Journal of Molecular and Genetic Medicine

Kondo et al., J Mol Genet Med 2017, 11:2
DOI: 10.4172/1747-0862.1000271

Research Article Open Access

Comparative Analysis of Intronic Noncoding RNA Genes among Organisms

Kondo Y[1], Hayashi C[2] and Miyazaki S[1*]

[1]Tokyo University of Science, Yamazaki, Noda-shi, Chiba, Japan

[2]Quintiles Transnational Japan K.K., Japan

Abstract

Development of sequencing techniques allowed us to determine genomic sequences in many organisms. Such a determined genome consists of not only protein-coding genes but also noncoding RNA (*ncRNA*) genes. We should analyze evolutional histories of such genes to estimate evolutional directions in future. Meanwhile, recent studies showed that some *ncRNA* genes are located in intragenic regions in protein-coding genes, which are called host genes. We considered that such information can help us to discuss gene evolutions. In this study, we constructed a database to analyze evolutions of protein-coding and noncoding genes based on gene locations in genomic sequences. We found that 547 out of 2,691 human host genes are orthologous to 546 out of 1,633 mouse host genes. Such orthologous host genes are involved in similar biological functions but some non-orthologous host genes have different functions. For example, non-orthologous host genes in human are annotated as neuron-related terms but such genes in mouse are not. Meanwhile, similarity searches for intronic microRNA (*miRNA*) genes between human and mouse showed that 85 out of the orthologous host genes have retained *miRNA* genes in the intronic regions. 64 out of such genes have retained intronic *miRNA* genes among human, mouse and rat. These results suggest that some orthologous genes have retained *ncRNA* genes in the intronic regions in the evolutionary process.

Keywords: Genome; Protein coding region; Gene evolution; Introns; Database

Introduction

A genomic DNA sequence has some kind of meaningful regions involved in gene expressions. Such a region is present sporadically in the DNA sequence. This region, for example, can become RNA molecules or regulate gene expressions. Converting from DNA to RNA sequence is called transcription, which usually makes an exact copy of the DNA sequence. This information flow is one of the important steps in gene expressions, which can make various transcripts playing a variety of roles. Some transcripts can work as the unprocessed form whereas some transcripts are further processed. Such processing makes a no more exact copy of the DNA sequence. One of the representative processing ways is elimination of stretches of RNA sequences like splicing, in which an inter-region of two exons (short for expressed regions) is eliminated from the precursor messenger RNA (pre-mRNA). The spliced region is called an intron (short for intragenic region) [1]. Splicing can create a mature form (mRNA) from the immature form (pre-mRNA). The mature mRNA is then converted into amino acids. This information flow converts an RNA sequence into an amino acid sequence based on a codon usage table. Therefore, the DNA sequence eventually converted into an amino acid sequence is called a protein-coding sequence. This is an essential part of expressions for protein-coding genes [2].

Meanwhile, other than protein-coding genes exist. Such genes are grouped as noncoding RNA (*ncRNA*) genes [3,4]. Some *ncRNA* s play important roles in protein biosynthesis. Such *ncRNA* s are, for example, small nuclear RNA (snRNA), small Cajal body-associated RNA (scaRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and small nucleolar RNA (snoRNA). snRNAs, which are modified by scaRNAs, are included in a spliceosome, which consists of many proteins and five snRNAs and works for splicing of pre-mRNAs [5]. A tRNA can deliver an amino acid to an mRNA on a ribosome, which works for elongating an amino acid sequence [6]. Such a ribosome includes rRNAs derived from precursor rRNAs (pre-rRNAs), which are processed and modified by snoRNAs [7]. In addition, some *ncRNA* s work for expression regulation of protein-coding genes. MicroRNAs (miRNAs) play a role in silencing protein-coding genes [7] and some miRNAs are involved in brain or neuron related diseases such as Alzheimer's disease [8]. Long ncRNAs (lncRNAs), which are longer than 200 bases, can serve as molecular signals and some l*ncRNA* s are upregulated or downregulated in cancers [9,10]. Piwi-interacting RNAs (piRNAs) are responsible for epigenetic inheritance in germ line and germline bordering somatic cells [11]. Thus, ncRNAs, which are classified into many subcategories, have many functions such as working with proteins and regulating gene expressions.

How genes were evolved is important to predict gene evolutions in future. Evolutionary lineages among protein-coding genes have been well discussed [12]. However, it is not well known how *ncRNA* genes were evolved. One of the reasons is that evolutional analysis for *ncRNA* genes is more difficult than protein-coding genes because alignment accuracy for *ncRNA* genes may be lower than amino acid sequences [13,14]. Sequence alignments of *ncRNA* genes cannot use information regarding a codon usage table like protein-coding genes. Therefore, it is difficult to identify important regions of the *ncRNA* gene. In addition, such low accuracy can be caused by the fact that the number of base types is lower than amino acid types.

Meanwhile, some *ncRNA* genes are located in intronic regions [15]. Such a protein-coding gene including an *ncRNA* gene in the intronic region is called a host gene. Host genes and intronic *ncRNA* genes are interesting to investigate how the transcriptions are regulated because such an *ncRNA* can be transcribed simultaneously with the host gene. This shows that a relationship between an intronic *ncRNA*

***Corresponding author:** Dr. Satoru Miyazaki, Department of Medicinal and Life Science, Faculty of Pharmaceutical Sciences, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba 2788510, Japan, Tel: 81471213630; E-mail: smiyazak@rs.noda.tus.ac.jp

Received May 24, 2017; **Accepted** June 14, 2017; **Published** June 19, 2017

Citation: Kondo Y, Hayashi C, Miyazaki S (2017) Comparative Analysis of Intronic Noncoding RNA Genes among Organisms. J Mol Genet Med 11: 271 doi:10.4172/1747-0862.1000271

gene and a host gene is useful to discuss how gene expressions are regulated. Expressions for intronic *ncRNA* genes are regulated by some mechanisms. Some intronic *ncRNA* genes depend on transcriptions of the host genes [16]. Therefore, such an *ncRNA* gene shares a transcription unit with the host gene. On the other hand, some intronic *ncRNA* genes are regulated by an independent transcription unit which has an own promoter [17]. These mechanisms for expression regulation of *ncRNA* genes suggest that some expression regulation of *ncRNA* genes depend on gene locations in genomic DNA sequences. Such researches for intronic *ncRNA* genes have shown that many *ncRNA* genes are located in intronic regions [18-21]. Therefore, we should classify *ncRNA* genes based on gene locations in genomic sequences.

In this study, we construct a database classifying *ncRNA* genes based on gene locations in genomic sequences by collecting information concerning some model organisms of the genomic sequences. The database stores orthologous relationships between protein-coding genes. We then summarize statistics for coding and noncoding genes and orthologous relationships between organisms. Moreover, what functions are enriched in the host genes is investigated by focusing on host genes of intronic *ncRNA* genes. Furthermore, sequences of *ncRNA* genes are compared in order to investigate whether host genes have re-tained *ncRNA* genes in the intronic regions. We discuss whether such gene locations in genomic sequences are effective to discuss gene evolutions.

Materials and Methods

Classification for *ncRNA* genes based on mRNA-transcribed locations

We propose a new classification for *ncRNA* genes based on gene locations in a genomic sequence. An outline of the classification is shown in Figure 1. The DNA sequence in Figure 1 is separated into two as pre-mRNA-transcribed and intergenic regions. This separation can divide *ncRNA* genes into three categories.

1. Intergenic (located on an intergenic region)
2. Intronic (located on a pre-mRNA-transcribed region)
3. Sense (located on a boundary region)

Construction of a database based on gene locations

The database consists of 9 tables shown in Figure 2. The database was designed to store all the data downloaded from the Ensembl genome database (release 87) [22]. The 'gene sets' table contains the data regarding gene sets downloaded from Ensembl in GTF format. This table only includes the data whose feature is 'transcript'. The '*ncRNA*' table contains the data regarding *ncRNA* genes downloaded from Ensembl in fasta format. The 'intronic *ncRNA*' table contains information regarding the host gene including the intronic *ncRNA* gene. The 'sense *ncRNA*' table contains information regarding the protein-coding gene overlapping with the sense *ncRNA* gene. The 'category' table stores the three categories of *ncRNA* genes. The 'organism' table stores 6 organisms: *Homo sapiens* (human), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (nematode) and *Saccha-romyces cerevisiae* (yeast). The 'coding gene' table contains the data regarding coding genes extracted from the 'gene sets' table. The 'ortholog' table contains the data regarding orthologs downloaded from Ensembl BioMart [23]. The 'type' table contains orthologous types such as one2one, one2many and many2many [12].

Gene set enrichment analysis for host genes

We conducted gene set enrichment analyses (GSEA) by Gene Ontology (GO) terms [24] using the GOstats package [25] in Bioconductor. All host genes and coding genes in human and mouse were extracted from the database. Then, the host genes were divided into orthologous and non-orthologous genes. We conducted GSEA of the host genes compared with all coding genes in human and mouse. In addition, we also extracted orthologous host genes possessing intronic miRNAs detected by BLAST searches described below. We conducted GSEA of the detected genes compared with all orthologous host genes in human and mouse. In these GSEA, detected GO terms were visualized by the tagcloud R package. We investigated whether the GO terms are identical or similar between human and mouse.

Sequence comparisons for intronic *miRNA* genes

Sequences of all intronic *miRNA* genes were extracted from the database. All combinations of a pair of BLASTN searches [26,27] were conducted by setting the word size as 14. We then investigated whether a detected pair of host genes possessing *miRNA* genes in the intronic regions is orthologous or not.

Results

Statistics of coding and noncoding genes

Our database stores 462,331 transcripts including 231,749 protein coding transcripts shown in Table 1. These transcripts are produced from 105,246 protein coding genes. On the other hand, other than protein coding transcripts relate with *ncRNA* s, pseudogenes and immunological products. Some pairs of protein-coding genes between organisms have a same evolutional origin. Our database stores such information as orthologous relationships shown in Table 2, which shows, for instance, 18,023 human genes are orthologous to 18,425 mouse genes. This shows that the numbers of human and mouse orthologous genes are different because of three types of orthologs; one-to-one, one-to-many and many-to-many.

All *ncRNA* genes on chromosomes were annotated with one of the three categories: intergenic, intronic or sense. In each organism, the intergenic *ncRNA* gene is the largest number and the intronic *ncRNA* gene is the second largest in the three categories as shown in Table 3. The intronic *ncRNA* genes have a variety of biotypes as shown in Table 4. The biotypes include *ncRNA* s with known and unknown functions. For instance, *ncRNA* s with known functions are miRNA, rRNA, ribozyme, scaRNA, snRNA, snoRNA and tRNA. *ncRNA* s with unknown functions are 3′ overlapping ncRNA, antisense, lincRNA, misc RNA, non-coding, ncRNA, processed transcript, retained intron, sense intronic, sense overlapping.

Table 3 shows 3,886 intronic *ncRNA* genes in human and Figure 3 shows 2,691 host genes in human. In addition, Table 3 shows 2,267 intronic *ncRNA* genes in mouse and Figure 3 shows 1,633 host genes in mouse. These results indicate that some host genes include one or more *ncRNA* genes in the intronic regions. Figure 3 also shows that 547 and 546 host genes have orthologous relations between human and mouse, respectively. These orthologs included one-to-many orthologous relations. For example, one human gene, ENSG00000104131, is orthologous to two mouse genes, ENSMUSG00000027236 and ENSMUSG00000043424. One mouse gene, ENSMUSG00000030738, is orthologous to two human genes, ENSG00000205609 and ENSG00000184110. One mouse gene, ENSGMUSG00000013701, is orthologous to two human genes, ENSG00000265354 and ENSG00000204152.

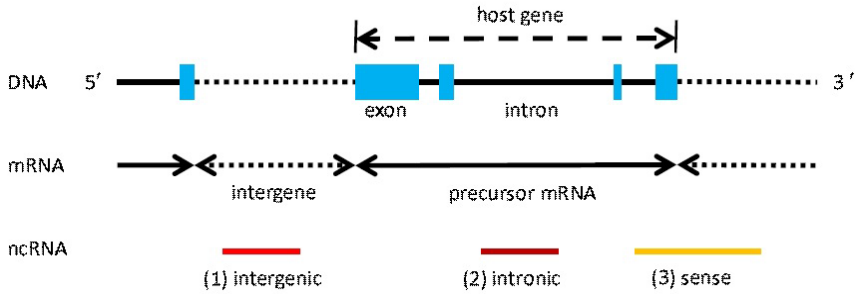

Figure 1: Classification for ncRNA genes based on mRNA-transcribed regions in a DNA sequence. ncRNA genes are classified into three categories: (1) intergenic, (2) intronic and (3) sense. An intergenic ncRNA gene is located on a region not transcribed into precursor mRNAs. An intronic ncRNA gene is located on a region trancribed into a precursor mRNA. A sense ncRNA gene is located on a boundary region between an intergene and precursor mRNA.

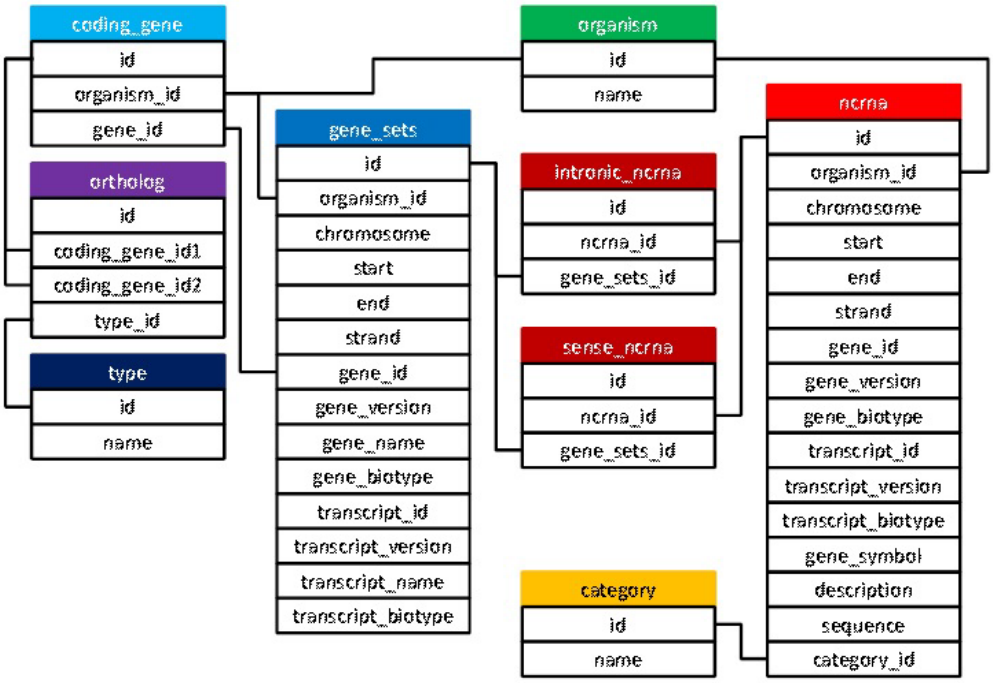

Figure 2: A database schema based on gene locations in genomic sequences. The colored or white boxes show table or column names, respectively. Two tables can be joined by the relation connected by the black line.

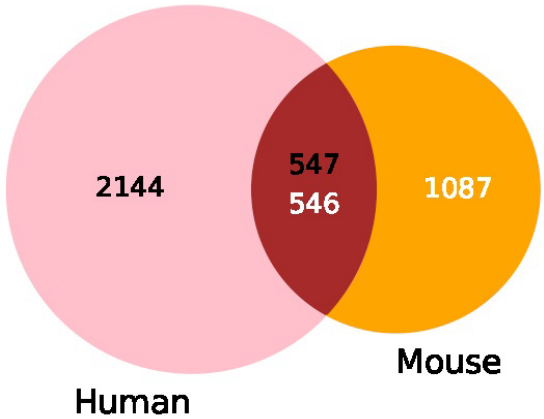

Figure 3: Orthologous genes in host genes.

Organism	Transcript	Coding transcript	Coding gene
Human	198,002	80,058	19,961
Mouse	123,063	54,336	22,050
Rat	40,459	28,736	22,263
Fly	34,740	30,353	13,918
Nematode	58,941	31,574	20,362
Yeast	7,126	6,692	6,692
Total	462,331	231,749	105,246

Table 1: The numbers of transcripts and protein coding genes.

Enrichment analyses for host genes

We conducted enrichment analyses for 2,144 non-orthologous host genes out of 19,961 coding genes in human and 1,087 non-orthologous host genes out of 22,050 coding genes in mouse by GO terms. Figure 4 shows results of the enrichment analyses. We can find

Ortholog						
Gene	**Human**	**Mouse**	**Rat**	**Fly**	**Nematode**	**Yeast**
Human gene	—	18,023	17,687	10,225	8,292	4,922
Mouse gene	18,425	—	19,926	10,254	8,308	4,905
Rat gene	18,621	20,192	—	10,755	8,841	5,419
Fly gene	7,318	7,319	7,272	—	5,507	3,668
Nematode gene	6,285	6,308	6,273	5,859	—	3,618
Yeast gene	2,363	2,367	2,356	2,234	2,199	—

Table 2: The numbers of orthologous genes.

Organism	Intergenic	Intronic	Sense	Total
Human	29,242	3,886	756	33,884
Mouse	15,545	2,267	176	17,988
Rat	7,950	1,204	43	9,197
Fly	3,190	806	47	4,043
Nematode	20,697	4,441	57	25,195
Yeast	391	13	8	412

Table 3: The numbers of intergenic, intronic and sense ncRNA genes.

that, for example, Figure 4A includes cellular component organization-, neuron development-, regulation of GTPase- and modification-related terms. Figure 4B includes cell morphogenesis-, glutamate receptor- and splicing-related terms. Figure 4C shows intracelullar-related terms. Figure 4D shows synapse- and lumen-related terms. Figure 4E shows GTP-related terms. Figure 4F includes glutamate receptor- or channel-related terms. The identical GO terms between human and mouse are 'postsynaptic density' and 'cell junction'.

Meanwhile, we also conducted enrichment analyses for 547 orthologous host genes out of 19,961 genes in human and 546 non-orthologous host genes out of 22,050 genes in mouse by GO terms. Figure 5 shows results of the enrichment analyses. Figure 5A includes neuron-related terms. Figure 5B includes cellular component organization-, neuron-, cytoskeleton-related terms. Figure 5C and 5D show junction-related terms. Figure 5E and 5F include binding-related terms. Except for transport-related terms in Figure 5A, the detected GO terms are identical or similar between human and mouse.

Similarity searches to detect conserved intronic *miRNA* genes

As shown in Table 4, our database stores 771 and 959 intronic *miRNA* genes in human and mouse, respectively. We conducted 771 × 959 (739,389) BLAST searches by setting the BLAST query as a human gene and subject as a mouse gene. The BLAST searches found 125 hits whose e-value is less than 10-14 as shown in Figure 6A. In the 125 BLAST hits, means of sequence identities and alignment lengths were 93.76% and 79, respectively. The 111 BLAST hits are intronic miRNA genes located in orthologous host genes between human and mouse. Figure 6B shows that the number of such orthologous host genes possessing intronic *miRNA* genes detected by the BLAST searches is 85 in human and mouse. On the other hand, the numbers of non-orthologous host genes are 6 and 5 in human and mouse, respectively. Meanwhile, out of the 111 hits, Table 5 shows 26 hits whose genes are located on X-chromosome. These 26 hits are hits regarding intronic *miRNA* genes located in 9 host genes which are orthologous between human and mouse.

We conducted enrichment analyses for 85 orthologous host genes possessing conserved *miRNA* genes out of 547 orthologous host genes in human and 85 out of 546 orthologous host genes in mouse. Figure 7 shows results of the enrichment analyses by GO terms. Figure 7A and 7B include process-related terms. This shows that many GO terms detected by the GSEA are identical or similar between human and mouse. On the other hand, Table 6 shows 14 hits whose host genes are not orthologous. Table 6 contains unique 11 host genes in human and mouse. GO terms associated with the 11 host genes in human and mouse were shown in Figures 8 and 9, respectively. This shows that some host genes have identical or similar GO terms in the pairs of host genes possessing miRNA genes conserved within human and mouse. For instance, Figures 8A and 9A show 4 identical GO terms. Figures 8B and 9B show some identical GO terms such as 'canonical Wnt signaling pathway' and 'ventricular cardiac muscle tissue morphogenesis'.

We conducted remained combinations of BLAST searches by using the intronic miRNA genes in 6 organisms shown in Table 4. Figure 10A and 10B show results of BLAST searches in human versus rat. Figures 10C and 10D show results of BLAST searches in mouse versus rat. Other combinations of organisms such as human versus fly or mouse versus nematode were not detected. Figure 10A shows that the BLAST searches found 110 hits whose e-value is less than 10-14. The 98 BLAST hits are intronic miRNA genes located in orthologous host genes between human and rat. Figure 10B shows that the number of such orthologous host genes possessing intronic miRNA genes detected by the BLAST searches is 66 in human and rat. The numbers of non-orthologous host genes are 5 and 4 in human and rat, respectively. Figure 10C shows that the BLAST searches found 970 hits whose e-value is less than 10-14. The 322 BLAST hits are intronic miRNA genes located in orthologous host genes between mouse and rat. Figure 10D shows that the numbers of such orthologous host genes possessing intronic miRNA genes detected by the BLAST searches are 232 and 233 in mouse and rat, respectively. The numbers of non-orthologous host genes are 63 and 66 in mouse and rat, respectively. These results show that the number of BLAST hits in mouse versus rat is the largest in the BLAST searches. In addition, by integrating Figure 6B, Figure 10B and 10D, we found 64 orthologous host genes possessing intronic miRNA genes among human, mouse and rat as shown in Figure 11.

Discussion

Our database is created from the Ensembl data. The Ensembl database can identify gene locations in genomic sequences. However, it is not categorized by location information among protein-coding and noncoding genes. Our database categorizes *ncRNA* genes based on genomic positions and can easily identify host genes of intronic *ncRNA* s. Our database stores all transcripts in Ensembl. The transcripts include protein-coding transcripts. Some protein-coding transcripts contain a 5′-UTR (untranslated region) and 3′-UTR but some transcripts are not. Therefore, some protein-coding transcripts only contain the coding regions. Our database stores all *ncRNA* genes in Ensembl. We assigned a category of *ncRNA* genes based on gene locations in genomic sequences. In intergenic, intronic and sense *ncRNA* genes, we focused on intronic *ncRNA* genes. Firstly, we investigated functions of *ncRNA* genes. Functions of some intronic *ncRNA* genes are known as shown in Table 4. However, functions of many *ncRNA* genes are unknown because they are annotated as li*ncRNA*, misc RNA, *ncRNA* and so on. This indicates that the number of *ncRNA* genes in each biotype may increase in the future.

We next focused on host genes of the intronic *ncRNA* genes. As

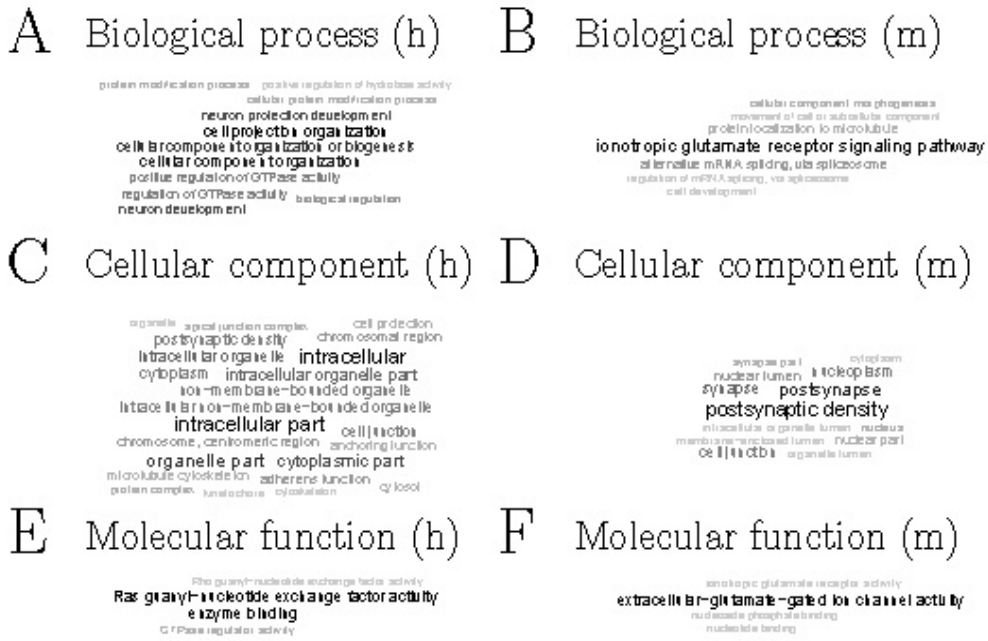

Figure 4: Gene set enrichment analyses of non-orthologous host genes by GO terms. The GO terms show the results of gene set enrichment analyses for 2,144 and 1,087 non-orthologous genes in human and mouse host genes, respectively. The p-value of each GO term is less than 10^{-4}. The character size of each GO term depends on the p-value: the larger the size is, the smaller the p-value is.

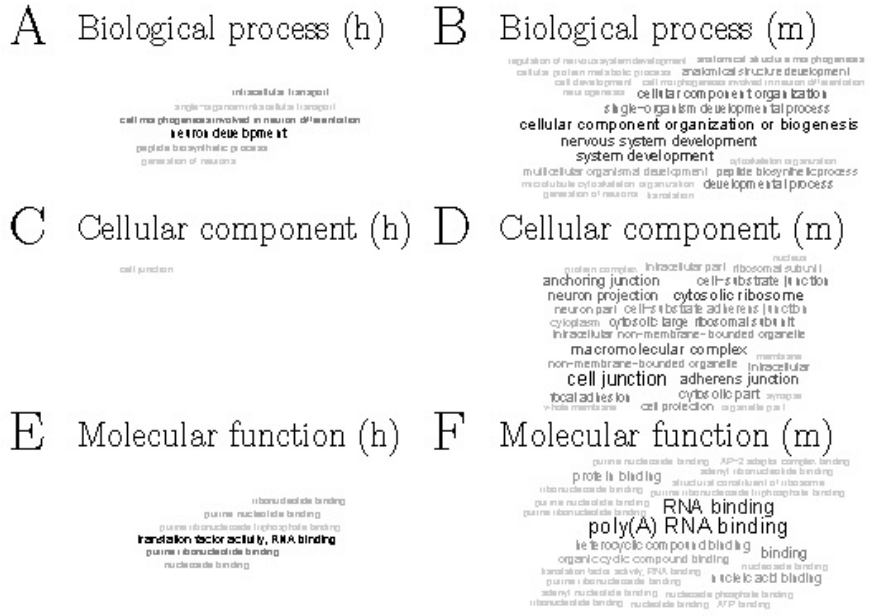

Figure 5: Gene set enrichment analyses of orthologous host genes by GO terms. The GO terms show the results of gene set enrichment analyses for 547 and 546 orthologous genes in human and mouse host genes, respectively. The p-value of each GO term is less than 10^{-4}. The character size of each GO term depends on the p-value: the larger the size is, the smaller the p-value is.

shown in Figure 3, the numbers of host genes are 2,691 and 1,633 in human and mouse, respectively. This indicates that the number of host genes in mouse is fewer than human. In addition, Table 3 shows that the total number of *ncRNA* genes in human or mouse is 33,884 or 17,988, respectively. This indicates that the total number of *ncRNA* genes in mouse is fewer than human. These results indicate that the number of intronic *ncRNA* genes in mouse may increase in the future. We divided the host genes by using orthologous relationships. We then conducted GSEA in order to investigate what genes are contained in the host genes, These GSEA show that some non-orthologous genes tend to have different functions but orthologous genes have similar functions. Moreover, in order to identify host genes associated with diseases,

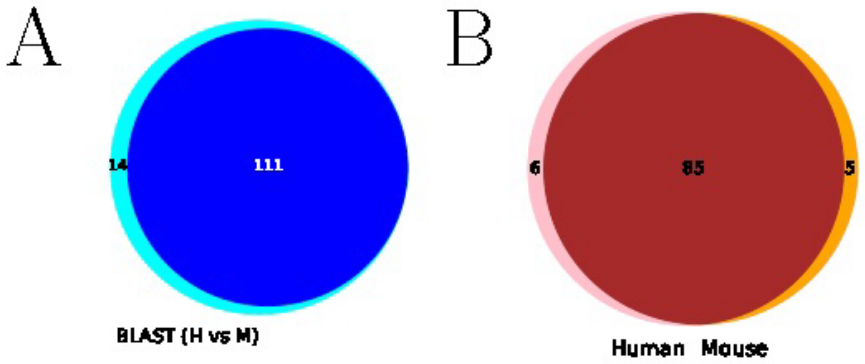

Figure 6: Results of BLAST searches in human vs. mouse. (A) BLAST hits of intronic miRNA genes possessed by orthologous host genes in all BLAST hits in human vs. mouse. (B) Orthologous host genes in all host genes possessing intronic miRNA genes detected by the BLAST searches in human vs. mouse.

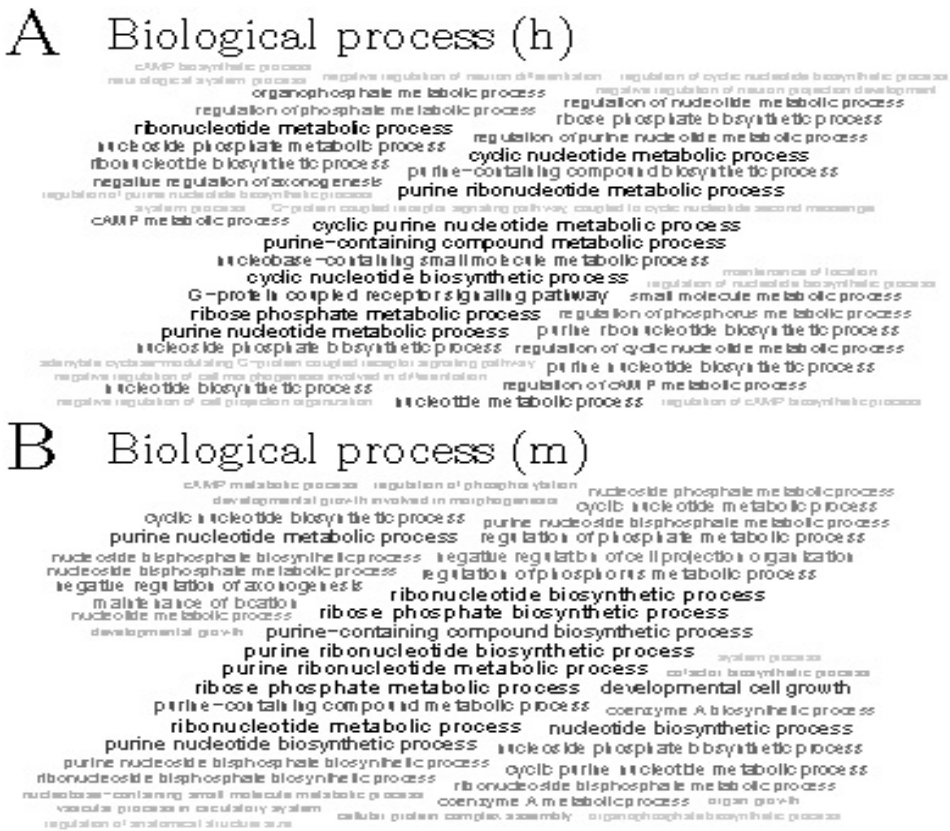

Figure 7: Gene set enrichment analyses of orthologous host genes possessing intronic miRNA genes by GO terms. The GO terms show the results of gene set enrichment analyses for 85 and 85 orthologous genes in human and mouse host genes, respectively. The p-value of each GO term is less than 0.01. The character size of each GO term depends on the p-value: the larger the size is, the smaller the p-value is.

we explored from our database which host genes are associated with diseases. Because the GSEA show that host genes are associated with some neuron-related terms, we searched host genes associated with Alzheimer's disease. We found that approximately 5% of host genes are associated with diseases. Moreover, 6 human host genes are involved in Alzheimer's disease. For example, a human gene (ENSG00000182240) is annotated by a GO term (GO:0050435, beta-amyloid metabolic process) and, therefore, it is involved in a term concerning Alzheimer's disease. Additionally, this human gene is a host gene of an *ncRNA* (ENST00000458830, snoRNA). This shows that our database is useful to identify *ncRNA* genes associated with diseases. Furthermore, 4 mouse host genes are involved in Alzheimer's disease. This shows that our database may guide to investigate relationship among gene evolutions and diseases in future research.

We conducted BLAST searches in order to investigate conservation

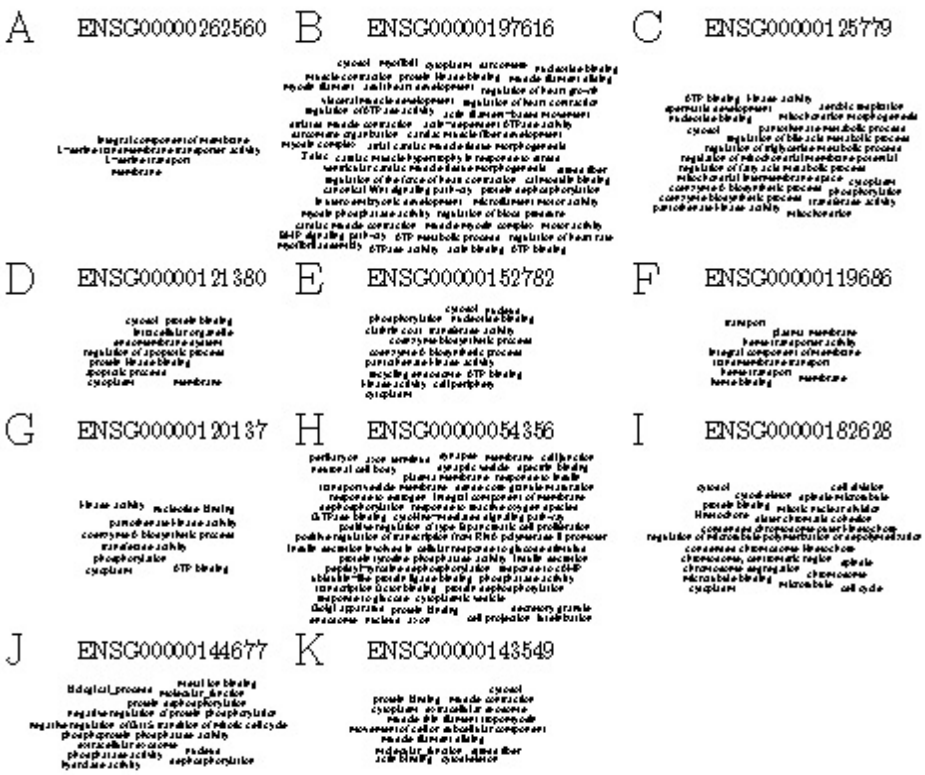

Figure 8: GO terms associated with a human gene.

Figure 9: GO terms associated with a mouse gene.

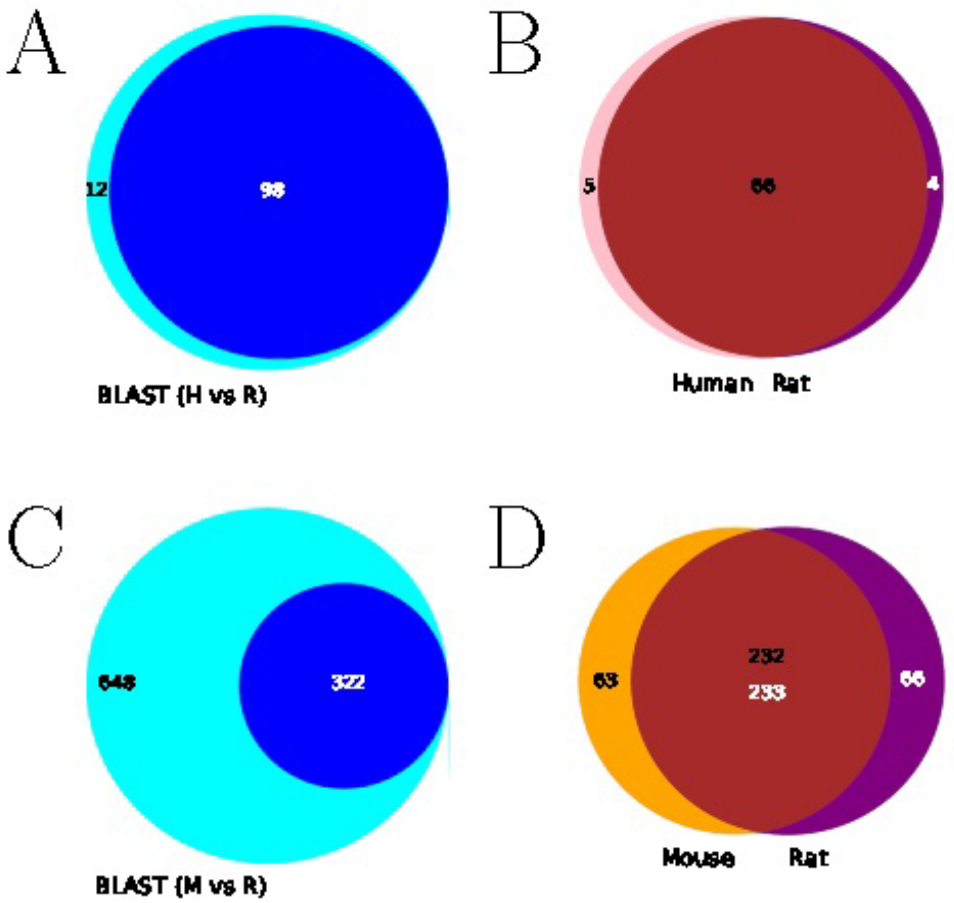

Figure 10: Results of BLAST searches in human vs. rat and mouse vs. rat. (A) BLAST hits of intronic miRNA genes possessed by orthologous host genes in all BLAST hits in human vs. rat. (B) Orthologous host genes in all host genes possessing intronic miRNA genes detected by the BLAST searches in human vs. rat. (C) BLAST hits of intronic *miRNA* genes possessed by orthologous host genes in all BLAST hits in mouse vs. rat. (D) Orthologous host genes in all host genes possessing intronic miRNA genes detected by the BLAST searches in mouse vs. rat.

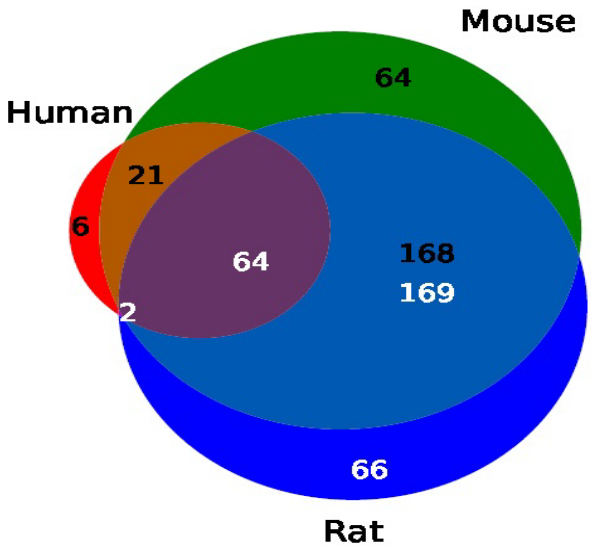

Figure 11: Orthologous host genes possessing intronic miRNA genes conserved within human, mouse and rat.

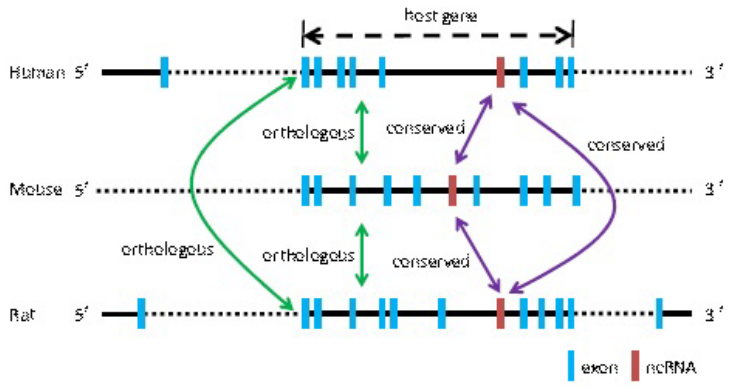

Figure 12: A schematic view of orthlogous host genes possessing intronic *ncRNA* genes conserved within human, mouse and rat.

Transcript Biotype	Human	Mouse	Rat	Fly
3'overlapping ncRNA	11	0	0	0
antisense	176	56	1	0
bidirectional promoter lncRNA	0	10	0	0
lincRNA	285	73	16	225
miRNA	771	959	452	168
misc RNA	642	116	76	0
Non-coding	2	0	0	0
ncRNA	0	0	0	0
piRNA	0	0	0	0
Pre-miRNA	0	0	0	123
processed transcript	39	46	2	0
pseudogene	4	0	0	0
rRNA	115	71	50	0
retained intron	32	14	0	0
ribozyme	2	3	2	0
sRNA	0	1	0	0
scaRNA	21	28	24	0
Sense intronic	899	279	10	0
sense overlapping	47	2	0	0
snRNA	489	259	209	10
snoRNA	351	348	362	224
TEC	0	2	0	0
tRNA	0	0	0	56

Table 4: Biotypes of intronic ncRNA genes.

of *ncRNA* genes in host genes. However, our database stores a variety of *ncRNA* genes. Therefore, comparisons of all *ncRNA* genes are very high computational costs. We then focused on miRNA genes because miRNA genes are relatively short genes than l*ncRNA* genes and so on. We conducted BLAST searches to all combinations of intronic *miRNA* genes in human and mouse. The BLAST hits show high sequence identities between intronic *miRNA* genes. This shows that some intronic miRNA genes are conserved in human and mouse. In addition, the most of BLAST hits are intronic miRNA genes located in orthologous host genes between human and mouse as shown in Figure 6A. This shows that many orthologous host genes possess conserved miRNA genes in their intronic regions. We then conducted GSEA in order to investigate what orthologous host genes possess the intronic *miRNA* genes conserved within human and mouse. The results show that such orthologous host genes are responsible for some processes such as biosynthetic process and metabolic process. On the other hand, the BLAST searches also found some host genes possessing conserved intronic miRNA genes but they are not orthologous within human and mouse. As shown in Figures 8 and 9), some these host genes are responsible for similar functions. This shows that host genes possessing conserved intronic miRNA genes have similar functions even if they are not orthologous.

In addition, we investigated pairs of miRNA genes between remained combinations of the 6 organisms. However, yeast is excluded because it does not have the data of intronic miRNA genes as shown in Figure 4.

miRNA (human)	miRNA (mouse)	Ident	Len	E-val	Host (human)	Host (mouse)
ENST00000385083	ENSMUST00000104655	90.62	64	2.00E-22	ENSG00000011677	ENSMUSG00000031343
ENST00000390228	ENSMUST00000103256	87.76	98	6.00E-31	ENSG00000011677	ENSMUSG00000031343
ENST00000385222	ENSMUST00000104655	87.5	64	5.00E-19	ENSG00000011677	ENSMUSG00000031343
ENST00000385277	ENSMUST00000083668	100	83	2.00E-43	ENSG00000086758	ENSMUSG00000025261
ENST00000606724	ENSMUST00000083602	93.52	108	2.00E-45	ENSG00000086758	ENSMUSG00000025261
ENST00000384901	ENSMUST00000093573	96.43	84	6.00E-39	ENSG00000101974	ENSMUSG00000062949
ENST00000385065	ENSMUST00000093602	96.1	77	4.00E-35	ENSG00000129682	ENSMUSG00000031137
ENST00000410389	ENSMUST00000175298	88.75	80	9.00E-27	ENSG00000147246	ENSMUSG00000041380
ENST00000616374	ENSMUST00000175381	94.2	69	5.00E-29	ENSG00000147246	ENSMUSG00000041380
ENST00000390811	ENSMUST00000103259	90.7	86	2.00E-30	ENSG00000147246	ENSMUSG00000041380
ENST00000410783	ENSMUST00000122764	93.06	72	9.00E-29	ENSG00000147246	ENSMUSG00000041380
ENST00000408783	ENSMUST00000116681	92.78	97	9.00E-39	ENSG00000147246	ENSMUSG00000041380
ENST00000362131	ENSMUST00000083516	96.3	108	8.00E-50	ENSG00000147246	ENSMUSG00000041380
ENST00000385278	ENSMUST00000102057	94.51	91	2.00E-39	ENSG00000157600	ENSMUSG00000047045
ENST00000390702	ENSMUST00000102443	91.04	67	1.00E-24	ENSG00000158813	ENSMUSG00000059327
ENST00000390204	ENSMUST00000093600	83.75	80	6.00E-19	ENSG00000171365	ENSMUSG00000004317
ENST00000390204	ENSMUST00000102296	88.37	86	3.00E-27	ENSG00000171365	ENSMUSG00000004317
ENST00000385051	ENSMUST00000093600	89.61	77	4.00E-26	ENSG00000171365	ENSMUSG00000004317
ENST00000385051	ENSMUST00000102296	86.15	65	1.00E-17	ENSG00000171365	ENSMUSG00000004317
ENST00000458843	ENSMUST00000093600	85.51	69	2.00E-18	ENSG00000171365	ENSMUSG00000004317
ENST00000458843	ENSMUST00000102296	84.85	66	4.00E-16	ENSG00000171365	ENSMUSG00000004317
ENST00000385034	ENSMUST00000083464	98.53	68	2.00E-33	ENSG00000171365	ENSMUSG00000004317
ENST00000385025	ENSMUST00000093630	94.25	87	3.00E-37	ENSG00000171365	ENSMUSG00000004317
ENST00000606349	ENSMUST00000093630	93.02	86	2.00E-34	ENSG00000171365	ENSMUSG00000004317
ENST00000385280	ENSMUST00000093592	93.55	62	3.00E-25	ENSG00000171365	ENSMUSG00000004317
ENST00000362181	ENSMUST00000083576	91.55	71	9.00E-26	ENSG00000188419	ENSMUSG00000025531

Table 5: BLAST hits between a pair of intronic miRNA genes in X-chromosome. The sequence identity (Ident), alignment length (Len) and e-value (E-val) show the BLAST searching result between the pair of miRNA genes. The host gene is a coding gene including the miRNA gene. This table only includes hits that genes are located on X-chromosome.

miRNA (human)	miRNA (mouse)	Ident	Len	E-val	Host (human)	Host (mouse)
ENST00000408865	ENSMUST00000116724	100	101	3.00E-53	ENSG00000262560	ENSMUSG00000046110
ENST00000362287	ENSMUST00000083498	95.77	71	8.00E-32	ENSG00000197616	ENSMUSG00000040752
ENST00000362154	ENSMUST00000083660	91.67	72	3.00E-27	ENSG00000125779	ENSMUSG00000033610
ENST00000636813	ENSMUST00000116835	87.64	89	1.00E-26	ENSG00000121380	ENSMUSG00000014232
ENST00000362127	ENSMUST00000083629	91.43	70	3.00E-26	ENSG00000152782	ENSMUSG00000037514
ENST00000625482	ENSMUST00000198352	100	50	2.00E-25	ENSG00000119686	ENSMUSG00000061080
ENST00000362154	ENSMUST00000083619	91.94	62	2.00E-22	ENSG00000125779	ENSMUSG00000018846
ENST00000362165	ENSMUST00000083629	91.94	62	2.00E-22	ENSG00000120137	ENSMUSG00000037514
ENST00000384914	ENSMUST00000083604	88.41	69	1.00E-20	ENSG00000054356	ENSMUSG00000056553
ENST00000362127	ENSMUST00000083619	87.88	66	2.00E-19	ENSG00000152782	ENSMUSG00000018846
ENST00000362165	ENSMUST00000083660	87.88	66	2.00E-19	ENSG00000120137	ENSMUSG00000033610
ENST00000385261	ENSMUST00000102333	84.85	66	1.00E-16	ENSG00000182628	ENSMUSG00000049916
ENST00000362205	ENSMUST00000083496	86.44	59	4.00E-15	ENSG00000144677	ENSMUSG00000078429
ENST00000401119	ENSMUST00000102424	82.86	70	1.00E-14	ENSG00000143549	ENSMUSG00000052698

Table 6: BLAST hits between a pair of intronic miRNA genes among non-orthologous host genes. The sequence identity (Ident), alignment length (Len) and e-value (E-val) show the BLAST searching result between the pair of miRNA genes. The host gene is a coding gene including the miRNA gene. This table only includes hits that the pair of host genes are not orthologous.

The BLAST searches found hits of human versus rat and mouse versus rat as shown in Figure 10. This shows that the number of host genes possessing conserved miRNA genes in the intronic regions is larger as the organisms have a close evolutionary relationship because the

number of hits in mouse versus rat is the largest. In addition, we found 64 orthologous host genes possessing intronic miRNA genes conserved within human, mouse and rat as shown in Figure 11. This indicates that an orthologous host gene possesses an intronic miRNA gene conserved within human, mouse and rat. Figure 12 shows a schematic view of such a relationship. This indicates that host genes have retained *miRNA* genes in the intronic regions within human, mouse and rat. On the other hand, host genes possess conserved intronic *miRNA* genes only among human, mouse and rat. In other words, the combinations regarding fly or nematode were not found by the BLAST searches. This suggests some possibilities. One possibility is that the host gene lost the *miRNA* gene from the intronic region in the evolutionary process. Another is that the intronic *miRNA* gene emerged in the intronic region of the host gene on the way of the evolutionary process.

By the use of the database, we find that orthologous protein-coding genes have retained intronic *miRNA* genes in the evolutionary process. Therefore, our database is useful to identify relationships among protein-coding and noncoding genes. Meanwhile, we have not clarified that other kinds of *ncRNA* s such as l*ncRNA* s have been retained in the host genes. Because our database stores not only *miRNA* genes but also the information concerning such *ncRNA* s, it should be necessary to investigate whether an intronic *ncRNA* gene is conserved among organisms and is located within orthologous protein-coding genes. In addition, we plan to associate l*ncRNA* s with diseases such as cancers and store such information into our database. This database is useful to identify *lncRNA* s involved in diseases.

Conclusion

In this study, we discussed evolutions of intronic *ncRNA* genes based on gene locations in genome sequences. We found that protein-coding genes which are orthologous between human, mouse and rat possess miRNA genes conserved within them in the intronic regions. This result suggests that some orthologous genes have retained *ncRNA* genes in their intronic regions in the evolutionary process.

References

1. Gilbert W (1978) Why genes in pieces. Nature 271: 501.
2. Orphanides G, Reinberg D (2002) A unified theory of gene expression. Cell 108: 439-451.
3. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157: 77-94.
4. Santosh B, Varshney A, Yadava PK (2015) Non-coding RNAs: Biological functions and applications. Cell Biochem Funct 33: 14-22.
5. Hoskins AA, Moore MJ (2012) The spliceosome: A flexible, reversible macromolecular machine. Trends Biochem Sci 37: 179-188.
6. Noller H (1991) Ribosomal-RNA and translation. Annu Rev Biochem 60: 191-227.
7. Holley CL, Topkara VK (2011) An introduction to small non-coding RNAs: miRNA and snoRNA. Cardiovasc Drugs Ther 25: 151-159.
8. Reddy PH, Tonic S, Kumar S, Vijayan M, Kandimalla R, et al. (2017) A critical evaluation of neuroprotective and neurodegenerative MicroRNAs in Alzheimer's disease. Biochem Biophys Res Commun 483: 1156-1165.
9. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Mol Cell 43: 904-914.
10. Jing F, Jin H, Mao Y, Li Y, Ding Y, et al. (2017) Genome-wide analysis of long non-coding RNA expression and function in colorectal cancer. Tumour Biol.
11. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, et al. (2008) An epigenetic role for maternally inherited piRNAs in transposon Silencing. Science 322: 1387-1392.
12. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) Ensembl compara gene trees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327-335.
13. Pignatelli M, Vilella AJ, Mu-ato M, Gordon L, White S, et al. (2016) ncRNA orthologies in the vertebrate lineage. Database (Oxford) 2016: bav127.
14. Nitsche A, Stadler PF (2017) Evolutionary clues in lncRNAs. Wiley Interdiscip Rev RNA.
15. Rodriguez A, Gri-ths-Jones S, Ashurst J, Bradley A (2004) Identification of mammalian mi-croRNA host genes and transcription units. Genome Res 14: 1902-1910.
16. Baskerville S, Bartel D (2005) Microarray profiling of microRNAs reveals frequent co-expression with neighboring miRNAs and host genes. RNA 11: 241-247.
17. Ramalingam P, Palanichamy JK, Singh A, Das P, Bhagat M, et al. (2014) Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing. RNA 20: 76-87.
18. Louro R, Smirnova AS, Verjovski-Almeida S (2009) Long intronic noncoding RNA transcription: Expression noise or expression choice? Genomics 93: 291-298.
19. St Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, et al. (2012) Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. BMC Genomics 13: 504.
20. Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, et al. (2013) Circular intronic long noncoding RNAs. Mol Cell 51: 792-806.
21. Osman I, Tay M, Pek J (2016) Stable intronic sequence RNAs (sisRNAs): A new layer of gene regulation. Cell Mol Life Sci: 1-13.
22. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic Acids Res 30: 38-41.
23. Kinsella RJ, Kaehaeri A, Haider S, Zamora J, Proctor G, et al. (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database (Oxford) 2011: bar030.
24. Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. (2000) Gene ontology: Tool for the unification of biology. Nat Genet 25: 25-29.
25. Falcon S, Gentleman R (2007) Using GO stats to test gene lists for GO term association. Bioinformatics 23: 257-258.
26. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
27. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10: 421.