Small RNA in rice genome

WANG Kai (王 凯)[1], ZHU Xiaopeng (朱小蓬)[2], ZHONG Lan (钟 兰)[1,3] & CHEN Runsheng (陈润生)[1,2]

1. Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China;
2. Laboratory of Bioinformatics, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;
3. College of Life Sciences, Peking University, Beijing 100871, China

Correspondence should be addressed to Chen Runsheng (email: crs@sun5.ibp.ac.cn)

Received July 27, 2002

Abstract Rice has many characteristics of a model plant. The recent completion of the draft of the rice genome represents an important advance in our knowledge of plant biology and also has an important contribution to the understanding of general genomic evolution. Besides the rice genome finishing map, the next urgent step for rice researchers is to annotate the genes and noncoding functional sequences. The recent work shows that noncoding RNAs (ncRNAs) play significant roles in biological systems. We have explored all the known small RNAs (a kind of ncRNA) within rice genome and other six species sequences, including *Arabidopsis*, maize, yeast, worm, mouse and pig. As a result we find 160 out of 552 small RNAs (sRNAs) in database have homologs in 108 rice scaffolds, and almost all of them (99.41%) locate in intron regions of rice by gene predication. 19 sRNAs only appear in rice. More importantly, we find two special U14 sRNAs: one is located in a set of sRNA ZMU14SNR9(s) which only appears in three plants, 86% sequences of them can be compared as the same sequence in rice, *Arabidopsis* and maize; the other conserved sRNA XLHS7CU14 has a segment which appears in almost all these species from plants to animals. All these results indicate that sRNA do not have evident borderline between plants and animals.

Keywords: ncRNA, sRNA, conserved sequences, plant-specific sRNA gene.

With the completion of rice genome working-draft[1,2], we have a chance to peek the difference between high-class animal genome and plant genome. What is the difference between them? To answer this question, we focus on noncoding RNA (ncRNA).

Noncoding RNA is involved in a good many of the most basic biological functions, including gene splicing, RNA nucleotide modification, protein transport and regulation of gene expression[3]. The recent work shows that, as a kind of noncoding RNA, small RNA (sRNA) plays a significant role in biological systems.

In *Caenorhabditis elegans*, *lin-4* and *let-7* encode 22- and 21-nucleotide (nt) RNAs, respectively, which function as key regulators of developmental timing[4]. The accumulating evidence makes us believe sRNA is indispensable for a life organism to regulate its system.

Plant and animal are totally different biological systems using different life strategies. However, they use almost the same proteins to construct their life systems. The difference between

them is the constructing way rather than the constructing materials. Therefore, comparing gene's regulators is more significant than comparing genes themselves.

The distribution of sRNA genes on genome reflects the specie's character in regulating system. Our intriguing approach is that by comparing the distribution of sRNA genes in different genomes we hope to discover the difference between plant and animal regulating systems. In order to do this, we systematically explored the known sRNA genes in rice genome and compared these rice-contained sRNA genes with other genomes. In this way, we identified some plant-specific sRNA genes. And we also found that the rice genome contains some sRNA genes which are not shown in any other known plant genomes. Here, we report the distribution of the sRNA genes in rice genome and the distribution of the rice-contained sRNA genes among species. This is the first step to research sRNA genes in rice.

1 Materials and methods

We used the program BLASTN[5], obtained from the National Center for Biotechnology Information (NCBI), to search for homologs of the known sRNAs from yeast, worm, plants and mammals. An important resource used to conduct this search is the sRNA database, which was built on January 19, 1999[6].

Rice scaffolds were generated in Beijing/Hangzhou Genomics Institute (BGI)[7]. Compared with the impending completion of the sequencing of the rice genome and entire genome of *Arabidopsis thaliana* and *Caenorhabditis elegans*, the maize data are only some EST contigs sequences (total 18260)[8], which were updated on May 10, 2002. The pig data, including 209967 sample reads, were produced by our center, BGI. The mouse, *Arabidopsis*, *C. elegans*,, and *S. cerevisiae* sequences were all obtained from NCBI.

The output of the program BLAST was parsed with a PERL program to extract the sRNA sequence corresponding to rice and other six species sequences. To make sense in which areas those homologs between rice data and sRNAs sequences are located, another PERL program was used to extract introns location information in rice, which resulted from prediction program of multiple genes FGENESH (Salamov A.A., Solovyev V.V., 1999), and tell us the location of these segments.

Also we used CLUSTAL W[9] to do the two multiple sequence alignments. One is about sRNA ZMU14SNRs with rice, *Arabidopsis* and maize, the other is about sRNA XLHS7CU14 with all six species, to see what are common and what are unique between the same or different species at the genome level.

2 Results

Using the 103044 rice scaffolds produced in our center (BGI) by shotgun sequencing as the database, we compared sRNA databases, including 552 small RNA sequences, by program BLASTN[5]. As a result, 160 small RNAs with the length from 36 bp to 347 bp (the average is 100 bp) were found in 108 rice scaffolds. 12 out of the 160 sRNAs are located in 11 rice ESTs.

99.41% of the rice sRNAs are located in the predicted intron regions.

Subsequently, each sRNA found in rice was used as a query to align to the other six genome sequences by BLASTN. The six genomes include two plants, *Arabidopsis* and maize, and yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabditis elegans*), mouse (*Mus musculus*), pig (*Sus scrofa*).

Table 1 shows the distribution of sRNA genes among these seven species. We can see the Ux family is the most common, while RPS3A and Y4 are rarer and are just found in a few species. The types and relative numbers of sRNA in rice look quite similar to those in *Arabidopsis*, especially from U1 to U6 sRNA genes. Pig is another species in which plentiful sRNAs have been found. Only a few sRNAs have been detected in maize and mouse genomes, such as U3, U5 and U14 in maize and RPS3A, U14 in mouse.

Table 1 Distribution of sRNA genes that had been compared with rice in the other six species

sRNA gene	Rice	*Arabidopsis*	Maize	Yeast	Worm	Mouse	Pig
Alu	14				8		10
RPS3A	1			1		1	1
U1	21	10					
U2	55	54		34	2		50
U3	10	6	1				
U4	5	5			1		
U5	7	7	1				
U6	33	33		2	31		33
U11	1						1
U14	10	9	9	1	2	2	
U17	2				2		2
Y4	1						
Sum	160	124	11	38	46	3	97

We compared the frequencies of sRNA found in rice, *Arabidopsis* and pig, especially in U2 gene family. They are divided into different sets by their names. Generally sRNAs in the same sets have the same frequency in one species, but we found two special sets of sRNAs in rice. One is Locus HUMUGA, G2A, G21, G20 and G2, as showed in fig. 1(a). Except for sRNA HUMUGA, the others all had the average frequency number of sRNA in rice (32). By comparing the numbers of sRNA HUMUGA in the three species we can see this sRNA did not appear in *Arabidopsis* at all and it also had low frequency in pig. Fig. 1(b) shows the other special set, including Locus S72337, S64585, S64581 and S64577. This set had the highest point in rice sRNA (Locus S72337), whose frequency is 70, and it also had the same diversity with the three sRNAs nearby.

Table 2 shows that 19 sRNAs were only found in rice genome, while those sRNAs were not distributed in the other six species. The lengths of the 19 sRNAs are between 100 and 400 bases. They mainly focus on Alu, U1 and U3 genes. When comparing the sRNAs in rice and *Arabidopsis* it is interesting to see that Alu can only be found in rice, and the number of U1 in *Arabidopsis* is

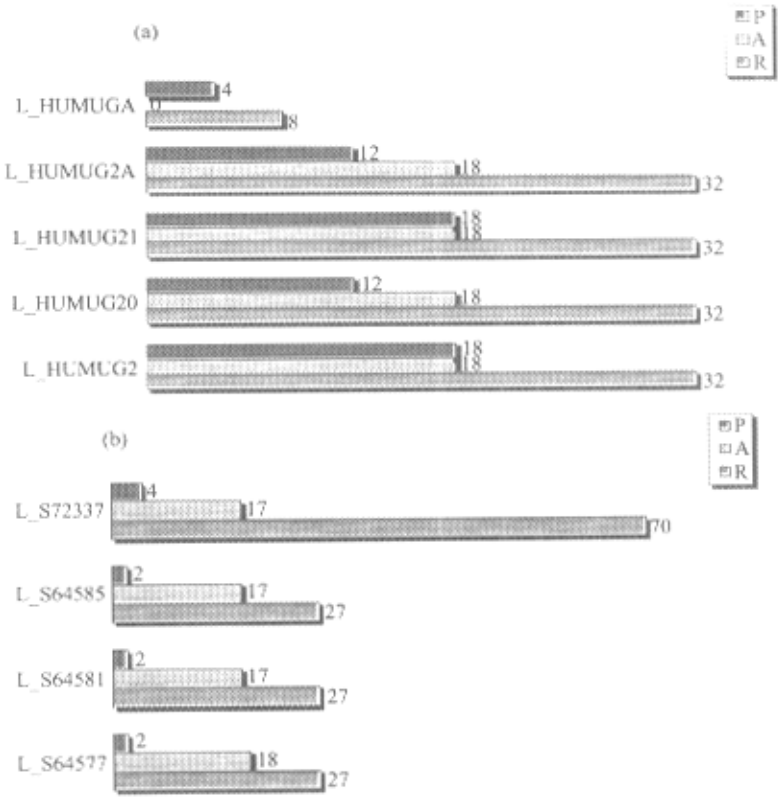

Fig. 1. sRNA distribution in rice, and *Arabidopsis* and pig.

Table 2 19 sRNAs only appear in rice sequences

Class	sRNA	Size/bp	Frequency
Alu	Locus_HUMSCALUA	121	9
Alu	Locus_HUMSCALUC	118	7
Alu	Locus_HUMSCALUF	120	9
Alu	Locus_HUMSCALUJ	119	8
Alu	Locus_HUMSCALUN	120	9
U1	Locus_S72336	162	15
U1	Locus_TAU1C	159	15
U1	Locus_TAU1D	161	15
U1	Locus_TAU1F	139	13
U1	Locus_TAU1G	142	15
U1	Locus_TAU1H	158	15
U1	Locus_TAU1I	161	16
U1	Locus_LESNRU13	368	3
U1	Locus_LESNRU14	388	3
U1	Locus_RNSNRU15	385	2
U1	Locus_LESNRU17	394	4
U3	Locus_TRLTGV3SNR	610	6
U3	Locus_ZMU3SNRNG	1007	6
Y4	Locus_HUMSCRNY4A	844	4

just half of that in rice.

We compared 552 sRNAs genes with 7 species genomes. 10 of them exist only in plants. Fig. 2 shows the plant-specific sRNA genes and their distribution. As the figure shows, the most abundant plant-specific sRNA gene is ZMSNORNA1 which is defined as "*Z. mays* small nuclear RNA genes snoR1.1, snoR2.2, snoR3.2". Compared with the other 9 plant-specific sRNAs, ZMSNORNA1 has the highest frequency in rice and maize. It also has relatively high frequency in *Arabidopsis*.

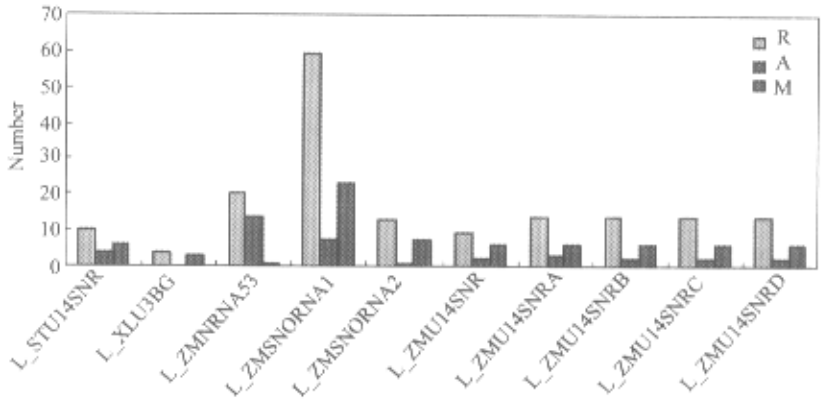

Fig. 2. Distribution of known sRNAs found only in rice, *Arabidopsis* and maize.

We can see that the last four sRNAs (Locus_ZMU14SNRA, B, C and D), whose sequence lengths are all between 124 and 129 bp, show almost the same results. Using CLUSTAL W[9] multiple alignments we can find more than 100 bases out of these four sRNAs (86%) have homologs in rice, *Arabidopsis* and maize. Table 3 shows parts of the results from CLUSTAL W. There are 11 entries in sRNA (5), rice (4), *Arabidopsis* (1) and maize (1) as representatives. Actually, the sRNA frequency in the three plants is about 14 in rice, 2 in *Arabidopsis* and 6 in maize.

There is another conserved sequence in sRNA, Locus XLHS7CU14[10], which is defined as "*X. laevis* DNA for intronic U14 small nucleolar RNAs from African clawed frog". This segment is the most impressed one in this sRNA database, because its segment from 1784 to 1955 base has approximately 20 homologs found not only in rice, *Arabidopsis* and maize, but also in yeast, worm and mouse. These homologs vary from 53 bp (maize) to 457 bp (mouse) in length.

We compared XLHS7CU14 with 7 databases using program BLASTN and the results are presented in table 4.

3 Discussions

Noncoding RNAs (ncRNAs) have been found to have roles in a great variety of processes[11]. This work is the first step toward the systematic identification and study of sRNA, as one kind of ncRNAs, in rice. Here we present the distribution of the sRNA genes in rice genome. The result shows that rice genome has some unique characteristics. In common sense, the rice genome should be more similar with plant genome than with animal genome. Actually, the rice genome

Table 3 Multiple alignments of ZMU14SNR sRNA, rice, *Arabidopsis* and maize

```
L_ZMU14SNR          -----TGCATTGCAAGTGATGATGAAGTC-AAGGCTTGTT---TCTCTACATTCGCAGTTGCCGCCTAAGAGC
L_ZMU14SNRA         ---------TATGGCAATGATGATGAAAGATAAGGCTTGTT---TCTCAACATTCGCAGTAGCCGCCTAAGAGC
L_ZMU14SNRB         -----TGCATTGCAAGTGATGACAAAATC-AAGGCTTGTT---TCTCGACATTCGCAGTTGCCGCCTAAGAGC
L_ZMU14SNRC         --CTGCATTGCAAATGATGCTATAATC-AAGGCTTGTTTC-TCATGACATTCGCAGTTGCCGCCTAAGAGC
L _ZMU14SNRD        ------------------------------------------------------------CTTGTT---TCTCTACATTCGCAGTTGCCGCCTAAGAGC
Rice_S5449          TCTGCATGGCAAATGATGCTAAAAGC-AAGGCTTGTTTCTCATAACATTCGCAGTTGCCGCCTAAGAGC
Rice_S6086          GATGGATGCCTTGTGATGCTAAAATC-AAGGCTTGTTTC-TCATGACATTCGCAGTTGCTGCCTAAGAGC
Rice_S72833         --------CGGTGCCCTATGATGACAAAATC-AAGGCTTGTT---TCTCTACATTCGCAGTTGCCGCCTAAGAGC
Rice_S9185          GATTATATGGCAATGATGATAAATTTAAGGCMTTGTTTC-TCATAACATTCGCAGTTGCCGCCTAAGAGC
Maize_16272.1       ----------TATGGCAATGATGTTGAAGTTAAAGGCTTGTT---TCTCAACATTCGCAGTAGCCGCCTAAGAGC
Arabidopsis_chr4    AATAGTATGGCAATGATGATAAATTT-AAGGCTTGTTTC-TCATAACATTCGCAGTTGCCGCCTAAGAGC
L_ZMU14SNR          TTTCGCCCT--GCCAGGCTTGAGAGCTAGTGCTGCC--AAATCCTTCCTTGGATGTCTGATGCAATGCA--
L_ZMU14SNRA         TTTCGCCAC--GCCAGGCTCGAGAGCTTGTGCTGTT--GAATCCTTCCTTGGATGTCTGAGCCATA--------
L_ZMU14SNRB         TTTCGCCCT--GCCAGGCTTGAGAGGTAGTGCTGCC--AAATCCTTCCTTGGATGTCTGACGCAATGCA-
L_ZMU14SNRC         TTTCGCCCT--GCCAGGCTTGAGAGCTAGTGCTGCT--AATTCCTTCCTTGGATGTCTGATGCCTAGCAA
L_ZMU14SNRD          TTTCGCCTT--GCCAGGCTTGAGAGCTAGTGCTGCT--AAATCCTTCCTTGGATGTCTGAT----------
-Rice_S5449         TTTCGCCTT--GCCAGGTTTGAGAGCT----------------------------------------------
Rice_S6086          TTTCGCCCT--GCCAGGCTTGAGGGTTAGTGCCACTG-----------------------------------
Rice_S72833         TTTCGCCCT--GCCAGGCTTGAGAGCTAATGCTGCA--GAATCCTTCCTTGGATGTCTGAGGGCCGCCG-
Rice_S9185          TTTCGCCCT--GCCAGGCTTGAGAG-----------------------------------------------
Maize_16272.1       TTTCGCCCT--GCCAGGCTTGAGAGCTTGTGCTGTT--TAATCCTTCCTTGGATGTCTGAGCCATA--------
Arabidopsis_chr4    TCTCGCCCT--GCCAGGCTTGAGAGCTAATGCTGTT--GATTCCTTCCTTGGATGTTTGAGCCATT---------
-
```

Table 4 Comparison results of sequences on XLHS7CU14 sRNA in six species by BLASTN

Species	Sequence name	Species segment		Homolog/length	SRNA Segment	
		from	to		from	to
Rice EST	Contig13764	666	847	152/182	1774	1955
	Contig12514	392	522	107/131	1784	1914
	Contig13571	712	883	138/172	1784	1955
	Contig10604	713	884	138/172	1784	1955
Rice	Scaffold15080	620	773	125/154	1802	1955
Arabidopsis	chr5	555412	555314	84/99	1774	1872
	chr1	20705541	2070574	140/174	1782	1955
Maize	TUC8398.1	728	909	148/182	1774	1955
	TUC11262.1	1708	1527	148/182	1774	1955
Yeast	ref|NC_001144.2|	96862	96799	58/64	1827	1890
	ref|NC_001133.1|	140806	140744	57/63	1828	1890
	ref|NC_001137.2|	365225	365281	53/57	1842	1898
Worm	chr_IV	15675311	1567547	144/167	1784	1950
	chr_X	3046321	3046403	75/83	1784	1866
	chr_I	13941520	1394153	58/64	1827	1890
Mouse	E00000039724	559	887	271/329	1757	2085
	E00000039737	559	887	271/329	1757	2085
	E00000045027	559	1117	457/560	1757	2316
	E00000015800	572	1130	457/560	1757	2316
	E00000007247	729	853	103/125	1784	1908

does have some sRNA genes found in animal genome but not found in plant genome, e.g. Alu, U17, U11 and RPS3A. However, most rice-contained sRNA genes also exist in *Arabidopsis* (U2, U6, U14). The most interesting thing is the sRNA gene U2. We found 55 U2 genes in rice genome, 50 of which exist in pig genome segments but none exists in mouse genome. However, as we all know, the mouse genome data are far more plentiful than the pig genome. All of these indicate that there is no clear boundary line between plant sRNA genes and animal sRNA genes.

What we are most interested in is the plant-specific sRNA gene. And here we report 10 sRNAs as plant-specific sRNA genes and their distribution among rice, maize and *Arabidopsis*. And the results show no significant distribution bias among these species.

As more and more evidence shows that sRNA is a very basic regulator for an organism, we believe that sRNA has existed since life began. But to our surprise, only one sRNA gene exists in all the six species. In order to interpret this phenomenon, we present a hypothesis: if we assume the sRNA gene can easily transfer from one genome to another, then all contradictions above will be solved. Considering the RNA characteristic, it may be the most possible way. We assume that sRNA can be reversed to transcript into DNA, and the sRNA gene is some kind of retrotransposons. Then, the distribution of sRNAs in genome mainly depends on the function of sRNA rather than on the origin of genome. That is why there is no clear boundary line between plant sRNA genes and animal sRNA genes. And we can interpret why there is only one sRNA gene existing in all six species.

Acknowledgements We acknowledge the work of all those who have participated in the Rice Genome Draft Sequencing as well as Pig Genome Project in both China and Denmark, which made possible the analysis presented here. We also thank all of our colleagues at BGI in Hangzhou Center for their work on rice EST analysis. We would like to thank Ye Chen, Huang Xiangang, Ren Xiaoyu and Zhang Jianguo for their help.

References

1. Yu, J., Hu, S. N., Wang, J. et al., A draft sequence of the rice (*Oryza sativa* ssp. *indica*) genome, Chinese Science Bulletin, 2001, 46(23): 1937—1951.
2. Yu, J., Hu, S. N., Wang, J. et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science, 2002, 296: 79—91.
3. Eddy, S. R., Noncoding RNA genes, Curr. Opin. Genet. Dev. , 1999, 9: 695—699.
4. Lagos-Quintana, M., Rauhut, R., Lendecket, W. et al., Identification of novel genes coding for small expressed RNAs, Science, 2001, 294: 853—858.
5. Altschul, S. F., Gish, W., Miller, W. et al., Basic local alignment search tool, J. Mol. Biol., 1990, 215: 403—410.
6. http://mbcr.bcm.tmc.edu/smallRNA/Database/
7. http://btn.genomics.org.cn/rice/
8. http://www.zmdb.iastate.edu/
9. Thompson, J. D., CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research, 1994, 22: 4673.
10. Xia, L., Liu, J., Sage, C. et al., Intronic U14 snoRNAs of Xenopus laevis are located in two different parent genes and can be processed from their introns during early oogenesis, Nucleic Acids Research, 1995, 23 (23): 4844—4849.
11. Storz, G., An expanding universe of noncoding RNAs, Science, 2002, 296: 1260—1263.