MINING *roX1* RNA IN *Drosophila* GENOMES USING COVARIANCE MODELS

K. Byron,* ,**,***** M.C.-Cervantes,*,*** J.T.L. Wang,*,**,*** W.C. Lin,**** and Y. Park*****

Abstract

Evolutionarily conserved functional domains of non-coding RNA on chromosome X (*roX1*) have been identified in eight *Drosophila* species in a prior study. Among our findings, three GUUNUACG repeats were localized in the 3′-end of the predicted *roX1* RNAs for these *Drosophila* species. In this study, we use a covariance model (CM) to search for the characteristic features of *roX1* functional domains as a way to predict new examples of these structured RNAs in other *Drosophila* species, as sequencing data become available. We scan whole genomes of *Drosophila* and identify search results in available "region" terms, i.e., "chromosome" or "scaffold", depending on the annotation status of the particular species being surveyed. With known *roX1* examples produced through our prior studies for support in comparative analysis, we hypothesize that it is possible to predict novel *roX1* functional domains accurately from sequence information alone. Annotating *roX1* on a genomic scale provides insight into evolutionary processes and phylogenetic relationships among the analyzed species. Based on our results, we confirmed that a CM search is effective in mining *roX1* RNA genes and, that due to its inherent flexibility, this mining approach will likely prove successful for similar endeavours in various other organisms.

Key Words

Non-coding RNA, covariance model, secondary structure, *Drosophila*, comparative genomics

1. Introduction

Non-coding RNAs (ncRNAs) are functional RNA transcripts that are not translated into protein (i.e., they are not messenger RNAs). Recent research has shown that ncRNAs perform a wide range of functions in the cell [1–4]. RNA on X (*roX1*) plays an essential role in equalizing the level of transcription on the X chromosome in *Drosophila* males (XY) to that of females (XX) [5]. Experiments have confirmed that *roX1* RNA exists in eight *Drosophila* species [6–8]. This leads to the hypothesis that there exists secondary structure conservation of the *roX1* gene among other *Drosophila* species [7, 8] for which specific *roX1* genomic coordinates are as yet unknown. Recent advances in the research of genomes from 12 *Drosophila* species [9] might contribute to support this hypothesis.

The covariance model (CM) method Infernal, used in the prediction of functional domain conservation in ncRNAs, is considered by experts to be one of the most accurate general tools [10]. A CM is a statistical representation or profile of a family of related RNAs that share a common consensus secondary structure [11]. The Infernal software package [12–14] contains a utility, cmbuild, for creating a CM from an alignment of sequences in the Stockholm format, and a utility, cmsearch, to search for sequences that are similar to the model. The cmsearch process is computationally expensive when a single-processor approach is used. However, by utilizing a parallel processing approach, search results can be obtained in efficient time frames.

We began by demonstrating the capability of using a CM in a genome-scale homology search. We obtained *roX1* sequences from eight species of *Drosophila* for which experimental evidence of *roX1* has been found, namely: *D. ananassae*, *D. erecta*, *D. melanogaster*, *D. mojavensis*, *D. pseudoobscura*, *D. simulans*, *D. virilis* and *D. yakuba* [7]. Subsequently, we focused on locating evidence of *roX1* sequences on the complete genomes of these eight species using a CM derived from our *roX1* sample sequences. We expected this search to be successful and it was in 6 of the 8 species. This was critical to demonstrate as a "proof of concept" that our method would find that which we know exist from empirical data. Subsequently, we used the same CM to search for evidence of these conserved structures in the complete genomes of the four remaining sequenced *Drosophila* species for which there are no transcript-derived *roX1* sequences. These four species are *D. grimshawi*, *D. persimilis*, *D. sechellia* and *D. willistoni*. Such a comparative genomics approach has been successful in the unicellular organism *Saccharomyces cerevisiae* [15].

* Bioinformatics Program, New Jersey Institute of Technology, Newark, NJ 07102, USA; e-mail: byron@njit.edu

** Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA; e-mail: wangj@njit.edu

*** Department of Biological Sciences, Rutgers University, Newark, NJ 07102, USA; e-mail: miguelcc@andromeda.rutgers.edu

**** Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, Republic of China; e-mail: wenlin@ibms.sinica.edu.tw

***** Department of Cell Biology and Molecular Medicine, University of Medicine and Dentistry of New Jersey – New Jersey Medical School, Newark, NJ 07103, USA; e-mail: parky1@umdnj.edu

(paper no. 210-1012)

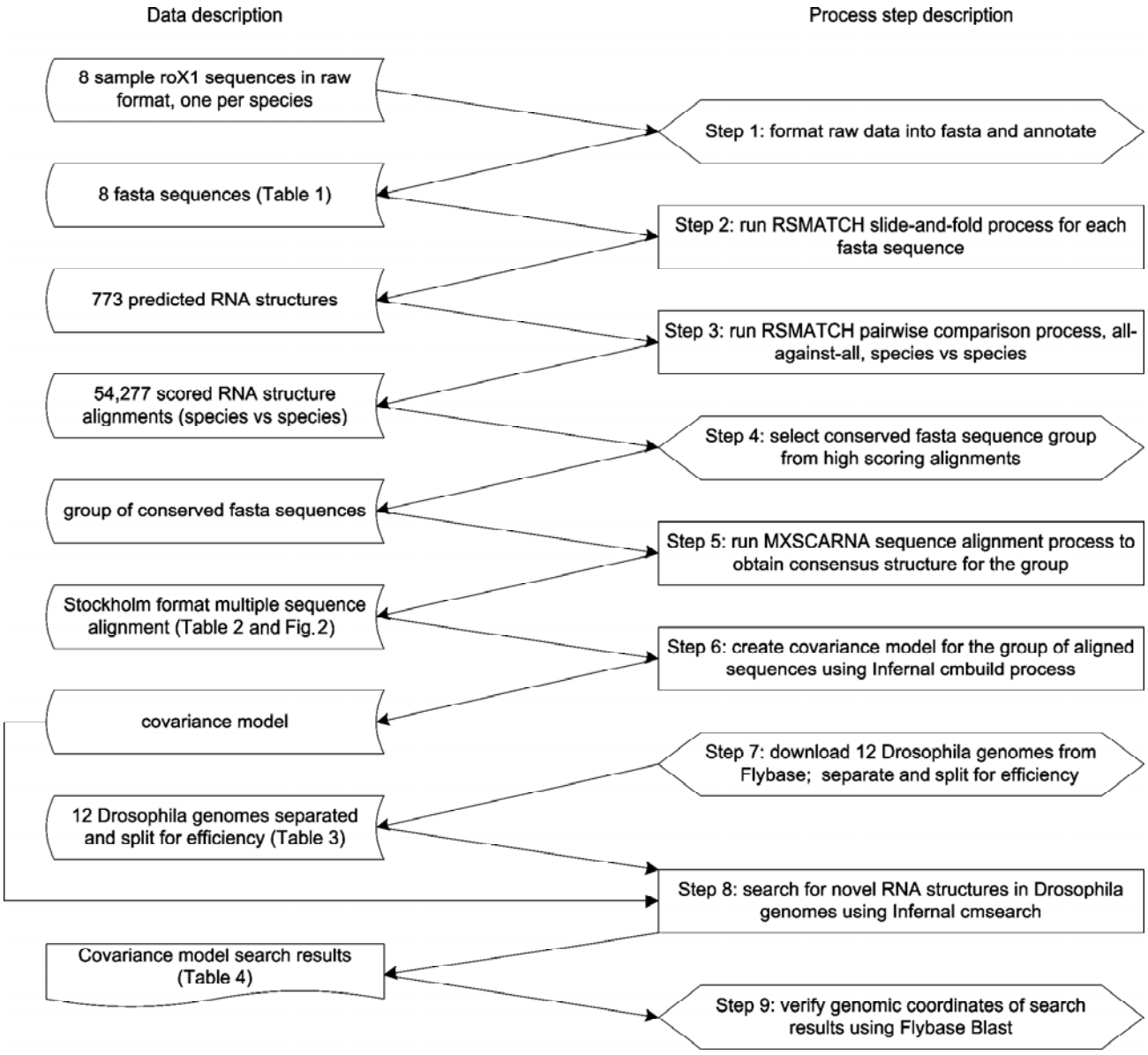

Figure 1. *Drosophila roX1* data mining process utilizing covariance model methodology. See methods section for description.

Using this approach, our results show strong evidence of the presence of *roX1* functional domains in the genome of *D. sechellia*. We believe this finding to be novel and significant in ongoing genomic studies of *Drosophila* and related taxonomic groups. This bioinformatics study lays the groundwork for future accurate and efficient CM searches where the model can be fine tuned as needed to vary the search for specific objectives.

2. Materials and Methods

2.1 *Drosophila* Sequence and Structure Prediction

Figure 1 summarizes our data mining approach and illustrates the CM methodology utilized. We obtained eight *roX1* RNA sequences experimentally (*i.e.*, non-predicted) from *Drosophila* species (*i.e.*, *D. ananassae*, *D. erecta*, *D. melanogaster*, *D. mojavensis*, *D. pseudoobscura*, *D. simulans*, *D. virilis* and *D. yakuba*) [7] (Table 1). In all cases, the sequences were obtained in raw format and were set up in standard FASTA format. Sequences were assigned arbitrary names as follows: in columns 1, 2 and 3, "yp1"; and in columns 4 and 5, a sequential 2-digit number. Starting and ending positions for all sequences were described in FASTA notation as "start:end" where "start" represents the first numeric position and "end" represents the last numeric position. For each of the original sequences, "start" had the value of 1 and "end" had the value of the length of the original sequence. The formatting of the starting and ending positions for each sequence was made compatible with RSmatch [16] so that when subsequences were extracted, the original FASTA annotation was preserved and additional subsequence position information was inserted for sequence tracking purposes.

As an illustration, the annotation for one FASTA sequence, named yp101, which was input to the RSmatch slide-and-fold process is shown here: ">yp101 (1:3493) droana rox1". Note that the length of this *roX1* gene

Table 1
Drosophila roX1 Sequences Used in This Study

Species	Length	Secondary Structures*	Similarities◇	FlyBase Region	Region Coordinates§
D. ananassae	3,493	77	5,666	scaffold_13117	695557 – 693154 693089 – 692300 692143 – 692065 692247 – 692215
D. erecta	3,462	98	6,785	scaffold_4690	1139892 – 1137083 1140318 – 1139928 1137036 – 1136857
D. melanogaster	3,468	106	7,249	X	3755987 – 3754338 3754043 – 3753143 3756379 – 3756024 3754304 – 3754082 3753108 – 3752929
D. mojavensis	3,768	99	6,864	scaffold_6328	3900419 – 3899115 3901467 – 3900566 3901937 – 3901499 3902390 – 3902000 3898736 – 3898624 3898929 – 3898874 3898845 – 3898810 3947396 – 3947375 246881 – 246900 700541 – 700522
D. pseudoobscura	3,469	92	6,385	XL_group 1e	6901185 – 6898994 6898915 – 6897910 6897801 – 6897717 1352025 – 1352045 10880750 – 10880730 476212 – 476239 2910133 – 2910114
D. simulans	3,439	101	7,049	X	2761962 – 2759151 2762379 – 2761996 2759122 – 2758943 9903425 – 9903446
D. virilis	3,623	97	6,854	scaffold_13042	4639617 – 4638455 4637622 – 4636608 4638333 – 4637894 4636532 – 4636064 4637736 – 4637672 4637835 – 4637787 4636035 – 4635995 4638396 – 4638367
D. yakuba	3,433	103	7,425	X	4658396 – 4661828 3710814 – 3710795
Total		773	54,277		

* Secondary structures were predicted by the Vienna RNA package.
◇ Similarities with other *Drosophila* species were computed by RSmatch.
§ Region and region coordinate information were obtained from FlyBase (http://www.flybase.org).

Table 2
Illustration of *Drosophila* Sequences Used in the Creation of a CM

```
# STOCKHOLM 1.0
dme_rox1:3102-3165 GGUUCGUGUUUCGGAAAACGCAUUAAAAGGCGUAAUUUUAAAUCGUUUUCCGAAAUGGGA
dsi_rox1:3079-3142 GGUUCGUGUUUCGGAAAACGCUCUAAAAGGCGCAAUUUUAAAUCGUUUUCCGAAAUGGGA
dya_rox1:3080-3143 GGUUCGUGUUUCGGAAAACGCACUAAAAGGCGUAGUUUUGAAUCGUUUUCCGAAAUGGGA
#=GC SS_cons       <<<<<.<<<<<<<<<<<<<<.<<<<<<<..<...>>>>>>>>.>>>>>>>>>>>>>>.>>
dme_rox1:3102-3165 AUCA
dsi_rox1:3079-3142 AUCA
dya_rox1:3080-3143 AUCA
#=GC SS_cons       >>>.
//
```

The alignment is shown in Stockholm format. The numeric range following the species code represents the portion of the *roX1* gene from which the sequence was extracted. The CM subsequently created was used in this study to search for other novel *roX1* conserved domains.

sequence is 3,493 nucleotides (nt). RSmatch extracted subsequences from the original yp101 sequence and produced properly annotated FASTA format sequences such as this: ">yp101:51–150 (1:3493) droana rox1". Note that this annotation clearly represents a 100 nt sequence which was extracted from positions 51 through 150 of the original yp101 sequence. All of the original FASTA annotation information was retained. Providing position information in the annotation of the extracted subsequence is a critical function performed by RSmatch.

In a similar manner, all eight *Drosophila roX1* sequences evaluated for this work were annotated for compatibility with RSmatch, thus preserving subsequence positions. RSmatch slide-and-fold process was run with the following parameters: sequence size = 100; overlap size = 50; minimum free energy = 0. We prepared RNA structures by the "slide and fold" method, as previously described [16]. Briefly, for each sequence, we took 100 nt subsequences at every 50 nt position from 5′ to 3′ resulting in consecutive subsequences overlapping with one another on a 50-nt segment. Subsequences shorter than 100 nt, *e.g.*, at the 5′ or 3′ ends, were also kept. We then folded all of the subsequences using the RNAsubopt function in the Vienna RNA package [17] with the setting "–e 0". With this setting, multiple structures with the same minimum energy can be generated. Using this method, we obtained 773 structures from the eight *Drosophila roX1* sequences.

2.2 RNA Structure Comparison

Pairwise comparisons of all RNA structures were carried out by RSmatch [16], with the "dsearch" function and default scoring matrices for single-stranded (ss) and double-stranded (ds) regions. Specifically, nucleotide match scores were 1 and 3 in ss and ds regions, respectively; and mismatch scores were –1 and 1, in ss and ds regions, respectively. The gap penalty was –6 for both ss and ds regions. This scoring scheme in effect gave more weight on matches in ds regions than those in ss regions. We extracted three unduplicated FASTA sequences from high-scoring pairwise alignments.

2.3 CM Creation

The Mxscarna [18] package was used to align sequences for the CM used in the study. The resulting alignment was rendered in the Stockholm format with predicted structure annotation (Table 2). This alignment was input to the Infernal package utility cmbuild to create a CM. Figure 2 shows the consensus secondary structure of the sequence alignment used to create our CM [19].

2.4 *Drosophila* Sequence Database

The CM search utility "cmsearch" was run against a database of *Drosophila* FASTA sequences. The genomes from 12 *Drosophila* species (*i.e.*, *D. ananassae*, *D. erecta*, *D. grimshawi*, *D. melanogaster*, *D. mojavensis*, *D. persimilis*, *D. pseudoobscura*, *D. sechellia*, *D. simulans*, *D. virilis*, *D. willistoni* and *D. yakuba*) were downloaded from Indiana University's FlyBase database (Table 3). Most *Drosophila* genomes have not been annotated into clearly defined chromosomes in FlyBase. As organismal research and sequencing efforts continue, we expect that the genomes of these 12 species of *Drosophila* will be fully annotated regarding specific chromosome identification.

2.5 CM Search

The Infernal package (version 1.0) utility "cmsearch" was used to locate structures in genomes with high degree of probability of matching the constructed CM [12]. To improve computational efficiency, large FASTA sequences were split into smaller, overlapping subsequences to facilitate independent parallel searching without negatively impacting results.

Figure 3 illustrates predicted secondary structures identified by the CM search as having high similarity relative to the profile that the CM represents. Out of the seven structures shown in Fig. 3, only the coordinates for *D. sechellia*'s *roX1* are not available for comparison. Given the structural similarity and high score result of the CM search, we propose that the *D. sechellia* sequence

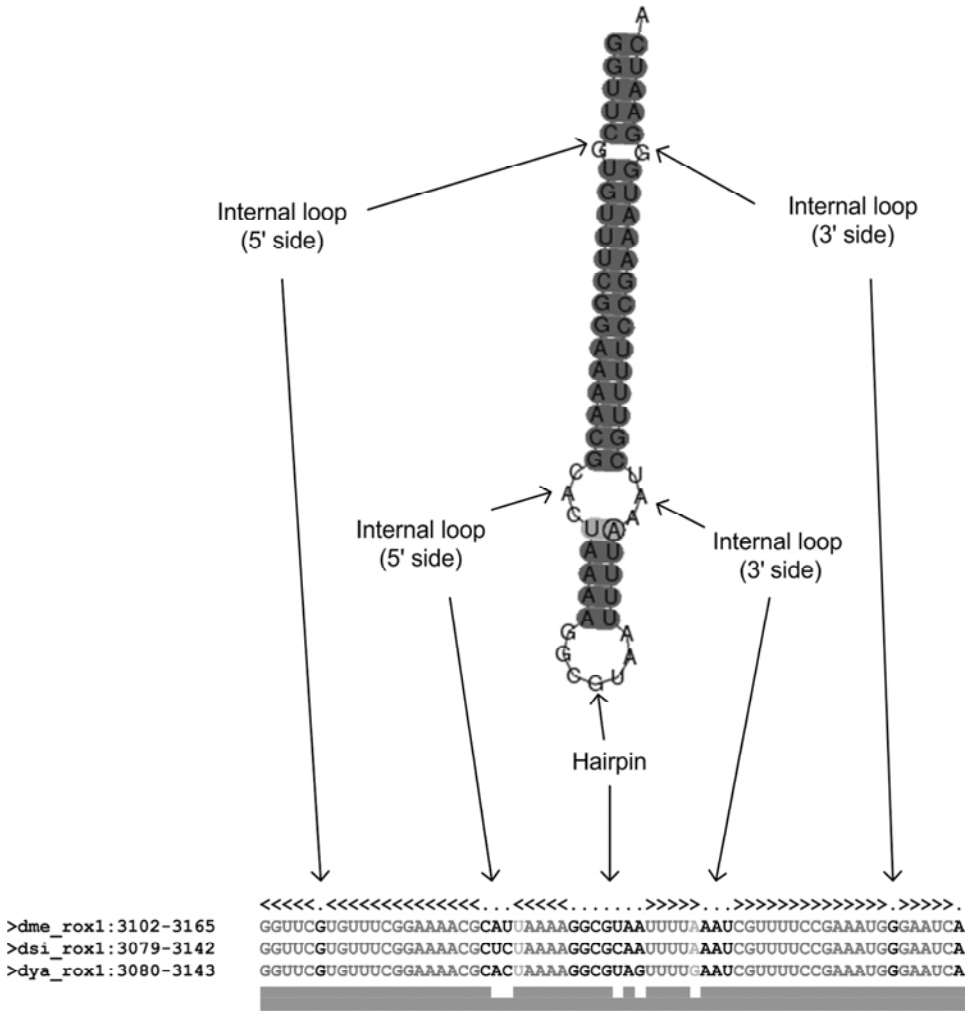

Figure 2. Illustration of the substructures of the RNA secondary structure representing the consensus structure of the alignment of three *Drosophila roX1* sequences from species *D. melanogaster*, *D. simulans* and *D. yakuba.*

discovered is likely to represent *roX1* functional domain characteristics.

3. Results

3.1 Mining *roX1* RNA Where the Presence of *roX1* is Known

Our purpose was to identify functional structure elements in genomes of *Drosophila* species in which the presence of *roX1* has been experimentally demonstrated. To an extent, our strategy was similar to that recently proven successful by Khaladkar and others [16, 20–22]. First, we obtained eight sequences of *roX1* RNA transcripts (Table 1). We then used a "slide and fold" method to construct RNA structures, as described in Methods. In this approach, subsequences 100-nt long or shorter were folded according to their thermodynamic properties using the Vienna RNA package [17]. Adjacent subsequences were overlapped by 50 nt. With this method, one can predict RNA structures accurately and efficiently for two reasons: (1) prediction of small ribonucleotide structures is more accurate and efficient than for large ones; and (2) structures with a size smaller than 50 nt were folded twice as subsequences of two different larger structures, further increasing the probability of obtaining accurate RNA structure predictions. We also used a setting in the Vienna package that yielded multiple RNA structure predictions with the same minimum free energy for a given sequence to further improve folding accuracy. This step resulted in 773 predicted RNA structures.

We then carried out species versus species pairwise comparisons using all 773 predicted RNA structures. To make our approach computationally efficient, we ran each alignment as a process independent of all others on a high performance computing (HPC) cluster at the New Jersey Institute of Technology, leveraging emerging grid computing capabilities [23]. This HPC system, a Sun Microsystems Discovery cluster, has 112 AMD Opteron dual-core Linux nodes with 2 GB of RAM per node. The operating system used was Red Hat Enterprise Linux AS release 4 Update 8. In this manner, approximately 520,000 pairwise alignments were completed in less than 5 min, whereas we would have expected this process to take several

Table 3
Description of 12 *Drosophila* Genomes Downloaded from FlyBase [29]

Species	Release #	Release date	Nucleotides	Original Sequences	Region Annotations	Original Files	Modified Sequences	Modified Files
D.ananassae	1.3	24Jul08	230,993,012	13,749	Sc	1	13,809	128
D.erecta	1.3	24Jul08	152,712,140	5,124	Sc	1	5,183	81
D.grimshawi	1.3	24Jul08	200,467,819	17,440	Sc	1	17,502	121
D.melanogaster	5.18	16May09	130,430,583	7	Ch	7	69	70
D.mojavensis	1.3	24Jul08	193,826,310	6,841	Sc	1	6,916	106
D.persimilis	1.3	24Jul08	188,374,079	12,838	Sc	1	12,874	121
D.pseudoobscura	2.4	19May09	152,738,921	4,896	Ch, Sc	1	4,952	87
D.sechellia	1.3	24Jul08	166,577,145	14,730	Sc	1	14,768	108
D.simulans	1.3	24Jul08	137,828,247	10,005	Ch, Sc	1	10,057	73
D.virilis	1.2	24Jul08	206,026,697	13,530	Sc	1	13,600	121
D. willistoni	1.3	24Jul08	235,516,348	14,838	Sc	1	14,907	155
D.yakuba	1.3	24Jul08	165,693,946	8,122	Ch, Sc	1	8,179	88
Total			2,161,185,247	122,120		18	122,816	1,259

To improve computational efficiency, files containing more than 2 Mbp were separated into smaller files by distributing whole sequences. In a case where a sequence was greater than 2 Mbp, the sequence was split into multiple overlapping segments. Overlap size is 5 K bases. Annotation: Ch=chromosome; Sc = Scaffold

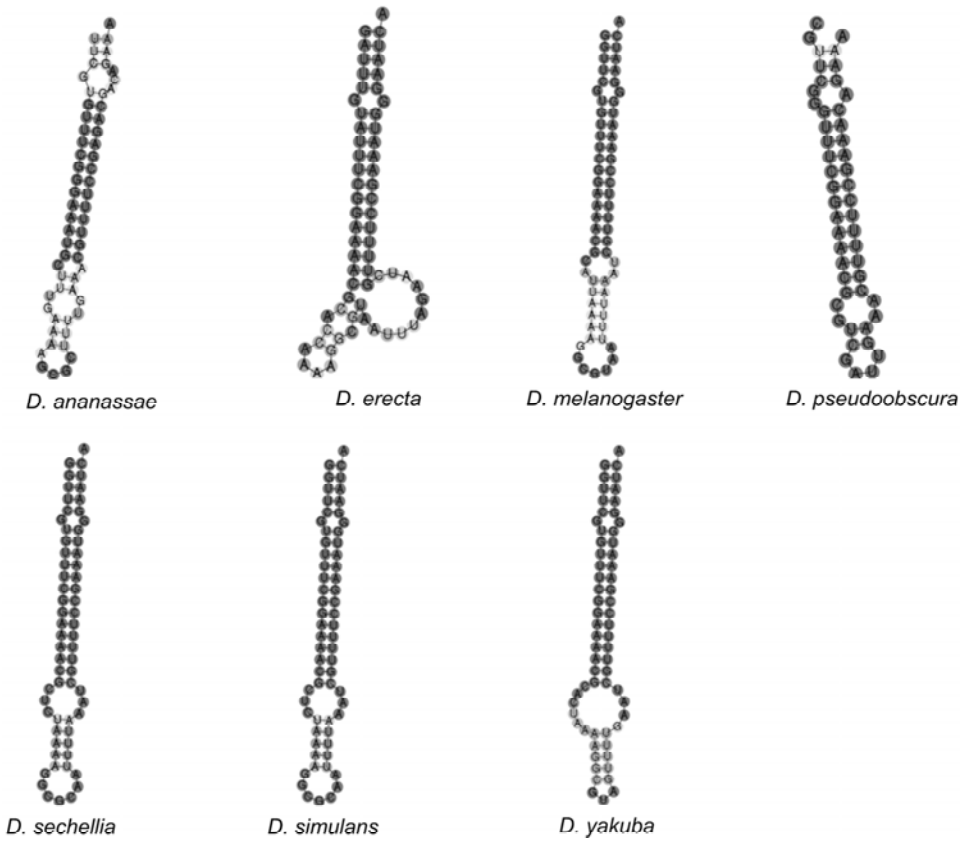

Figure 3. Samples of secondary structures of CM search results [12].

Table 4
Summary of Homologues Found in the Seven *Drosophila* Species

ID	Genome Searched	CM Score	FlyBase Region	Region Coordinates	Strand	Within *roX1*?
1	*D. ananassae*	32.26	scaffold_13117	692432 – 692373	–	Y
2	*D. erecta*	72.78	scaffold_4690	1137235 – 1137172	–	Y
3	*D. melanogaster*	88.84	chromosome X	3753295 – 3753232	–	Y
4	*D. pseudoobscura*	29.4	Unknown_ group_410	14965 – 14898	–	N
5	*D. pseudoobscura*	29.11	Unknown_ group_260	63165 – 63089	–	N
6	*D. pseudoobscura*	28.28	XL_group1e	6898105 – 6898042	–	Y
7	*D. sechellia*	88.1	scaffold_4	2954091 – 2954154	+	N/A
8	*D. simulans*	88.1	chromosome X	2759303 – 2759240	–	Y
9	*D. yakuba*	88.69	chromosome X	4661475 – 4661538	+	Y

hours using a single-processor approach. Each comparison yielded an alignment score. We then selected a group of three structures that were scored similarly and whose sequence lengths were at least 40 nt. At this step, RNA structures were obtained from *D. melanogaster*, *D. simulans* and *D. yakuba*.

3.2 CM Creation and Search

We created a CM from this group of structures by first aligning the sequences into the Stockholm format (Table 2) and then executing the cmbuild utility. The complete genomes of eight *Drosophila* species for which the presence of *roX1* ncRNA transcripts has been confirmed were used as targets in CM searches. All complete genomes used in this study were obtained from Indiana University's FlyBase database (http://www.flybase.org) [21]. These genomes were the most current releases at the time the study was conducted (Table 3). A CM search located the *roX1* genes precisely where they were known to be present in six *Drosophila* species, i.e., *D. ananassae, D. erecta, D. melanogaster, D. pseudoobscura, D. simulans* and *D. yakuba*. However, the CM search failed to locate the known *roX1* ncRNAs on the remaining two *Drosophila* species, *i.e.*, *D. mojavensis* and *D. virilis* (Table 4). In five of the six successful searches, the highest scoring search result represented a sequence within the known range of the *roX1* genomic coordinates for that species. The sixth successful search, on *D. pseudoobscura*, produced the third highest scoring search result that represented a sequence within the known range of the *roX1* genomic coordinates for that species. We hypothesize that the two highest scores for *D. pseudoobscura* represent sequences with conserved *roX1* functionality.

To make our searching approach computationally efficient, we separated the downloaded genome files into smaller files with approximately two megabase pairs (Mbp) per file maintaining small FASTA sequences intact. FASTA sequences larger than 2 Mbp were split into smaller FASTA sequences which overlapped one another by 5 kilobase pairs (Kbp) to prevent loss of accuracy in the study. This approach is similar to the slide-and-fold approach described in the RSmatch discussion in this paper. We performed concurrent identical cmsearch runs on different genome segments using NJIT's HPC cluster. In this manner, a CM search of an entire genome required only about 10 min, whereas we would have expected a genome search to take several hours using a single-processor approach.

3.3 Mining *roX1* RNA Where the Presence of *roX1* is Unknown

We wanted to identify functional structure elements in genomes of the four *Drosophila* species in which the presence of *roX1* transcripts has not been confirmed, namely, *D. grimshawi, D. persimilis, D. sechellia* and *D. willistoni*. We downloaded the most current release of these complete genomes from the FlyBase database. We used the same CM to search for presence of *roX1* functional domains. While scoring results were not significant for three of the four species, we received a strong score result from the CM search on the *D. sechellia* genome (Tables 4 and 5). We propose that this high score indicates strong evidence of a *roX1* functional domain in a specific area of the *D. sechellia* genome, namely scaffold_4. Furthermore, in spite of the *D. sechellia's* incomplete annotation, this result might indicate that this region of the genome may be located in the X chromosome of *D. sechellia*. These findings need to be experimentally confirmed.

To investigate possible *roX1* homology between species, we obtained *roX1* gene sequences FBgn0019661 (for *D. melanogaster*) and FBgn0255860 (for *D. sechellia*) from the FlyBase database and performed a pairwise alignment on the two sequences. We used the program DiAlign [24] with the "-n" option for nucleic acid sequence comparison. The result indicated a 94% similarity between the two gene sequences indicating high probability of conserved *roX1* functionality between the two species.

We have designed a systematic and computationally efficient approach to mine *roX1* RNA structure elements conserved in *Drosophila* species. This approach consists of three major steps: (1) comparison of RNA structures among all *roX1* RNAs; (2) selection of RNA structure

Table 5
Homologous RNA Sequences Found in *Drosophila* Species

ID	RNA Sequence	Length
1	UUC—GUGUUUCGGGAAAUGCUUUGAAAAGCG—CUUUUGAAACGUUUUCCGAGACGACAGAAA	60
2	GAUUUGUAUUUCGGAAAACGCACCAAAAGGCGUAAUUUAGAAUCGUUUUCCGAAAUGGGAAUCA	64
3	GGUUCGUGUUUCGGAAAACGCAUUAAAAGGCGUAAUUUUAAAUCGUUUUCCGAAAUGGGAAUCA	64
4	GACCACUCCUUCGGGUACCUCAAAAAAAaagGGCAUAGgUAUUUGGGAGGUACCCGAAGGAGUGGUCU	68
5	UCCACACGUUUCCAACUUCGUUUCCACACGC**********GUGUGGAAACGAAGUUGGAAACGCguGUGGAA	77
6	CGUUCGGGUUUCGGAAAACGCGUCGA**********UUGAAACGUUUUCCGAAAC-AGAA—A	64
7	GGUUCGUGUUUCGGAAAACGCUCUAAAAGGCGCAAUUUUAAAUCGUUUUCCGAAAUGGGAAUCA	64
8	GGUUCGUGUUUCGGAAAACGCUCUAAAAGGCGCAAUUUUAAAUCGUUUUCCGAAAUGGGAAUCA	64
9	GGUUCGUGUUUCGGAAAACGCACUAAAAGGCGUAGUUUUGAAUCGUUUUCCGAAAUGGGAAUCA	64

An asterisk (*) indicates a base that is left unaligned with a CM counterpart; a minus sign (−) indicates that no base is present to align with a CM counterpart (not included in the sequence length) and a lowercase letter represents a base on the genome that is added with respect to the CM.

groups significantly associated with those in other species and (3) utilization of a highly regarded structure-searching methodology (*i.e.*, CMs) which, in addition to being highly sensitive and specific, is also very flexible. The statistical representation of a cluster of RNA structures can be fine tuned as needed by adding or removing structures from the cluster. Using parallel processing contributes to overcoming the burden of lengthy processing times. We applied this method to mining small RNA structures chiefly because they can be more accurately forecast by those RNA prediction programs that only use thermodynamic parameters. As more powerful RNA structure prediction programs become available, particularly those reliant on phenetics for structure prediction, this approach can be extended to larger RNA structures.

3.4 Tool Comparison: Infernal versus BLAST

For the purpose of tool-effectiveness analysis, we wanted to see how a simple BLAST search might perform compared with Infernal when searching for conserved structural motifs. As BLAST is not designed to detect base pairing so critical to form an RNA secondary structure, we expected Infernal to perform better than BLAST. Table 6 presents the results of a simple test that confirmed our expectations. We used each of the three sequences from our CM and used FlyBase BLAST to search for homologues in the complete genomes of all 12 *Drosophila* species downloaded from FlyBase. Every homologue detected by BLAST was also detected by Infernal. However, BLAST failed to detect *roX1* evidence in *D. ananassae* and *D. pseudoobscura*, while such evidence was detected by Infernal. This simple experiment provides an insight into the complexity involved in the mining of ncRNA motifs.

3.5 Evaluation of the *Drosophila* Genus Complex

By conducting homology search on a complete genome, one can confirm whether a functional domain is present throughout the genome of a species or rather at specific sequence locations (*i.e.*, genomic coordinate ranges) within a specific genomic region such as a chromosome, a scaffold, *etc.* A stem-loop structure was previously predicted in *roX1* RNA on the X chromosome of *D. melanogaster* [25], and we subsequently determined that this structure is conserved in several species of *Drosophila* [7]. Our study confirms that among seven different *Drosophila* species, the *roX1* functional domain is only present on the X chromosome but also absent in any chromosome other than X. As genome annotation matures and "scaffold" regions are translated into "chromosome" regions, we will see whether this observation continues to hold.

4. Conclusion

In this study, we demonstrated that the tools RSmatch and Infernal are effective in identifying novel ncRNAs. Homology searching is a most ubiquitous undertaking in bioinformatics, yet some of the most popular homology search methods such as BLAST and FASTA, are often the least accurate [26]. Homology search tasks are more challenging for ncRNAs than a regular sequence homology search. This stems from an evolutionary perspective: ncRNA secondary structures due to intramolecular base pairs are conserved to a higher degree with respect to their primary structure, i.e., their nucleotide sequence.

An Infernal search requires a large amount of computer time [12]. For instance, Freyhult *et al.* [26] estimated that with a search query for a tRNA (a type of ncRNA), Infernal would take about 96 days to search the entire human genome on a single processor. Innovative methodologies including HMM filtering and sequence-based heuristics [27, 28] have been employed as appropriate to improve computational efficiency. In this study, as described, parallel processing with an HPC cluster was utilized for improved throughput.

Table 6
FlyBase BLAST Results Using twelve *Drosophila* Genomes As the Targets With Each of Three CM Sequences Used As a Search Query

Species Genome Searched	*D. melanogaster* Sequence Query	*D. simulans* Sequence Query	*D. yakuba* Sequence Query	Infernal *roX1* Hit Score
D. ananassae	no hit	no hit	no hit	32.26
D. erecta	1137227 (–) 1137172 52/56 (93%)	1137227 (–) 1137172 51/56 (91%)	1137227 (–) 1137172 53/56 (95%)	72.78
D. melanogaster	3753295 (–) 3753232 64/64 (100%)	3753295 (–) 3753232 61/64 (95%)	3753295 (–) 3753232 61/64 (95%)	88.84
D. mojavensis	no hit	no hit	no hit	no hit
D. pseudoobscura	no hit	no hit	no hit	28.28 (3rd)
D. simulans	2759303 (–) 2759240 61/64 (95%)	2759303 (–) 2759240 64/64 (100%)	2759303 (–) 2759240 60/64 (94%)	88.1
D. virilis	no hit	no hit	no hit	no hit
D. yakuba	4661475 (–) 4661538 61/64 (95%)	4661475 (–) 4661538 60/64 (94%)	4661475 (–) 4661538 64/64 (100%)	88.69
D. grimshawi	2 @ 18nt	1 @ 17nt	2 @ 17nt	26.13 (?)
D. persimilis	1 @ 21nt	1 @ 17nt	1 @ 19nt	39.51 (?)
D. sechellia	1 @ 61nt	1 @ 64nt	1 @ 64nt	88.1 (?)
D. willistoni	2 @ 19nt	2 @ 18nt	1 @ 18nt	35.19 (?)

Notes: Infernal search results shown for comparison. (?) = *roX1* hit unknown; "no hit" = not found within *roX1* coordinates.

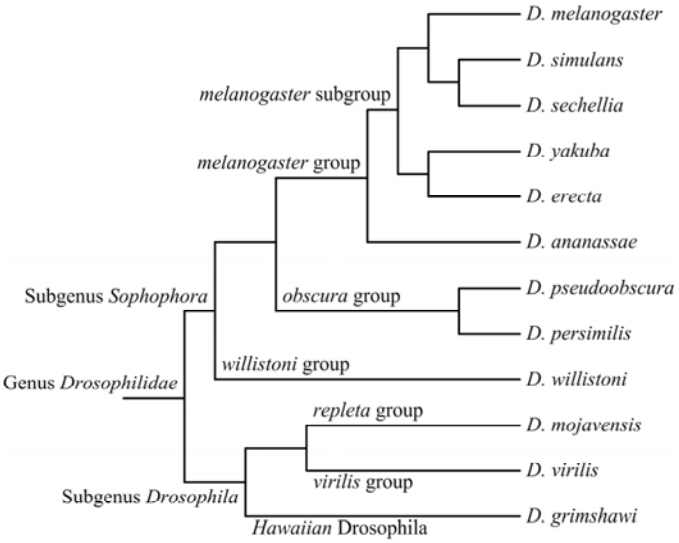

Figure 4. Phylogram of the genus *Drosophila* showing the evolutionary relationships among the 12 sequenced *Drosophila* species. Distances between nodes do not reflect base substitution span. Adapted from Stark *et al.* [9], and FlyBase [29].

We evaluated the whole genomes of all 12 species of *Drosophila* that have been completely sequenced to date. Figure 4 displays a phylogram of the genus *Drosophila* complex, which will contribute to understanding the phylogenetic relationships among the 12 *Drosophila* species evaluated in this study. All 12 species are believed to have a common ancestor that existed about 40 million years ago [7]. Phylogenetic relationships are based on the premise that species which evolved most recently will have more genetic similarities among them than those species that evolved much earlier. As a result of our study, the presence of the *roX1* ncRNA is confirmed as previously reported by other authors in six *Drosophila* species [7]. In addition, we found strong evidence of the presence of *roX1* in *D. sechellia* which, to the best of our knowledge, has not been previously reported. Through such comparative genomics, each discovery of similarities or differences among closely related species within a genus contributes to unravelling the mysteries of evolution and inborn errors of metabolism leading to human disease.

In summary, we expect that further studies will verify that *roX1* ncRNA structures can be predicted where there is evidence of *roX1* functionality. We also expect that

this bioinformatics study will lay ground work for similar accurate and efficient ncRNA mining in other organisms.

References

[1] S.R. Eddy, Non-coding RNA genes and the modern RNA world, *Nature Reviews Genetics*, *2*, 2001, 919–929.

[2] G. Storz, An expanding universe of noncoding RNAs, *Science*, *296*, 2002, 1260–1263.

[3] J.S. Mattick & I. V. Makunin, Non-coding RNA, *Human Molecular Genetics*, *15* Spec No 1, 2006, R17–R29.

[4] F.F. Costa, Non-coding RNAs: Lost in translation? *Gene*, *386*, 2007, 1–10.

[5] Y. Park & M.I. Kuroda, Epigenetic aspects of X-chromosome dosage compensation, *Science*, *293*(5532), 2001, 1083–1085.

[6] Y. Park, R.L. Kelley, H. Oh, M.I. Kuroda, & V.H. Meller, Extent of chromatin spreading determined by *roX* RNA recruitment of MSL proteins, *Science*, *298*(5598), 2002, 1620–1623.

[7] S. Park, Y.I. Kang, J.G. Sypula, J. Choi, H. Oh, & Y. Park, An evolutionarily conserved domain of *roX2* RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome, *Genetics*, *177*(3), 2007, 1429–1437.

[8] S. Park, M.I. Kuroda, & Y. Park, Regulation of histone H4 Lys16 acetylation by predicted alternative secondary structures in *roX* noncoding RNAs. *Molecular and Cellular Biology*, *28*(16), 2008, 4952–4962.

[9] A. Stark, *et al.*, Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, *450*, 2007, 219–232.

[10] A. Wang, W. Ruzzo, & M. Tompa, How accurately is ncRNA aligned within whole-genome multiple alignments? *BMC Bioinformatics*, *8*, 2007, 417.

[11] R. Durbin, S. Eddy, A. Krogh, & G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and Nucleic acids* (Cambridge UK: Cambridge University Press 1998), 277–297.

[12] E.P. Nawrocki, D.L. Kolbe, & S.R. Eddy, Infernal 1.0: Inference of RNA alignments, *Bioinformatics*, *25*(10), 2009, 1335–1337.

[13] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, & S.R. Eddy, Rfam: An RNA family database, *Nucleic Acids Research*, *31*, 2003, 439–441.

[14] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, & A. Bateman, Rfam: Annotating non-coding RNAs *in* complete genomes, *Nucleic Acids Research*, *33*, 2005, D121–D124.

[15] L.A. Kavanaugh & F.S. Dietrich, Non-coding RNA prediction and verification in *Saccharomyces cerevisiae, PLoS Genetcis* *5*(1), 2009, e1000321.

[16] J. Liu, J.T.L. Wang, J. Hu, & B. Tian, A method for aligning RNA secondary structures and its application to RNA motif detection, *BMC Bioinformatics 6*, 2005, 89.

[17] I.L. Hofacker, Vienna RNA secondary structure server, *Nucleic Acids Research 31*, 2003, 3429–3431.

[18] Y. Tabei, H. Kiryu, T. Kin, & K. Asai, A fast structural multiple alignment method for long RNA sequences, *BMC Bioinformatics*, *9*, 2008, 33.

[19] A.R. Gruber, R. Neubock, I.L. Hofacker, & S. Washietl, The RNAz web server: Prediction of thermodynamically stable and evolutionarily conserved RNA structures, *Nucleic Acids Research*, *35*, 2007, W335–W338.

[20] M. Khaladkar, J. Liu, D. Wen, J.T.L. Wang, & B. Tian, Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment, *BMC Genomics*, *9*, 2008, 189.

[21] M. Khaladkar, V. Patel, V. Bellofatto, J. Wilusz, & J.T.L. Wang, Detecting conserved secondary structures in RNA molecules using constrained structural alignment, *Computational Biology and Chemistry*, *32*, 2008, 264–272.

[22] M. Khaladkar, V. Bellofatto, J.T.L. Wang, B. Tian, & B.A. Shapiro, RADAR: A web server for RNA data analysis and research, *Nucleic Acids Research*, *35*, 2007, W300–W304.

[23] L. Peng, L.K. Ng, & S. See, YellowRiver: A flexible high performance cluster computing service for grid, *Proceedings Eighth IEEE International Conference on High-Performance Computing in Asia-Pacific Region*, 2005, 553–558.

[24] A. Subramanian, M. Kaufmann, & B. Morgenstern, DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment, *Algorithms for Molecular Biology*, *3*, 2008, 6.

[25] C. Stuckenholz, V.H. Meller, & M.I. Kuroda, Functional redundancy within *roX1*, a noncoding RNA involved in dosage compensation in *Drosophila melanogaster*, *Genetics*, *164*, 2003, 1003–1014.

[26] E.K. Freyhult, J.P. Bollback, & P.P. Gardner, Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA, *Genome Research*, *17*(1), 2007, 117–125.

[27] Z. Yao, Z. Weinberg, & W.L. Ruzzo, CMfinder – A covariance model based RNA motif finding algorithm, *Bioinformatics*, *22*(4), 2006, 445–452.

[28] Z. Weinberg & W.L. Ruzzo, Sequence-based heuristics for faster annotation of non-coding RNA families, *Bioinformatics*, *22*, 2006, 35–39.

[29] The FlyBase Consortium, FlyBase – The *Drosophila* database, *Nucleic Acids Research*, *22*, 1994, 3456–3458.

Biographies

Kevin Byron received his B.S. degree in Computer Science from New Jersey Institute of Technology, USA in 1983 and his M.S. degree in computer science from Stevens Institute of Technology, USA in 1987. He is Director of Core Systems and Operations at New Jersey Institute of Technology, USA. He is currently a Ph.D. candidate in Computer Science at New Jersey Institute of Technology, USA. His research interests are bioinformatics and *Drosophila* genomic data mining.

Miguel Cervantes-Cervantes studied Biology as an undergraduate at the Instituto Politécnico Nacional in Mexico City, where he also received a Master of Science degree in Biochemistry. His interest in the inner workings of the plant cell took him to the Waksman Institute of Rutgers, The State University of New Jersey, where he obtained a Ph.D. in biochemistry in 1991. After training as a post-doctoral associate in plant cell biology at Rutgers-Newark, he worked at the City University of New York from 1996 to 2007. Currently, Miguel is the coordinator of undergraduate studies at the Federated Department of Biological Sciences of Rutgers-Newark and the New Jersey Institute of Technology. He continues doing research on cereal seed germination and metabolic engineering, focusing on the isoprenoid biosynthetic genes and enzymes, and has started utilizing bioinformatics tools in his research on metabolic networks as well as in graduate courses in plant science.

Jason T.L. Wang received a B.S. in Mathematics from National Taiwan University, Taipei, Taiwan, and a Ph.D. in Computer Science from the Courant Institute of Mathematical Sciences at New York University. He is a Professor of computer science and bioinformatics at the New Jersey Institute of Technology. His research interests include data mining, computational biology and bioinformatics. He has published over 120 peer-reviewed papers, 5 books, has been a program committee member of over 100 conferences, and is on the editorial board of 15 journals as well as the Executive Editor of the World Scientific book series on science, engineering and biology informatics. He is currently the principal investigator of an NSF-funded project on RNA genomics.

Wen-chang Lin received his B.S. degree in Agricultural Chemistry from National Taiwan University, Taiwan in 1986 and his Ph.D. degree in Molecular Biology and Microbiology from Case Western Reserve University, USA in 1991. After training as a Post-Doctoral Associate in tumor immunology and gene therapy at Pittsburgh Cancer Institute from 1991 to 1993, he returned to Taiwan in 1993 and joined Institute of Biomedical Sciences, Academia Sinica, Taiwan. He is currently an Associate Research Fellow at Academia Sinica. Wen-chang's research interests include tumor progression and biomarker discovery; bioinformatic studies on human genome annotation and transcriptome data-mining (subtle wobble splicing); microRNA discovery and functional characterization.

Yongkyu Park received a B.S. in Biology from Korea University, Seoul, Korea, and a Ph.D. in microbiology from Korea Advanced Institute of Science and Technology (KAIST) in 1999. From 1999 to 2004, he was a Postdoctoral Researcher in molecular genetics at Baylor College of Medicine, Houston, Texas, and then a Postdoctoral Researcher at Harvard Medical School, Boston, Massachusetts. He is currently an Assistant Professor in the Department of Cell Biology and Molecular Medicine of the New Jersey Medical School in the University of Medicine and Dentistry of New Jersey. His researcher interest is focused on noncoding RNA (ncRNA) in *Drosophila*. He is the principal investigator of an NSF-funded project on roles of ncRNA in global chromatin organization.