Chinese Science Bulletin 2005 Vol. 50 No. 11 1126—1130

Origin and evolution of new exons in the rodent zinc finger protein 39 gene

PENG Lixin[1,2*], ZHENG Hongkun[3*], LI Xin[1,2], YANG Shuang[2,4], CHEN Hong[1,5] & WANG Wen[2]

1. College of Animal Science and Technology, Northwest Sci-Tech University of Agriculture and Forestry, Yangling 712100, China;
2. CAS-Max Planck Junior Scientist Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China;
3. Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China;
4. Graduate School of the Chinese Academy of Sciences, Beijing 100039, China;
5. Institute of Cellular and Molecular Biology, Xuzhou Normal University, Xuzhou 221116, China

Correspondence should be addressed to Wang Wen (email: wwang@mail.kiz.ac.cn)

Abstract **The origin of new structures and functions is an important process in evolution. In the past decades, we have obtained some preliminary knowledge of the origin and evolution of new genes. However, as the basic unit of genes, the origin and evolution of exons remain unclear. Because young exons retain the footprints of origination, they can be good materials for studying origin and evolution of new exons. In this paper, we report two young exons in a zinc finger protein gene of rodents. Since they are unique sequences in mouse and rat genome and no homologous sequences were found in the orthologous genes of human and pig, the young exons might originate after the divergence of primates and rodents through exonization of intronic sequences. Strong positive selection was detected in the new exons between mouse and rat, suggesting that these exons have undergone significant functional divergence after the separation of the two species. On the other hand, population genetics data of mouse demonstrate that the new exons have been subject to functional constraint, indicating an important function of the new exons in mouse. Functional analyses suggest that these new exons encode a nuclear localization signal peptide, which may mediate new ways of nuclear protein transport. To our knowledge, this is the first example of the origin and evolution of young exons.**

Keywords: new exons, origin and evolution, zinc finger protein, nuclear localization signal peptide.

DOI: 10.1360/982004-743

With rapid accumulation of genomic data, we have known that there are great gene number variations among organisms[1]. In the past decades, we have started to understand the molecular mechanisms of new gene origination and their evolutionary process[1,2]. However, as the basic unit of gene structure, exon's origin and evolution remain largely unclear. It has been proposed that the following mechanisms are possibly involved in the creation of new exons: exaptation of transposable elements[3], retroposition[4], exon duplication[5] and exonization of intronic sequences[6]. In 1994, Makalowski et al.[7] first reported that the *DAF* gene of human contains an Alu sequence. Recently, Nekrutenko and Li[3] found that in human protein-coding genes, 4% of the exons are possibly created by transposable elements. For exon duplication, Letunic et al.[5] estimated that 10% of genes include duplicated exons in human. Since the genomes of eukaryotes have high proportional intronic sequence, it is conceivable that exonization of intronic sequences could be an important mechanism for creating new exons. In deed using bacteria and yeast as outgroups, Kondrashov et al.[8] identified a number of exons in eukaryotes originating from non-coding intronic sequences.

As we have demonstrated in the studies of origin of new genes[9,10], in order to understand origin and evolution of new exons in detail it is necessary to identify young exons first. Here we shall report the first example of young exons identified in the zinc finger protein 39 gene (*zfp39*) in rodents. *zfp39* belongs to KRAB-*zfp* gene family that can be divided into four sub-families according to different KRAB domains[11,12]. These genes code for transcription repression factors that have important function in gene regulation network[13]. We analyzed the origination, evolution and function of these young exons in detail.

1 Materials and methods

1.1 Identification of new exons

To get the orthologous gene of KRAB-*zfp* in human, mouse and rat, we performed the BLASTP search in the Refseq datatbase (http://www.ncbi.nlm.nih.gov/RefSeq/) by best reciprocal BLAST hit principle[14]. The best reciprocal BLAST hit principle is to blast the genomic sequence of specie Y using gene A in specie X as query sequence. If gene B in specie Y is the best match of gene X, gene B is used to blast the genomic sequence of specie X. If gene A is also identified as the best match of gene B, then gene A and gene B are called the best reciprocal hits, and they are considered to be orthologs. Comparing the structures of the three orthologous genes, we found two extra exons in the two rodent species. These two exons have no homologous sequences either in the ESTs of pig, suggesting that the two exons are newly evolved young exons in rodents.

1.2 Population samples of mouse

Thirteen mice (*Mus musculus*) collected from Shannxi (3), Henan (1), Shanxi (3), Hubei (3), Yunnan (3) provinces of China were kindly provided by Prof. Zheng and Dr. Jing of Kunming Institute of Zoology.

* These authors contributed equally to this work.

1.3 PCR and sequencing

Genomic DNA was extracted from mouse liver with the Gentra DNA isolation Kit (Gentra system Inc.). Primers were designed for the second new exon of mouse from the published sequences of *zfp39* in mouse (GenBank accession number: NM_011758). Forward primer is 5′-GAACGGAGTGCCAGCTCATC-3′ and reverse primer is 5'-TTTCCCCTAACATCGTAAAATCTC-3′ (Shanghai Sangon). The following primers were used to amplify the old fifth exon in this gene: forward primer is 5′-ACGAAACCAGCCAGGACAAT-3′ and reverse primer is 5′-GACATTCAGGTCCGACTTACAGTAG-3′. PCRs were performed in 50 μL volume containing 5 μL 10× Buffer, 1 μL 10 pmol/μL primer, 4 μL 2.5 nmol/μL dNTP, and 0.25 μL 100 U/μL r*Taq* DNA Polymerase. Amplification was performed for 35 cycles (30 s at 94℃, 1 min at 56－58℃, 1 min at 72℃). PCR products were purified with Promega purification kit (Promega Corp.). All PCR products were sequenced with the BigDye Terminator Cycle Sequencing Ready Reaction kit (ABi Inc.) using PCR primers.

1.4 Data analyses

Using the program Seqman 5.0 in DNAStar, we assembled and manually checked the obtained sequences. With MEGA 3.0[15] we calculated nonsynonymous substitution rate (*Ka*) and synonymous substitution rate (*Ks*) between mouse and rat sequences using Modified Nei-Gojobori method and performed *Z* test to test whether *Ka*/*Ks* ratio is significantly apart from 1 (the neutrality). *Ka*/*Ks* significantly being greater than 1 indicates that the exon is subject to strong positive selection; while *Ka*/*Ks* being less than 1 suggests negative selection. Nucleotide diversity (π) was estimated with DnaSP 3.99[16]. Tajima's D Test[17], Fu and Li's Test[18] were also performed with DnaSP 3.99[16] to test whether the sequences within population have been evolving under neutrality. These two tests are performed to detect whether a certain locus evolves under nature selection using its population data. The general principle is to estimate the population genetics parameter $\theta(\theta = 4Ne\mu$, Ne is the effective population size, and μ is the mutation rate of nucleotide per sequence per generation) using the segregation site numbers and the average numbers of nucleotide difference between two sequences randomly selected within population, respectively. If the two estimators are equal, then the locus evolves neutrally; if they differ from each other significantly, then the locus evolves selectively[19].

2 Results and discussion

2.1 The structure and origin of new exons

The *zfp39* and its orthologous genes are annotated as zinc finger protein genes (Table 1). In the KRAB-*zfp* gene family, KRAB and zinc finger protein domains are often encoded by three exons[11]. The *zfp39* and its orthologous genes are similar to this structure (Fig. 1). Comparing the orthologous genes in human, we see that there are two extra exons at the 5′ terminal of the rodent *zfp39*. Exon1 is

Table 1 *zfp39* and its orthologous genes in mouse, rat and human

Species	Gene feature	Accession number	Chromosome location	Number of exons	Number of alternative splicing forms
Homo sapiens	Kruppel like zinc finger factor X17	XM_496648	3	3	1
Mus musculus	Zinc Finger protein 39	NM_011758	11	5	2
Rattus norvegicus	similar to Zinc Finger protein 11b (kox2)	XM_220504	10	5	2

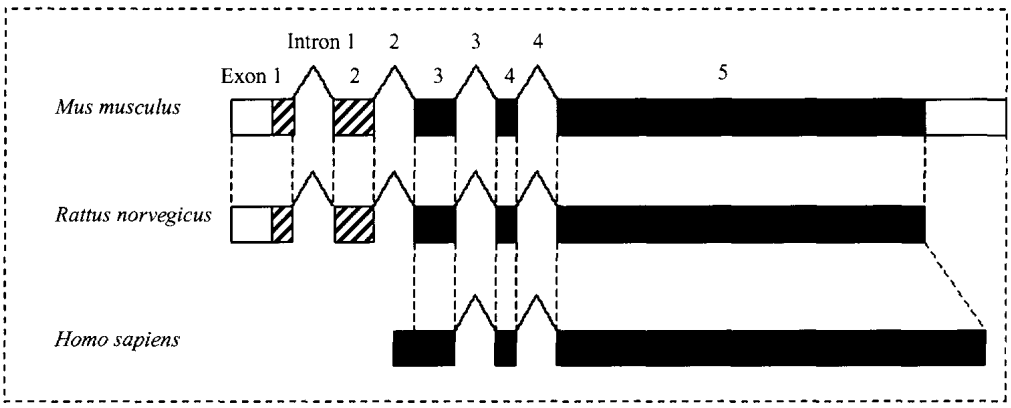

Fig. 1. Gene structures of the *zfp39* gene and its orthologous exons in human and rat. The orthologous exons in mouse, rat and human are shown with black boxes; blank boxes represent the UTR; hatched boxes indicate the new exons in rodents. Exons and introns are marked by number index respectively.

112 bp which only encodes the initiation amino acid, and Exon2 is 165 bp which encodes a peptide of 55 amino acids. Both the new exons have intact splicing sites. Since they are unique sequences in the rodent genomes and no homologous sequences were found in human and pig, we speculate that the young exons must have originated after the divergence of primates and rodents (80 Ma ago)[20] and before the divergence of mouse and rat (16 Ma ago)[21] probably through exonization of intronic sequences. The Repeatmasker program (http://repeatmasker.genome.washington.edu/) was used to search the repeat sequences in this gene. The repeat sequences amount to 34.69% and 48.76% in introns of mouse and rat *zfp39* gene, respectively. This high proportion of repeat sequences indicates that the locus may be an insertion hot spot of repeat sequences. Therefore we cannot rule out the possibility that transposable elements may be involved in forming these new exons. Unfortunately we were unable to amplify the orthologous exon in other rodents out of mouse and rat because of too distant divergence. To reveal the detailed origination process we need to clone *zfp39* genes in other primitive rodents in the future.

2.2 Subsequent evolution of new exons

Although many studies on new gene origination revealed that positive selection is a common phenomenon during evolution of new genes[1,2], whether the origin of new exons is also subject to positive selection remains unclear. To shed light on the evolution of new exons in *zfp39*, we calculated the *Ka* and *Ks* and performed *Z* test on different exons between mouse and rat (Table 2).

Table 2 *Ka*, *Ks* and *Ka/Ks* values on exons between mouse and rat

Exon	Coding length/bp	Ka	Ks	Ka/Ks
Exon1	3	—	—	—
Exon2	165	0.180	0.068	2.647[a)]
Exon3	127	0.034	0.029	1.172
Exon4	93	0.090	0.154	0.584
Exon5	1765	0.099	0.230	0.430

a) $p< 0.05$.

Because Exon1 only encodes initiation amino acid, we calculated *Ka/Ks* values of the other four exons. For Exon2, the *Ka* is significantly higher than the *Ks* at the 0.05 level, which suggests that this exon was subject to positive selection and functional divergence between mouse and rat. We further analyzed the population data of this new exon in mouse. First, we performed Tajima's D, Fu and Li's test on this exon and its flanking intronic sequences. Because of geographic subdivision in mouse population, we divided this population into north and south populations (Table 3). The Tajima's D, Fu and Li's test did not significantly deviate from neutral expectations, indicating that mutations within species were selectively neutral. We also calculated the average nonsynonymous substitution rate and average synonymous substitution rate on this exon. The *Ka* (0.006) was lower than the *Ks* (0.010). But because there were only two polymorphism sites on the new exon, it was difficult to get a meaningful result. Then we compared the nucleotide diversities (π) between the new exon and introns. The π value in introns (0.00735) was significantly higher than that in Exon2 (0.00179, $p<0.001$). π value (0.00373) in the partial sequences of Exon5 was significantly lower than that in the introns ($p<0.001$) and significantly higher than that in Exon2 ($p<0.05$). These results demonstrate that both Exon2 and Exon5 are under strong negative selection. And by comparing to the π value of Exon5, a stronger functional constraint is detected in Exon2, implying that this exon has an important function.

Table 3 Tajima's D, Fu and Li's test in mouse population

	Tajima's D	Fu and Li's D	Fu and Li's F
South population	–0.9493	–0.6758	–0.9017
North population	–0.9081	0.1918	–0.2175

2.3 Function of new exons

The above results suggest that these new exons have undergone functional constraint and functional divergence after separation of mouse and rat. We want to know what function of these new exons had, and what role they played in the functional divergence of *zfp39* between mouse and rat. We analyzed the hypothetic peptides encoded by these two new exons. Interestingly by searching the Prosite (http://www.expasy.org/prosite) and PredictNLS (http://cubic.bioc.columbia.edu/predictNLS/) databases (Fig. 2)[22] we found the peptides contain nuclear localization signals (NLS). As we know, NLS plays a key role in mediating the transportation of nuclear protein into the nucleus[23,24]. Two classes of NLS have been identified so far. The first class, named monopartite NLS, consists of a short stretch of basic amino acids[25]. The second is composed of two clusters of basic residues separated by

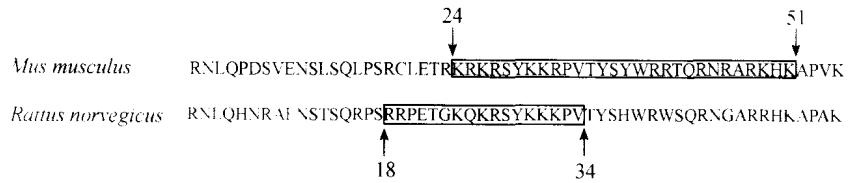

Fig. 2. The NLS encoded by new exons in rodents.

Table 4 The NLS in rodent ZFP*39* and other similar NLS. Black font indicates basic amino acids of NLS, and the number behind the NLS represent the location of NLS in the corresponding proteins

Name	Structure feature	Name	Structure feature
ZFP39 (Mus musculus)	RCLETRKRKRSYKKRPV 18	ZFP39 (Mus musculus)	**KRKR**SYKKRPVTYSYWRRTQRNRAR**K**H**K** 24
ZFP39 (Rattus norvegicus)	**RR**PETGKQKRSY**KKK**PV 18	ZFP39 (Rattus norvegicus)	KQKRSYKKKPVTYSHWRWSQRNGARRHK 24
hSlu7 (Homo sapiens)[28]	**RK**GACENCGAMTH**KKK** 133	U2AF (Mus musculus)	**RRRR**SQHLTRGAKEEQGGWIRSPCHE**KKK** 59
HCDA (Homo sapiens)[29]	**KR**PACTLKPECVQQLLVCSQEA**KK** 27	TP2A (Mus musculus)	**RKRK**PSSSDSSDSDFERAISKGATSK**KA**K 1452
Nucleoplasmin (Xenopus laveis)[30]	**KR**PAATKKAGQAKKKK 171	ATRX (Mus musculus)	RKRKNSTSGSDFDTKKGKSTETSIISKKK 791
RB (Homo sapiens)[31]	**KR**SAEGSNPPKPL**KKLR** 77	ZNP40 (Homo sapiens)	**KRKKI**VAENHLKKIPKSPLRNPLQA**K**H**K** 48

spacer sequence of about 10 aa in length and is called bipartite NLS[26]. Previous studies showed that all members of KRAB-*zfp* family have NLS located in their ZFP domain[27]. In the rodent *zfp39* genes, there are two alternative splicing forms. One is lacking the new exons, as nuclear protein indicates that the protein should have NLS within ZFP domain too. Both pieces of evidence remind us that this new NLS encoded by these two new exons may mediate in a new way the protein transport.

To understand functional divergence between the two species, we also compared the NLS structure of mouse and rat ZFP39. In mouse, the NLS includes basic residues at 24－27 (KRKR), and 48－51 (RKHK) residues separated by 20-aa spacer sequence, and in rat the NLS is composed of basic residues at 18－19 (RR) and 30－32 (KKK) residues separated by 9-aa spacer sequence. Obviously they belong to different bipartite NLS forms (Table 4).

Recently, numerous NLS proteins are described and they are conservative in the basic amino acid motifs. Many mutant NLS demonstrate that the basic amino acids are important residues mediating the transport of nuclear proteins into the nucleus[29]. It has been found that the different conformations of NLS can mediate three different transporting patterns: (i) the efficiency of nuclear protein transporting can be regulated by different modifications of phosphate sites of NLS[23]; (ii) the affinity difference of different NLS to their different receptors can influence the dosage of functional nuclear proteins; (iii) NLS bind to specific receptors thus allowing proteins to be selectively imported into nucleus[32]. Because the members of KRAB-ZFP can exhibit transcription repression effect at high dosages but activate transcription at low dosages[33,34], the divergence of these new exons between mouse and rat might influence the functions of *zfp39*. Both the study of Noce et al.[35] and ESTs data of *zfp39* show that this *zfp39* expresses in spermatocytes, oocytes, brain and other tissues. From the point of view of gene expression regulation network, the appearance of these new exons might profoundly enhance the complexity of gene expression regulation.

Although the concrete function of the new exons awaits further investigation, our results suggest that these new exons should play important roles in the evolution of *zfp39*. We believe that by studying more young exons, we could have a comprehensive and profound understanding of the origin and evolution of exons.

Acknowledgements We thank Prof. Zheng Xiaoguang and Dr. Jing Meidong for providing mouse samples, and specially Yu Haijing, Li Yan, Zhou Qi, Jiang Huifeng, Zhao Ruoping, Zhang Yue and Ding Yun for generous discussions and technical support. This work was supported by the program of the CAS-Max Planck Junior Scientist Group, the key project of the CAS (Grant No. KSCX2-SW-121) and the National Natural Science Foundation of China (Grant Nos. 30325016 & 30430400).

References

1. Li, X., Yang, S., Peng, L. X. et al., Origin and evolution of new genes, Chinese Science Bulletin, 2004, 49(16): 1219－1225.
2. Long, M., Betran, E., Thornton, K. et al., The origin of new genes: Glimpses from the young and old, Nat. Rev. Genet., 2003, 4(11): 865－875.
3. Nekrutenko, A., Li, W. H., Assessment of compositional heterogeneity within and between eukaryotic genomes, Genome Res., 2000, 10(12): 1986－1995.
4. Rogalla, P., Kazmierczak, B., Flohr, A. M. et al., Back to the roots of a new exon--the molecular archaeology of a SP100 splice variant, Genomics, 2000, 63(1): 117－122.
5. Letunic, I., Copley, R. R., Bork, P., Common exon duplication in animals and its role in alternative splicing, Hum. Mol. Genet., 2002, 11(13): 1561－1567.
6. Ast, G., How did alternative splicing evolve? Nat. Rev. Genet., 2004, 5(10): 773－782.
7. Makalowski, W., Mitchell, G. A., Labuda, D., Alu sequences in the coding regions of mRNA: A source of protein variability, Trends Genet., 1994, 10(6): 188－193.
8. Kondrashov, F. A., Koonin, E. V., Evolution of alternative splicing:

Deletions, insertions and origin of functional parts of proteins from intron sequences, Trends Genet., 2003, 19(3): 115－119.
9. Wang, W., Brunet, F. G., Nevo, E. et al., Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*, Proc. Natl. Acad. Sci. USA, 2002, 99(7): 4448－4453.
10. Wang, W., Yu, H., Long, M., Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species, Nat. Genet., 2004, 36(5): 523－527.
11. Looman, C., Abrink, M., Mark, C. et al., KRAB zinc finger proteins: An analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution, Mol. Biol. Evol., 2002, 19(12): 2118－2130.
12. Shannon, M., Hamilton, A. T., Gordon, L. et al., Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters, Genome Res., 2003, 13(6A): 1097－1110.
13. Urrutia, R., KRAB-containing zinc-finger repressor proteins, Genome Biol., 2003, 4(10): 231.
14. Jordan, I. K., Wolf, Y. I., Koonin, E. V., No simple dependence between protein evolution rate and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly, BMC Evol. Biol., 2003, 3: 1.
15. Kumar, S., Tamura, K., Nei, M., MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, Brief Bioinform., 2004, 5(2): 150－163.
16. Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. et al., DnaSP, DNA polymorphism analyses by the coalescent and other methods, Bioinformatics, 2003, 19: 2496－2497.
17. Tajima, F., Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics, 1989, 123(3):585－95.
18. Fu, Y. X., Li, W. H., Statistical tests of neutrality of mutations, Genetics, 1993, 133(3): 693－709.
19. Zhou, Q., Wang, W., Detecting natural selection at the DNA level, Zoological Research, 2004, 25 (1): 73－80.
20. O'Brien, S. J., Eizirik, E., Murphy, W. J., Genomics. On choosing mammalian genomes for sequencing, Science, 2001, 292(5525): 2264－2266.
21. Springer, M. S., Murphy, W. J., Eizirik, E. et al., Placental mammal diversification and the Cretaceous-Tertiary boundary, Proc. Natl. Acad. Sci. USA, 2003, 100(3): 1056－1061.
22. Cokol, M., Nair, R., Rost, B., Finding nuclear localization signals, EMBO Rep., 2000, 1(5): 411－415.
23. Dingwall, C., Laskey, R. A., Protein import into the cell nucleus, Annu. Rev. Cell Biol., 1986, 2: 367－390.
24. Fontes, M. R., The, T., Toth, G. et al., Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-alpha, Biochem. J., 2003, 375: 339－349.
25. Moede, T., Leibiger, B., Pour, H. G. et al., Identification of a nuclear localization signal, RRMKWKK, in the homeodomain transcription factor PDX-1, FEBS Lett., 1999, 461(3): 229－234.
26. Boulikas, T., Nuclear localization signals (NLS), Crit. Rev. Eukaryot. Gene Expr., 1993, 3(3): 193－227.
27. Pandya, K., Townes, T. M., Basic residues within the Kruppel zinc finger DNA binding domains are the critical nuclear localization determinants of EKLF/KLF-1, J. Biol. Chem., 2002, 277(18): 16304－16312.
28. Shomron, N., Reznik, M., Ast, G., Splicing factor hSlu7 contains a unique functional domain required to retain the protein within the nucleus, Mol. Biol. Cell., 2004, 15(8): 3782－3795.
29. Somasekaram, A., Jarmuz, A., How, A. et al., Intracellular localization of human cytidine deaminase. Identification of a functional nuclear localization signal, J. Biol. Chem., 1999, 274(40): 28405－28412.
30. Dingwall, C., Dilworth, S. M., Black, S. J. et al., Nucleoplasmin cDNA sequence reveals polyglutamic acid tracts and a cluster of sequences homologous to putative nuclear localization signals, EMBO J., 1987, 6(1): 69－74.
31. Efthymiadis, A., Shao, H., Hubner, S. et al., Kinetic characterization of the human retinoblastoma protein bipartite nuclear localization sequence (NLS) *in vivo* and *in vitro*. A comparison with the SV40 large T-antigen NLS, J. Biol. Chem., 1997, 272(35): 22134－22139.
32. Fahrenkrog, B., Aebi, U., The nuclear pore complex: Nucleocytoplasmic transport and beyond, Nat. Rev. Mol. Cell Biol., 2003, 4(10): 757－766.
33. Sauer, F., Fondell, J. D., Ohkuma, Y. et al., Control of transcription by Kruppel through interactions with TFIIB and TFIIE beta, Nature, 1995a, 375(6527): 162－164.
34. Sauer, F., Jackle, H., Heterodimeric *Drosophila* gap gene protein complexes acting as transcriptional repressors, EMBO J., 1995b, 14(19): 4773－4780.
35. Noce, T., Fujiwara, Y., Sezaki, M. et al., Expression of a mouse zinc finger protein gene in both permatocytes and oocytes during meiosis, Dev. Biol., 1993, 153 (2): 356－367.

(Received December 6, 2004; accepted April 11, 2005)