster. We repeat the procedure on a randomised version of our samples in order to assess the statistical significance of our results against a random background where the identity of each precursor is shuffled, thereby randomising the position of a precursor on the principal components space, but preserving the `LocARNA` cluster it belongs to. The statistical significance of the results is determined by comparing the results obtained for the regular and randomised samples using Welch's two-sample t-test. In each case, the normality of both sets of results (regular

and randomised) is checked using the Kolmogorov-Smirnov test.

In order to compare the results for all the considered vectorial representations across the k-levels ranging from 0.0 to 0.9 we take the symmetric of the logarithm of the p-values of our statistical test. The larger this value the more significant are the results. Tables 1 and 2 show these values for the *A. gambiae* and *D. melanogaster* datasets, respectively.

For low values of k (up to 0.4), in both datasets, TripletL obtains the best results, which means that, for mostly homogeneous clusters, this vectorial representation outperforms all others. If we allow for more heterogeneous clusters (larger values of k), other vectorial representations take the lead but in an inconsistent way, since we obtain different results on both datasets or for different values of k.

For the *D. melanogaster* dataset, the best vectorial representation for $k = 0.5$ becomes TripletB and then TripletS for $k > 0.5$, whereas for the *A. gambiae* dataset, the best vectorial representation changes for $k > 0.7$ to StructureL. In both cases, the transition is to a vectorial representation encoding less information about the hairpins (either structural information in the case of *D. melanogaster* or sequence information for *A. gambiae*), which is consistent with clusters grouping increasingly heterogeneous hairpins.

The TripletL representation emerges as the best choice, since it exhibits the best results for the greater range of k levels and, even though it is outperformed by other representations for the larger values of k, it maintains a very good relative performance.

It is interesting to note that all representations including quintuplets, although encoding more structural information, fail to yield top performances. This might be explained, in part, by the very large number of dimensions and also by the sparsity of the information across the vectors (where most positions will have zeroes) and the implications it has on the principal components analysis. On the representations that exclude sequence information, all except the one distinguishing left/right-hand pairings and stem arm/terminal loop unpaired positions have relatively poor performances, which underlines the importance of including sequence information.

	k									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TripletS	207.9	204.9	191.4	165.4	147.3	108.1	83.9	63.4	56.2	52.5
TripletB	198.2	199.8	191.3	174.1	147.4	125.8	95.7	71.3	58.1	49.7
TripletL	251.9	242.3	224.0	197.0	180.0	157.2	117.8	83.5	65.9	60.7
QuintupletS	140.0	149.2	146.4	134.6	116.5	89.4	55.8	43.0	37.7	32.5
QuintupletB	139.2	140.2	129.9	120.9	113.4	85.9	54.0	31.4	26.8	22.3
QuintupletL	105.3	105.7	97.8	94.0	91.4	80.0	58.8	38.7	34.0	30.3
StructureS	14.3	12.8	12.8	12.4	12.7	10.0	8.4	7.9	7.8	8.3
StructureB	79.5	74.8	72.5	65.1	58.7	44.7	30.0	27.9	27.7	27.7
StructureL	213.4	212.5	204.3	184.7	150.2	121.3	81.3	74.6	73.9	73.3

Table 1: Table showing the -log(p-values) for the statistical significance of the correct assignment rate across all considered vectorial representations and k-levels for the *A. gambiae* dataset

	k									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TripletS	266.1	256.7	244.4	223.1	201.3	182.7	148.2	131.1	114.2	96.3
TripletB	286.6	275.7	258.7	247.5	220.6	193.1	140.4	123.8	99.1	83.8
TripletL	322.1	302.3	279.5	253.9	224.6	184.9	145.7	125.9	101.0	89.2
QuintupletS	229.7	222.2	217.3	204.4	174.5	139.1	113.2	103.2	90.3	78.8
QuintupletB	212.4	204.0	195.4	179.0	162.3	129.6	95.8	72.2	56.1	50.1
QuintupletL	149.8	154.6	148.4	142.3	122.8	101.3	81.5	60.3	41.5	36.7
StructureS	33.1	28.2	26.0	22.8	24.1	22.6	17.6	17.0	15.7	14.0
StructureB	113.3	107.7	102.4	97.8	93.6	79.8	53.6	43.1	38.4	32.1
StructureL	233.9	220.4	211.9	206.4	187.2	162.0	114.0	99.5	95.8	85.4

Table 2: Table showing the -log(p-values) for the statistical significance of the correct assignment rate across all considered vectorial representations and k-levels for the *D. melanogaster* dataset

A.2 Illustration of the distribution of candidates and known pre-miRNAs in a projection of the feature space

Fig. 1 and 2 show the distribution of precursors candidates and annotated pre-miRNAs in a projection of the feature space over the first three principal components of the vectorial representations of the hairpins for the datasets of *A. gambiae* and *D. melanogaster*, respectively.

The distributions on the left-side show the entire dataset, whereas the distributions on the right-side only depict the positions of the annotated pre-miRNAs for each dataset. The graphs show that the known precursors are relatively close to each other in both cases, with respect to the area occupied by all the candidates. It is also easy to see from these distributions that the portion of the feature space where pre-miRNAs are located is also densely populated by other candidates. For this reason, this region cannot be identified by an unsupervised approach.

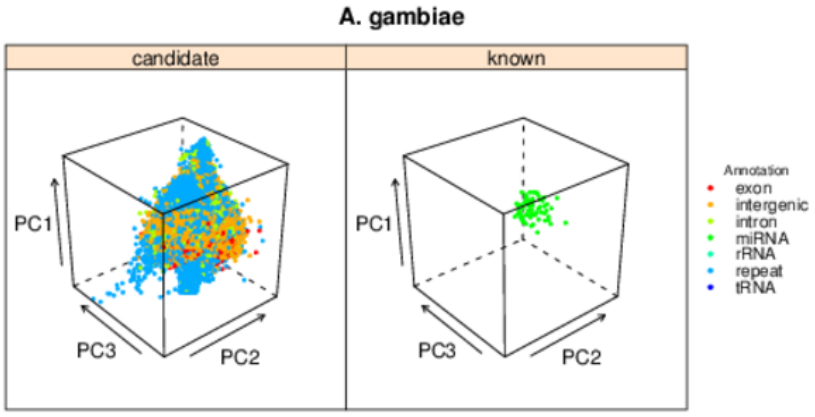


Figure 1: The spatial distribution of candidates and known precursors across the three-dimensional space defined by the first three principal components of the vectorial representation of the hairpins of *A. gambiae*. The annotation information of each hairpin is also indicated.

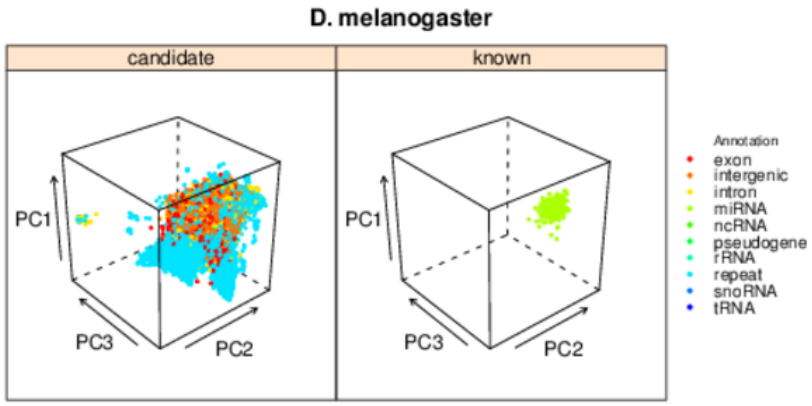
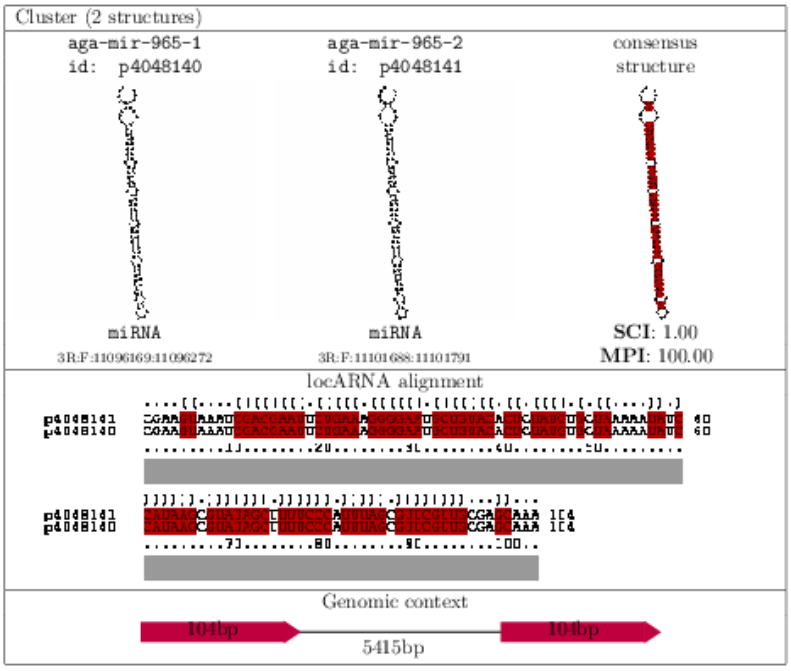


Figure 2: The spatial distribution of candidates and known precursors across the three-dimensional space defined by the first three principal components of the vectorial representation of the hairpins of *D. melanogaster*. The annotation information of each hairpin is also indicated.

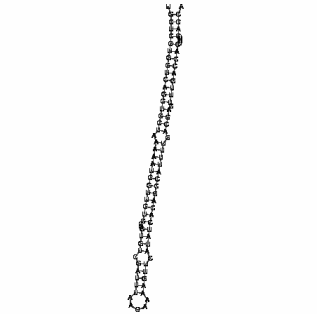
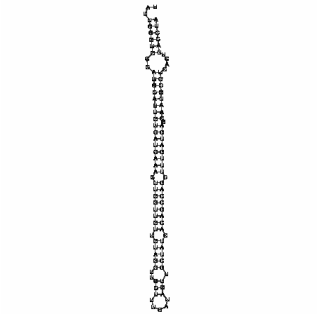
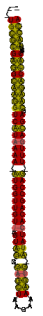
A.3 Genomic clusters identified in *A. gambiae* and *D. melanogaster*

The genomic clusters listed below are a selection of clusters determined using each of the two approaches described in the paper. The first approach enumerates genomic cluster found within structural clusters including at least one known precursor, and the second approach enumerates genomic clusters found within the top-scoring leaves of the similarity tree in terms of SCI (structure conservation index) which were not listed in the first approach. In each list, the genomic clusters are ordered by their MPI (mean pairwise identity), which, for genomic cluster containing only two stem-loops, corresponds to the degree of identity of the two primary sequences. The selection was obtained by considering a visual inspection of the secondary structure, the annotation information available for each candidate, the genomic distance between each member of the cluster and the structural conservation index.

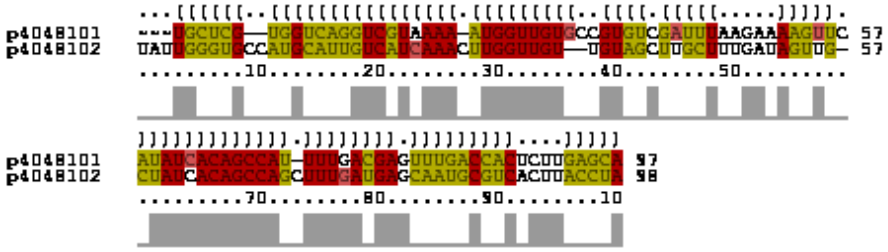
A.3.1 Genomic clusters including at least one known precursor in *A. gambiae*



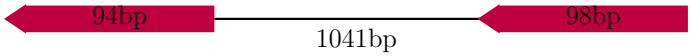
Cluster (2 structures)

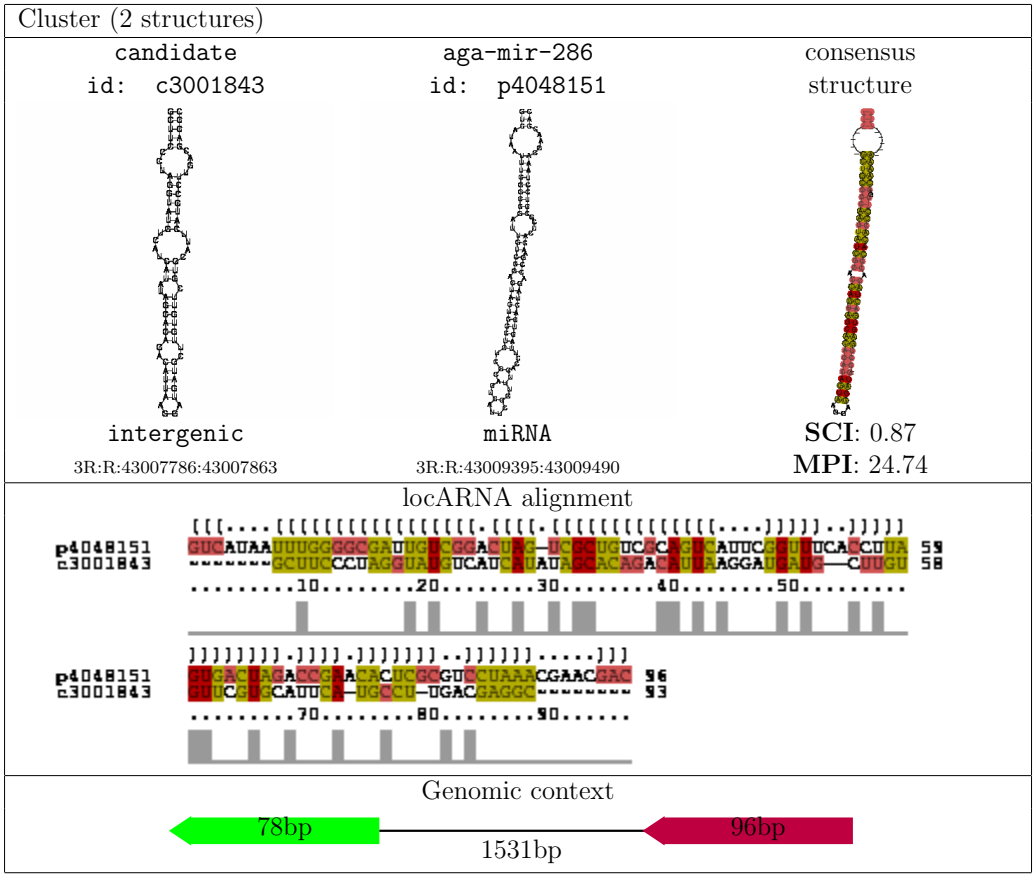
aga-mir-13b	aga-mir-2-2	consensus
id: p4048101	id: p4048102	structure
		
miRNA	miRNA	**SCI**: 1.12
2L:R:37757564:37757657	2L:R:37758699:37758796	**MPI**: 50.50

locARNA alignment

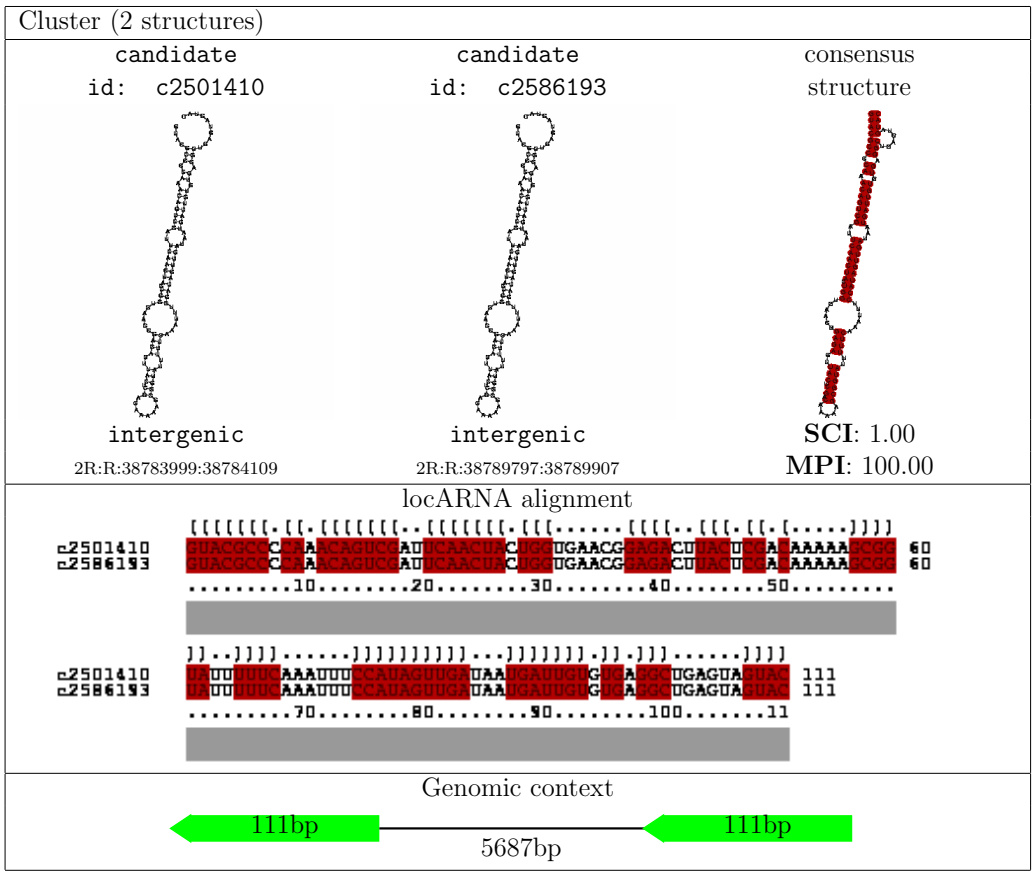


Genomic context




Cluster (2 structures)
candidate
id: c3001843
aga-mir-286
id: p4048151
consensus
structure
intergenic
3R:R:43007786:43007863
miRNA
3R:R:43009395:43009490
SCI: 0.87
MPI: 24.74
locARNA alignment
p4048151
c3001843
78bp
96bp
1531bp
Genomic context

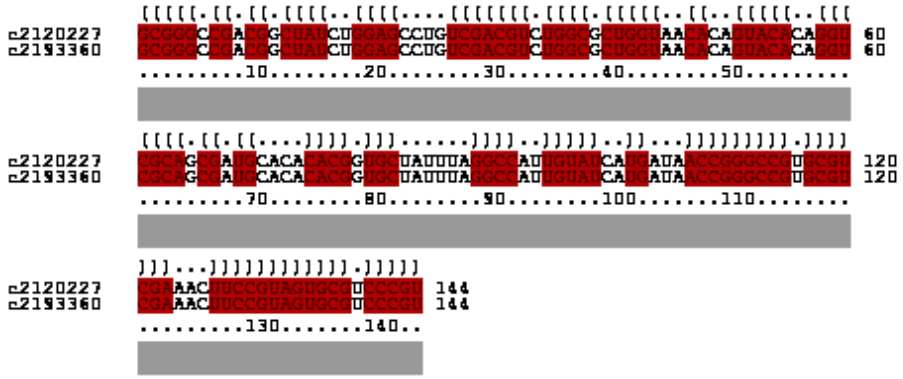
A.3.2 Genomic clusters in top-scoring structural clusters in *A. gambiae*



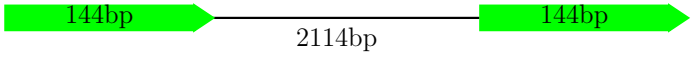
Cluster (2 structures)

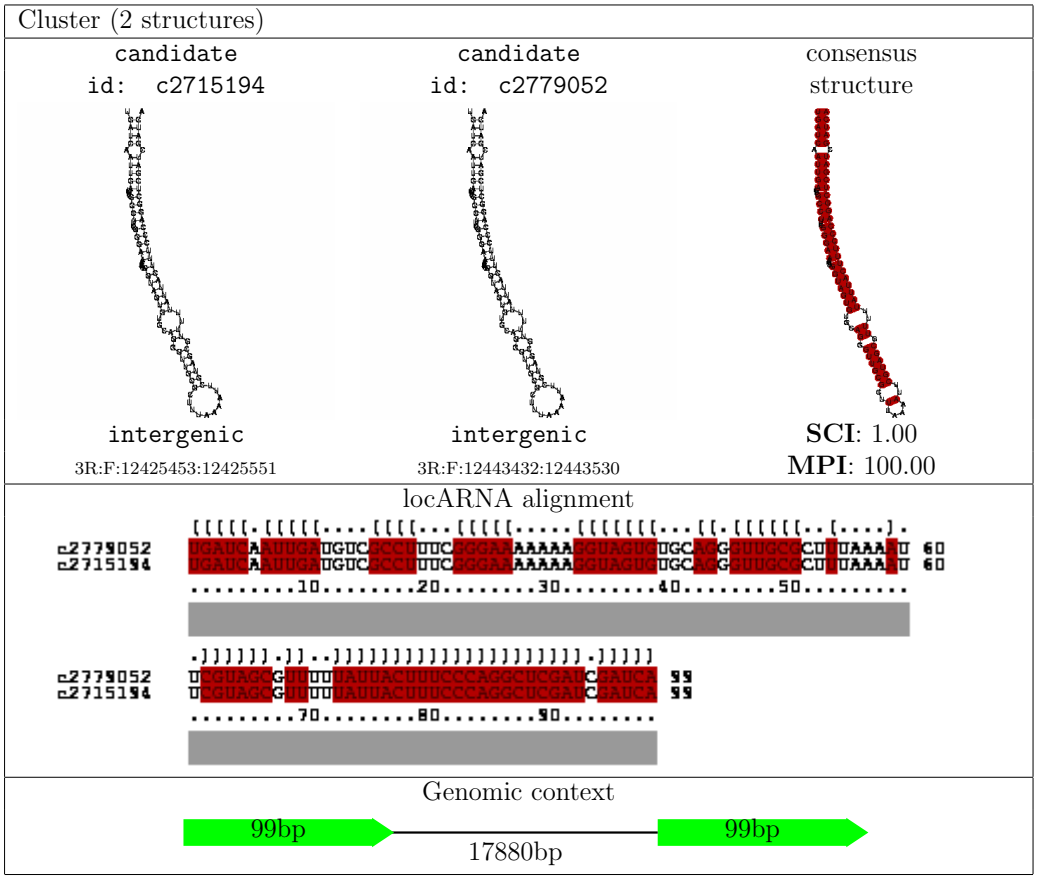
candidate id: c2120227	candidate id: c2193360	consensus structure
		
intergenic	intergenic	**SCI**: 1.00
2R:F:13087682:13087825	2R:F:13089940:13090083	**MPI**: 100.00

locARNA alignment

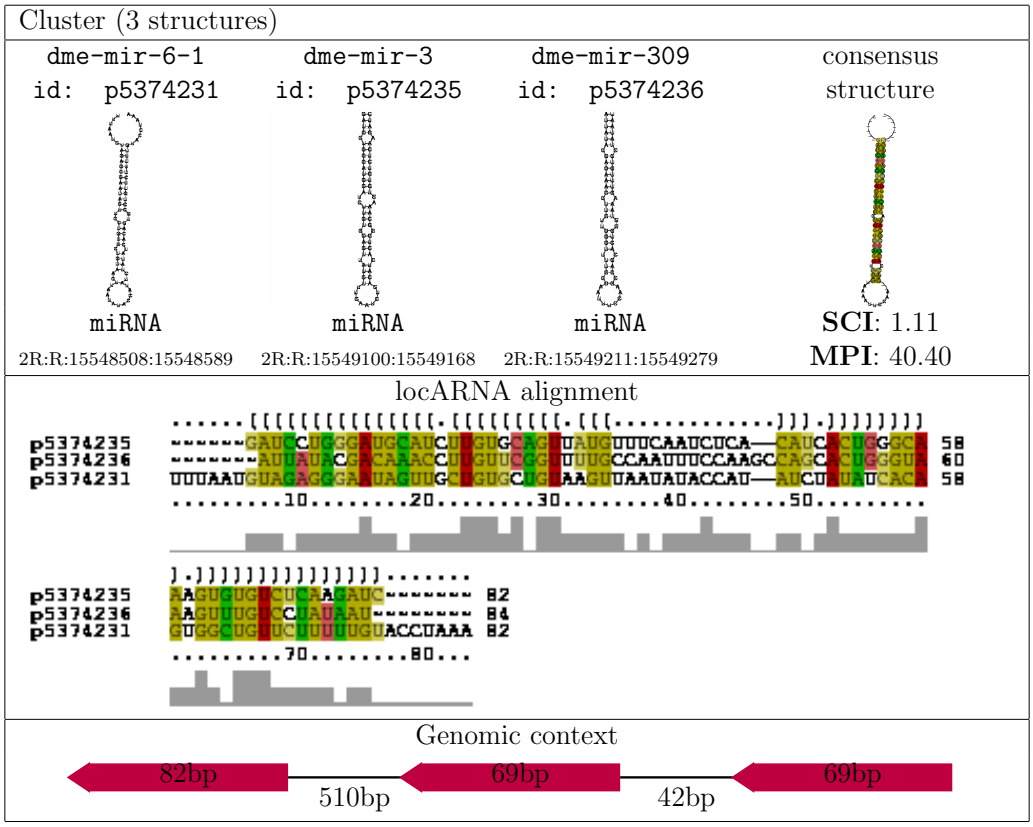


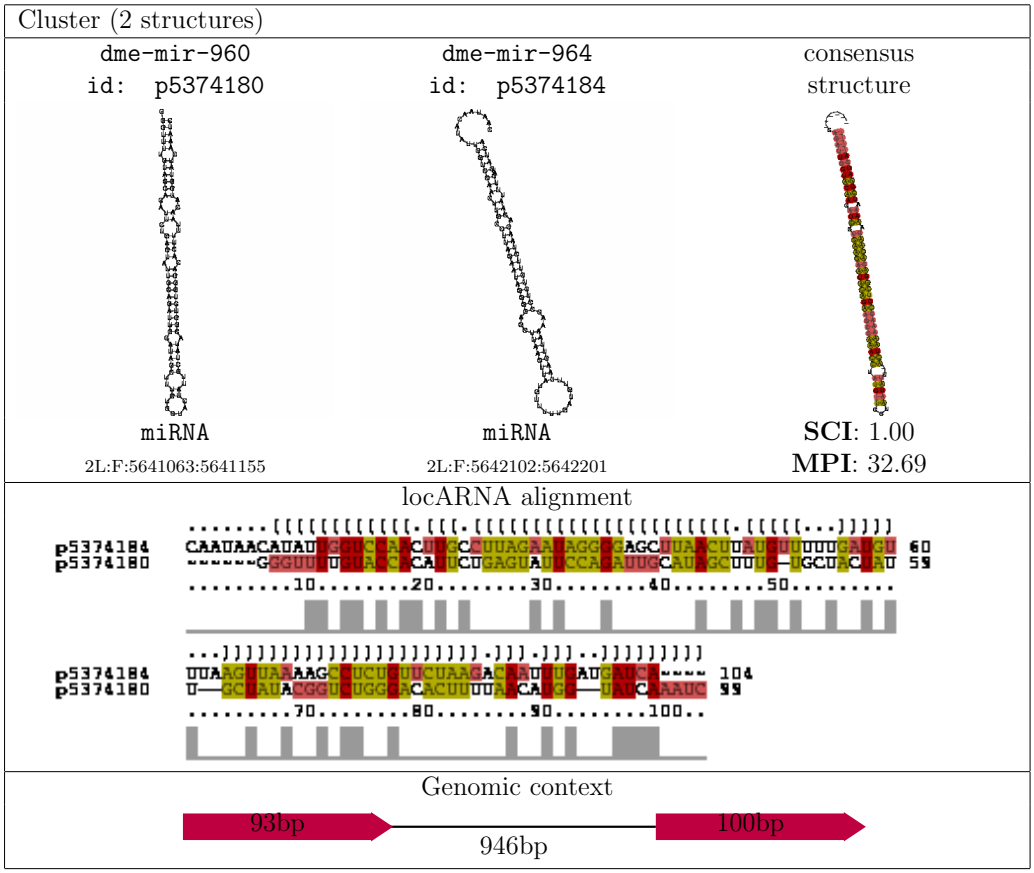
Genomic context

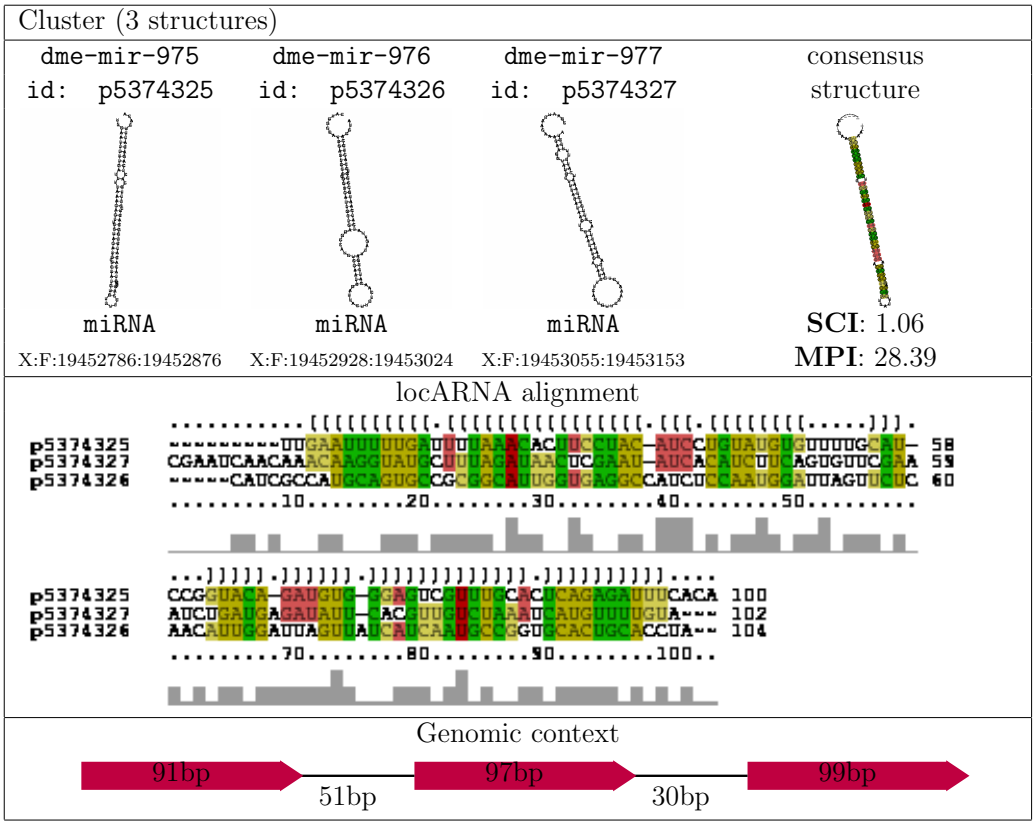



Cluster (2 structures)
candidate
id: c2715194
candidate
id: c2779052
consensus
structure
intergenic
3R:F:12425453:12425551
intergenic
3R:F:12443432:12443530
SCI: 1.00
MPI: 100.00
locARNA alignment
c2779052
c2715194
99bp
17880bp
99bp

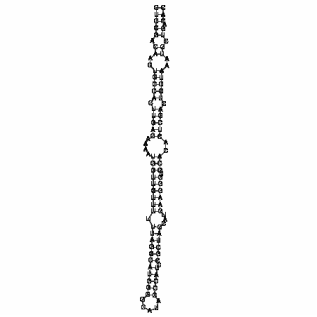
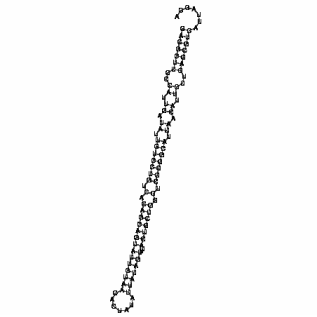
A.3.3 Genomic clusters including at least one known precursor in *D. melanogaster*



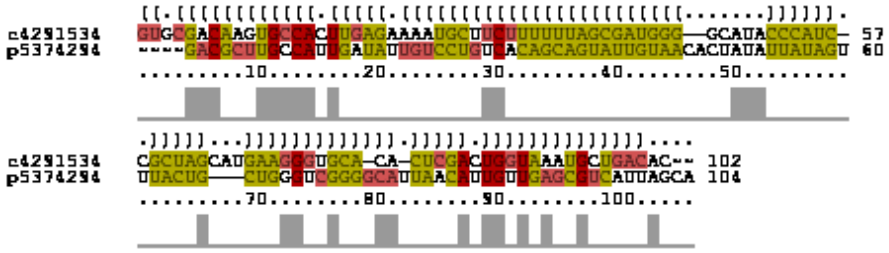

Cluster (2 structures)
dme-mir-960
id: p5374180
dme-mir-964
id: p5374184
consensus
structure
miRNA
2L:F:5641063:5641155
miRNA
2L:F:5642102:5642201
SCI: 1.00
MPI: 32.69
locARNA alignment
Genomic context
93bp
946bp
100bp


Cluster (3 structures)
dme-mir-975
id: p5374325
dme-mir-976
id: p5374326
dme-mir-977
id: p5374327
consensus
structure
miRNA
miRNA
miRNA
X:F:19452786:19452876
X:F:19452928:19453024
X:F:19453055:19453153
SCI: 1.06
MPI: 28.39
locARNA alignment
p5374325
p5374327
p5374326
Genomic context
91bp
51bp
97bp
30bp
99bp

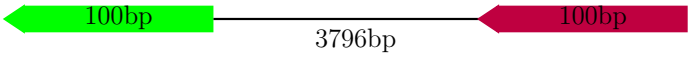
Cluster (2 structures)

candidate id: c4291534	dme-mir-1000 id: p5374294	consensus structure
		
intergenic	miRNA	**SCI**: 1.02
3R:R:21410685:21410784	3R:R:21414581:21414680	**MPI**: 25.23

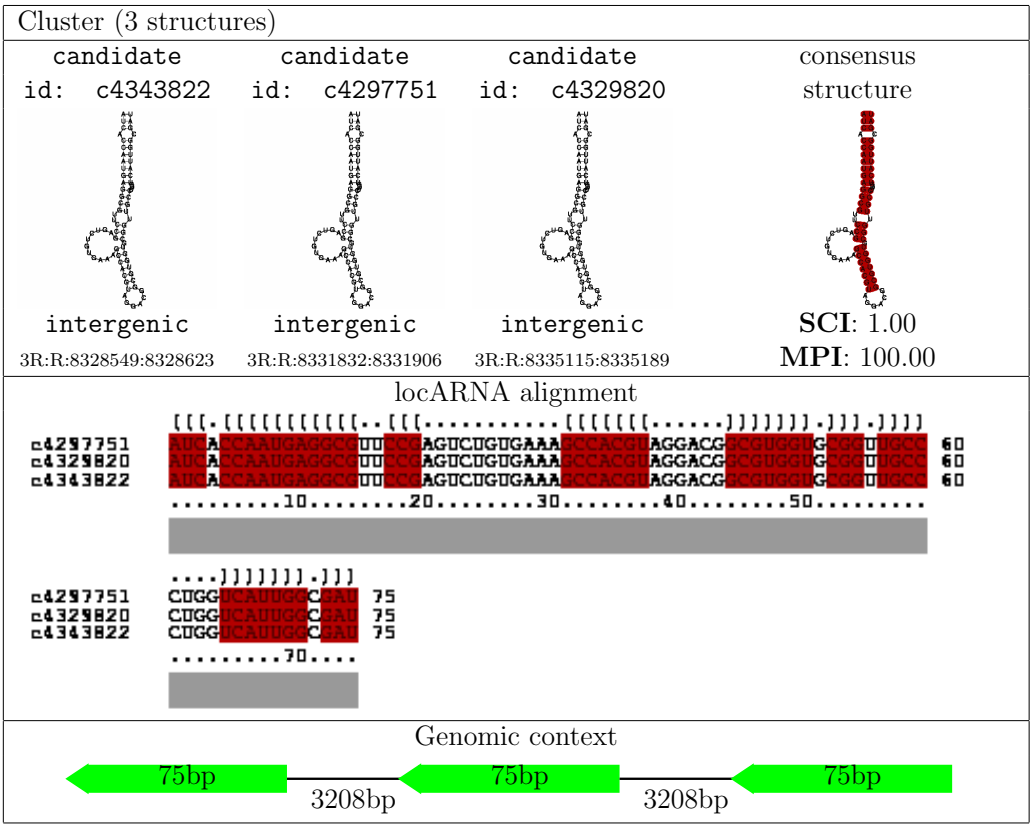
locARNA alignment



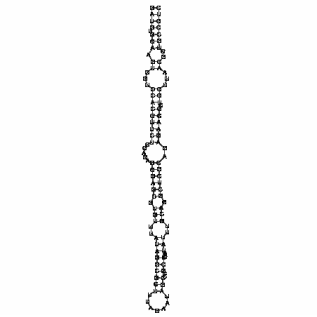
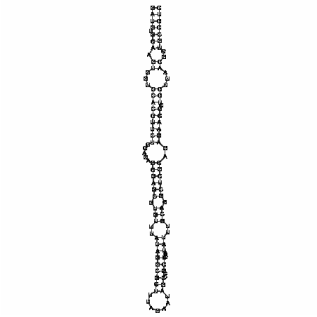
Genomic context



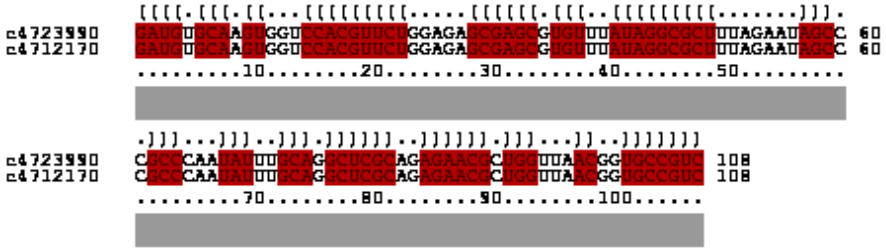
A.3.4 Genomic clusters in top-scoring structural clusters in *D. melanogaster*



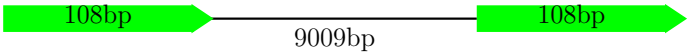
Cluster (2 structures)

candidate id: c4723990	candidate id: c4712170	consensus structure
		
intergenic	intergenic	**SCI**: 1.00
2L:F:20448117:20448224	2L:F:20457234:20457341	**MPI**: 100.00

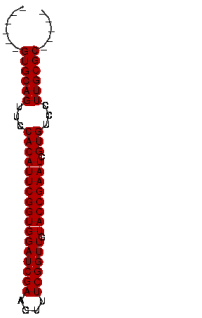
locARNA alignment



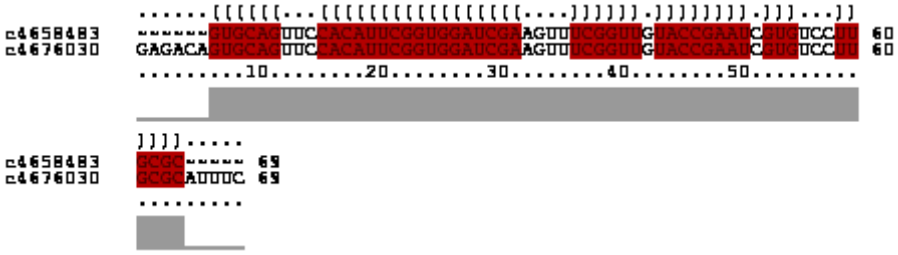
Genomic context



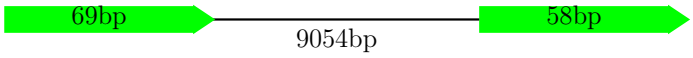
Cluster (2 structures)

candidate id: c4676030	candidate id: c4658483	consensus structure
		
intergenic	intergenic	**SCI**: 1.00
2L:F:20448380:20448448	2L:F:20457503:20457560	**MPI**: 84.06

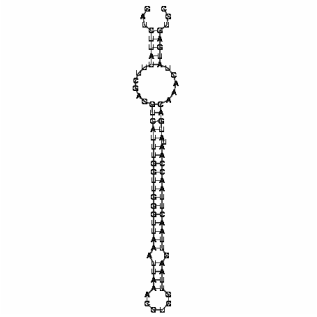
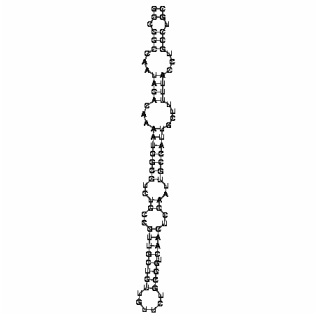
locARNA alignment



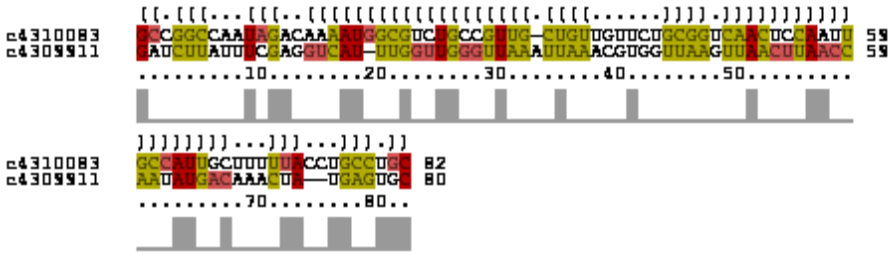
Genomic context



Cluster (2 structures)

candidate id: c4309911	candidate id: c4310083	consensus structure
		
intergenic	intergenic	**SCI**: 1.09
3R:R:27591293:27591372	3R:R:27602897:27602978	**MPI**: 30.12

locARNA alignment



Genomic context