Genomics & Informatics Vol. 2(2) 86-91, June 2004

High Correlation between Alu Elements and the Conversion of 3' UTR of mRNAs Processed Pseudogenes

Hyeong Jun An, Dokyun Na, Doheon Lee, Kwang Hyung Lee and Jonghwa Bhak*

Department of BioSystems, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea

Abstract

Even though it represents 6-13% of human genomic DNA, Alu sequences are rarely found in coding regions. When in exon region, over 80 % of them are found in 3' untranslated region (UTR). Pseudogenes are an important component of human genome. Their functions are not clearly known and the mechanism of how they are generated is still debatable. Both the Alu and Pseudogenes are important research problems in molecular biology. mRNA is thought to be a prime source of pseudogene and active research is going on its molecular mechanism. We report, for the first time, that mRNAs containing Alu repeats at 3' UTR has a significantly high correlation with processed pseudogenes, suggesting a possibility that Alu containing mRNAs have a high tendency to become processed pseudogenes. It is known that about 10% of all human genes have been transposed. Transposed genes at 3' UTR without Alu repeat have about two processed pseudogenes per gene on average while we found with statistical significance that a transposed gene with Alu had over three processed Pseudogenes on average. Therefore, we propose Alu repeats as a new and important factor in the generation of pseudogenes.

Keywords: Alu, LINE, pseudogene, processed pseudogene, Alu and pseudogenes

Introduction

Alu elements have been amplified in primate genomes through an RNA-dependent mechanism, termed retrotransposition, and have reached a copy number in excess of 500,000 copies per human genome (Deininger *et al.*, 1999). Alu sequences represent around 6-13% of human genomic DNA (Mighell *et al.*, 1997; Boeke 1997). Alu sequences were identified in 5% of 1616 human full-length cDNA, 82% of which were found in the 3' UTR, 14% lie in the 5' UTR, and very rarely in coding region (Yulung *et al.*, 1995). Alu sequences are postulated to be retrotransposon that have been inserted into the human genome via a single-stranded RNA generated by RNA polymerase III transcription, but the mechanisms and factors about retrotransposition are very poorly understood (Mighell *et al.*, 1997; Moran *et al.*, 1999).

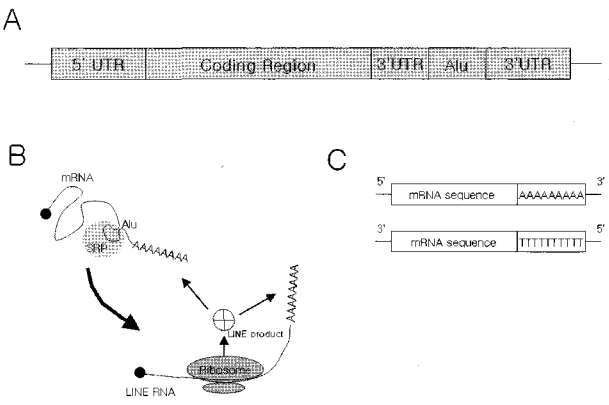

Fig. 1. The schematic diagrams of mRNA containing Alu and the steps of the generation of processed pseudogene of the mRNA. (A) an mRNA containing Alu at 3' UTR. (B) SRP bound to Alu at the 3' UTR of mRNAs and SRP brought itself to the LINE product. The product adhered to not only the poly A of LINE, but also the poly A of near mRNA. (C) an mRNA was inserted into human genome by the product, therefore the mRNA became processed pseudogene.

Pseudogenes are regarded as disabled copies of genes that are known to have no important function and do not code any full-length proteins. Pseudogenes are known to be a consequence of gene duplication which can occur in two fundamentally different ways: firstly, by retrotransposition, and secondly, via the duplication of genomic DNA. Pseudogenes occurred by retrotransposition are known as processed pseudogenes (Maestre *et al.*, 1955; Mighell *et al.*, 2000). Human pseudogenes on chromosome 21 and 22 first had been discovered by Gerstein group in USA (Harrison *et al.*, 2000). Over 8,000

*Corresponding author: E-mail biopark@kaist.ac.kr, Tel +82-42-869-4318, Fax +82-42-869-4358
Accepted 9 May 2004

processed pseudogenes of human were identified by the same group. 20% of processed pseudogenes are from highly expressed ribosomal proteins and about 10% of genes are known to be originated in the discovered processed pseudogenes. There are three factors affecting biogenesis of processed pseudogenes. First, mRNA expression level is the most deciding factor. Higher number of mRNA transcripts in the germ-line cell is proportional to the higher chance of retrotransposition compared to those genes that are less well transcribed. Second, relatively GC-poor ribosomal protein (RP) genes have more processed pseudogenes than GC-rich RP genes. Third, the reverse-transcription and insertion processes are less efficient for longer mRNA transcripts than shorter mRNA transcripts (Zhang and Madden 2003).

Long Interspersed Nucleotide Element (LINE) has a few functions in human. For examples, autonomous retrotransposons, processed pseudogene formation, exon shuffling and double-strand-break repair (Moran *et al.*, 2001; Ostertag and Kazazian, 2001; Kazazian and Goodier, 2002). Both Alu and mRNAs are transposed by the products of LINE because the products bind to polyA of not only LINE, but also mRNAs and Alu. Although the mechanisms of transposition of both of Alu and mRNAs are the same, the efficiency of the transpositions of Alu is over 300 times higher than that of mRNA (Esnault *et al.*, 2000; Kazazian, 2004).

The signal recognition particle (SRP) binding to Alu RNA brings the Alu RNA into proximity with ribosomes and nascent L1 proteins on L1 RNA (Dewannieux *et al.*, 2003; Kazazian, 2004). This particle has been known as the factor of increasing the efficiency of transposition of Alu. This study reports a correlation between Alu that are included in the 3' UTR of mRNAs and the production of processed pseudogenes. We support that hypothesis that Alu can enhance the conversion of mRNA into processed pseudogenes.

Methods

Dataset

The list of processed pseudogenes was downloaded in the website (*http://www.pseudogene.org/*). We removed about 2,000 processed pseudogenes originated from ribosomal proteins among all the 8,000 processed Pseudogenes. This is because we only consider mRNA sequences. The list was published in the human genome draft build 34. Blastn program of NCBI was used to match processed pseudogene sequences and the sequences in the human chromosome. The reference sequences (RefSeq) of NCBI are reviewed sequences that are more reliable than most other databases (*ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot*).

The IDs of processed pseudogenes are named after SwissProt (*http://www.expasy.org/*) IDs. We searched for the original mRNA sequences of processed pseudogenes by mapping IDs between RefSeq IDs and SwissProt IDs in the website; GeneCard (*http://bioinformatics.weizmann.ac.il/cards/*).

881 mRNA sequences were matched to 2150 processed pseudogenes. In previous studies, processed pseudogenes were searched against only the protein coding regions and their UTRs (untranslated region) were not considered as they converted protein sequences to DNA sequences as data source. However, as most of Alu sequences in mRNAs are in the 3' UTR region, we chose to use 3' UTR sequences of mRNAs from RefSeq. The same protocol was applied to the processed Pseudogenes for the 3' UTR sequence regions. Additionally, for higher quality data, we removed 3' UTR sequences that were not alignable from the RefSeq and processed pseudogenes. The reason why there are unalignable sequences between the two sets (RefSeq derived and processed Pseudogenes) is that there could have been deletions at 3' UTR of pseudogenes for the last few hundred thousand years. The final dataset had 1214 processed pseudogenes covered by 581 mRNA species.

The following are the stepwise summary of the procedure:

1) RefSeq consisted of about 20,000 mRNA sequences and processed pseudogenes were about 6,000 sequences excluding 2,000 originated from ribosomal proteins.
2) About 10% of RefSeq mRNAs contain 20,000 processed pseudogenes. It is consistent with the previous study (Zhang *et al.*, 2003). The pseudogene database had only Swissprot IDs (e.g. chr2_O75317.1, chrY_O75317.1, and chrX_O75317.1), however, The Genbank format of RefSeq had no Swissprot IDs. Therefore, we make the simple java source that compared Swissprot IDs with RefSeq IDs in the the website; GeneCard (*http://bioinformatics.weizmann.ac.il/cards/*). As a result, 881 out of 2,000 RefSeq IDs were matched with 2,150 out of about 6,000.
3) 881 out of 2,000 mRNAs contained 2150 pseudogenes (on average 2.4 pseudogenes per mRNA)
4) We collected 3' UTR sequences of both mRNAs and pseudogenes with annotation information of RefSeq.
5) Sequence alignment with bl2seq program (from NCBI) for quality checking.

6) If there are no matching regions between the two 3'UTR sequences, we removed the 3' UTRs from the mRNA and pseudogene data sets. When pseudogenes were discovered and published, the length criterion of them was 70% of the length of original genes. If 3' sequence of pseudogene was truncated, the pseudogene had no 3' UTR; therefore, we made an alignment between 3'UTR of mRNA and that of pseudogenes. In this process, 300 mRNAs containing 936 pseudogenes were excluded (resulting in 581 mRNAs).
7) Finally, 581 mRNAs and 1214 pseudogenes were in our dataset (on average 2.09 pseudogenes per mRNA)

Significance analysis

518 mRNAs had 2.09 processed pseudogenes on average and 37 mRNAs (n) with Alu at 3' UTR mRNAs had 3.11 pseudogenes on average; therefore, the average of the population of whole mRNAs was 2.09 (u), and the average of the population of mRNAs with Alu is 3.11 (X). And also, the standard deviation of the population was 3.32 among 518 whole mRNAs. Therefore, P value was 0.0336.

$$Z = \frac{X - u}{\delta/\sqrt{n}} = \frac{3.11 - 2.09}{3.32/\sqrt{37}} = 1.83$$

Results

9.5% of processed pseudogenes had Alu repeats, but 6.4% of mRNAs of reference sequences (RefSeq) supported by National Center for Biotechnology Information (NCBI) had Alu over 100 bases at 3' UTR in our dataset (see Methods). The average length of Alu is known to be about 280bp and the previous study used the minimum length of Alu as around 60bp and the minimum of 70% identity (Weiner *et al.*, 1986; Yulung *et al.*, 2003). In our study, we set a more rigorous threshold of 100bp and 70% identity. We found that there were few Alu at 5' UTR in both mRNAs and the processed pseudogenes because of short lengths of 5'UTRs (Table 1) and it is consistent with a previous study (Yulung *et al.*, 1995). On average an mRNA had 2.09 processed

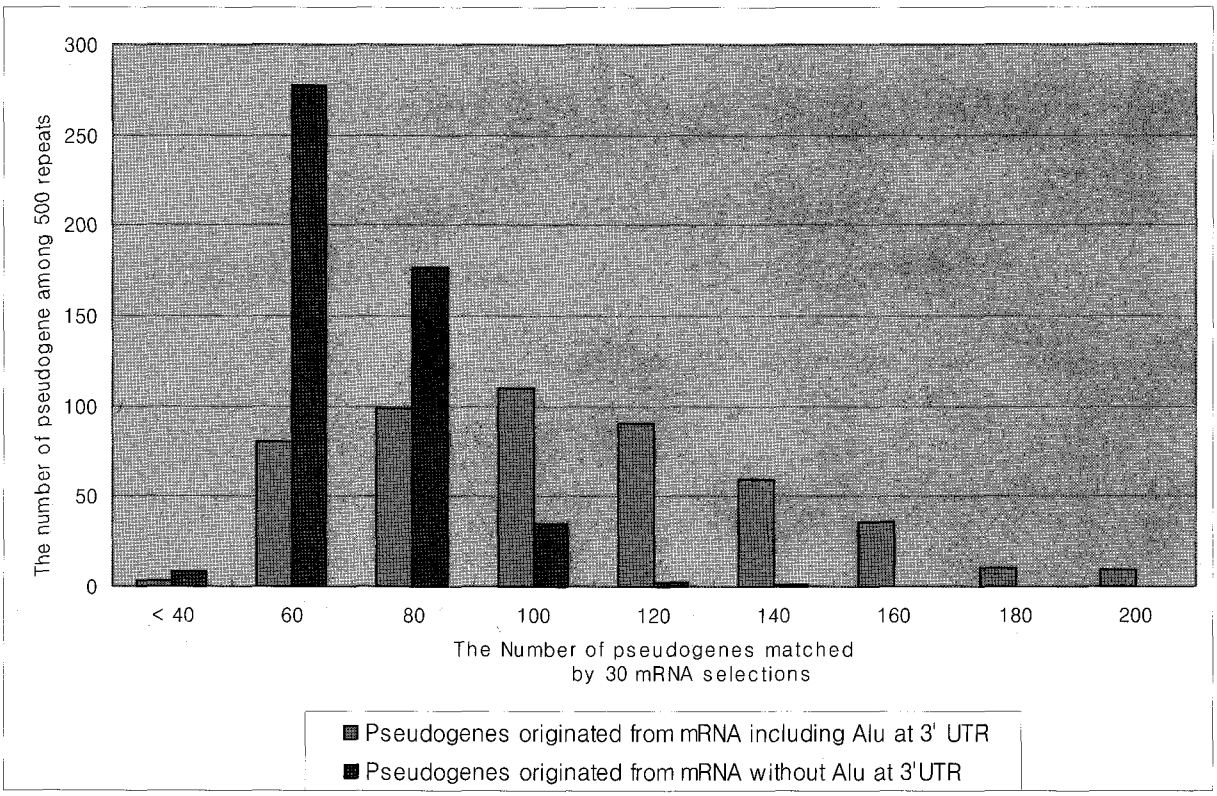

Fig. 2. The two distributions of the number of pseudogenes detected in tbe 30 mRNA selections. One distribution is for the mRNAs containing Alu and the other is mRNAs containing no Alu at 3' UTR out of 500 randomly sampled repeats. In the case of mRNAs without Alu, 40-60 processed pseudogenes per random 30 mRNAs appeared about 270 times out of 500.

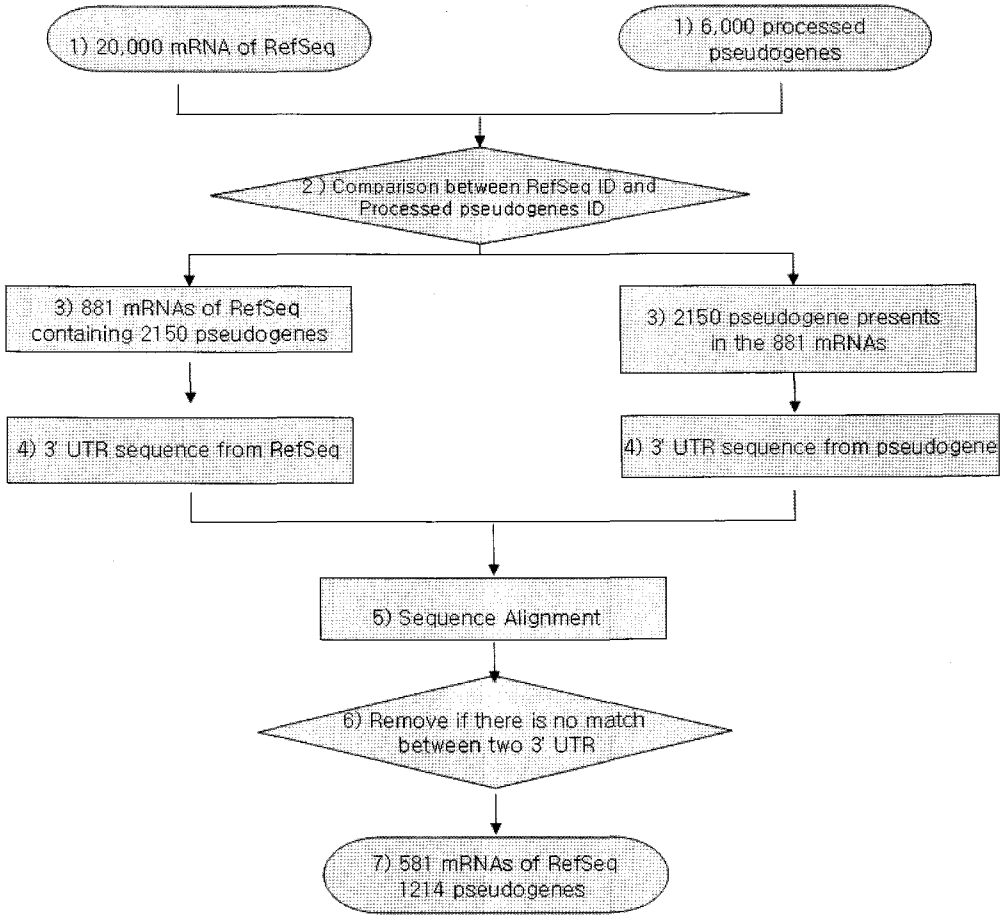

Fig. 3. The schematic diagram of the protocol for test data preparation.

pseudogenes in our dataset. On the other hand, an mRNAs without Alu at 3' UTR had 2.02 processed pseudogenes and 3.11 processed pseudogenes on average per mRNA with Alu at 3' UTR ($Z=1.83$, $P<0.034$).

Table 1. The numbers of mRNAs and processed pseudogenes including Alu over 100 bases at 3' UTR and 5' UTR.

	No. of mRNAs of RefSeq	No. of Processed pseudogene
3' UTR over 100 bp	37	115
5' UTR over 100 bp	5	8
The number of sequences	581	1214

We have selected 30 random mRNAs for each data set from two types of data; one containing Alu and the other not-containing Alu at 3' UTR. The reason for selecting 30 randomly generated mRNA was for a comparison under same conditions. These random mRNAs removed possible bias due to one or two mRNAs containing a great number of pseudogenes. If our result was due to such bias by mRNAs containing a great number of pseudogenes, normal distribution like Figure 1 would not be shown and the bias would be detected in the distribution statistics. I.e., Figure 1 showed that there is no such bias by unusually high number of Pseudogenes. The random sampling process has been iterated 500 times to eliminate statistical bias. We compared the randomly generated mRNAs between two data sets. The distribution of the numbers of Pseudogenes in the randomly generated sets of mRNAs is shown in Figure 1. The mRNAs containing Alu at 3' UTR had at most 80-100 processed pseudogenes while 40-60 processed pseudogenes were found in the mRNAs that do not contain out of 500 randomly sampled

repeats. As shown in the distribution, in the range of 40 to 100 pseudogene copies in the X-axis, the number of pseudogenes without Alu decreases, but the number of those with Alu increases. After around 100 pseudogene copies, there are few processed pseudogene without Alu at 3' UTR while the number of processed pseudogenes with Alu at 3' UTR rises to 200 pseudogenes copies.

Conclusion

Alu is one of the representative transposons in human genome. Alu sequences are known to be randomly inserted in human genome. Some of them are located in UTR and are transcribed with the host mRNAs containing the Alu repeats. The transposition of Alu element is known to be 300 times more efficient (Dewannieux *et al.*, 2003) on average than that of any mRNAs because SRP binding at Alu is suggested to bring itself into ribosome and in the vicinity of the products of LINE. We suspected that the Alu within mRNA had a similar effect as Alu alone. It is also suggested that once Alu is inserted in human genome, it stays in the genome except in the case of homologous recombination mechanism (Lander *et al.*, 2001; Kazazian, 2004 and personal communication). Therefore, most processed Pseudogenes should have the same tract of Alu as that of the original functional genes. This tendency of stable preservation of Alu is the foundation of our experiment in calculating the tendency of pseudogene generation when Alu sequences are inserted in mRNA. Few Alu repeats at 5' UTR were found in our study, and it is consistent with previous survey (Yulung *et al.*, 1995). This is mainly because the length of 5' UTR is much shorter than 3' UTR. Consequently, the ratio of the Alu insertion was very low. Therefore, we used Alu at 3'UTR only. Here, we put forward a hypothesis that Alu is another important factor that enhances the conversion of an mRNA to a processed pseudogene. An mRNA with Alu element at 3' UTR has 1.53 times higher likelihood of becoming a pseudogene.

Acknowledgements

AHJ, ND, LD, and LKH are supported by the Korean Systems Biology Research Grant (M1-0309-02-0002) from the Ministry of Science and Technology of Korea. BJH is supported by IMT-2000-C4-3 grant of ministry of information and communication of Korea. AHJ is thankful to Dr. Kazazian for answering my questions about Alu. We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM SUR program for providing research and computing facilities.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Boeke, J.D. (1997). LINE and Alus – the polyA connection. *Nature Genet.* 16, 6-7

Deininger, P. L. and Batzer, M. A. (1999). Alu repeats and Human Disease. *Melecular Genetics and Metabolism* 67, 183-193

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics* 35, 41-48

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotranspsposons generate processed pseudogenes. *Nature Genet* 24, 363-367

Gish, W. and States, D.J. (1993). Identification of protein coding regions by database similarity search. *Nature Genet.* 3:266-272.

Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone P., Echols, N., Johnson, T., and Gerstein, M. (2002). Molecular fossils in the Human Genome: identification and Analysis of the pseudogenes in chromosomes 21 and 22, *Genome Res.* 12, 272-280

Kazazian, H.H. Jr. (2004). Mobile elements: drivers of genome evolution. *Science* 303, 1626-1632.

Kazazian, H.H. Jr. and Goodier J.L. (2002). LINE drive. retrotransposition and genome instability. *Cell* 110, 277-280.

Kurose K., Hata K., Hattori, M., and Sakaki, Y. (1995). RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. *Nucleic Acids Res.* 23, 3704-3709.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., and Zody, M.C., (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Madden, T.L., Tatusov, R.L., and Zhang, J. (1996). Applications of network BLAST server. *Meth. Enzymol.* 266:131-141.

Maestre, J., Tchenio, T., Dhellin O., and Heidmann T. (1995). mRNAs retroposition in human cells: processed pseudogene formation. *EMBO J.* 14, 6333-6338.

Mighell, A.J., Markham, A.F., and Bobinson, P.A. (1997). Alu sequences. *FEBS Lett.* 417, 1-5.

Mighell, A.J., Smith, N.R., Robinson, P.A., and Markham, A.F. (2000). Vertebrate pseudogenes. *FEBS Lett.* 468, 109-114

Moran, V., DeBerardinis, J., and Kazazian, H.H. Jr. (1999). Exon Shuffling by L1 Retrotransposition. *Science* 283, 1530-1543.

Ostertag, M. and Kazazian, H.H. Jr. (2001). Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* 35, 501-538.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics* 13, 163.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H. Jr., Boeke, J.D., and Moran, J.V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* 21, 1429-1439.

Weiner, A. M., Deininger, P. L., and Efstratiadis, A.. (1986). Nonviral retroposons: Genes, pseudogenes and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* 55, 631-662.

Yulung, G., Yulung, A. and Fisher, E.M.C. (1995). The frequency and position of Alu repeats in cDNAs, as determined by database searching. *Genomics* 27, 544-548

Zhang, J. and Madden, T.L. (1997). PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* 7:649-656.

Zhang, Z., Harrison, P. M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* 13, 2541-2558