The abundance of processed pseudogenes derived from glycolytic genes is correlated with their expression level

Laura McDonell and Guy Drouin

Abstract: The abundance of processed pseudogenes in different vertebrate species is known to be proportional to the length of their oogenesis. However, this hypothesis cannot explain why, in a given species, certain genes produce more processed pseudogenes than others. In particular, one would expect that all genes of the glycolytic pathway would generate roughly the same number of processed pseudogenes. However, some glycolitic genes generate more processed pseudogenes than others. Here, we show that there is a positive correlation between the abundance of processed pseudogene generated from glycolytic genes and their level of expression. The variation in expression level of different glycolytic genes likely reflects the fact that some of them, such a *GAPDH*, have functions other than those they play in glycolysis. Furthermore, the age distribution of *GAPDH*-processed pseudogenes corresponds to the age distribution of LINE1 elements, which are the source of the reverse transcriptase that generates processed pseudogenes. These results support the hypothesis that gene expression levels affect the level of processed pseudogene production.

Key words: glycolytic genes, GAPDH, processed pseudogenes, transcription, gene expression.

Résumé : L'abondance des pseudogènes remaniés dans différentes espèces vertébrées est proportionnelle à la durée de l'ovogenèse chez ces espèces. Cependant, cette hypothèse ne peut expliquer pourquoi, dans une espèce donnée, certains gènes produisent plus de pseudogènes remaniés que d'autres. En particulier, on pourrait s'attendre à ce que tous les gènes de la voie glycolytique généreraient des nombres similaires de pseudogènes remaniés. Toutefois, certains gènes glycolytiques génèrent beaucoup plus de pseudogènes remaniés que d'autres. Ici, nous montrons qu'il y a une corrélation positive entre l'abondance des pseudogènes remaniés générés à partir des gènes de la glycolyse et leur niveau d'expression. La variation du niveau d'expression de différents gènes glycolytiques reflète probablement le fait que certains d'entre eux, tel celui codant pour GAPDH, ont des fonctions autres que celles qu'ils jouent dans la glycolyse. Par ailleurs, la répartition par âge des pseudogènes remaniés *GAPDH* correspond à la répartition par âge des éléments LINE1 qui sont la source de la transcriptase réverse qui génère les pseudogènes remaniés. Ces résultats sont consistants avec l'hypothèse que les niveaux d'expression des gènes affectent leur niveau de production de pseudogène remaniés.

Mots-clés : gènes de la glycolyse, GAPDH, pseudogènes remaniés, transcription, expression des gènes.

Introduction

Processed pseudogenes are created by the reverse transcription of processed RNAs followed by the genomic integration of the resulting cDNA (Vanin 1985). They typically lack introns and have a poly-A tail. Although processed pseudogenes are similar to their parent gene, most of them are not functional because they lack a 5′ promoter region (Graur et al.1989; Gu and Li 1995). Thus processed pseudogenes are released from selective constraints and as a result can undergo genetic drift (Gojobori et al. 1982). They accumulate frame disruptions and evolve faster than their functional paralogs (Ophir et al. 1999). However, not all retrotransposed genes share the same fate (Brosius 1999). For example, the gene coding for phosphoglycerate kinase (*PGK*) generated two intronless sequences. The first, *ΨPGK-1*, is a processed pseudogene but the second, *PGK2*, produces a transcript that is expressed in the germ line during spermatogenesis (McCarrey et al. 1996).

Processed pseudogenes have been found in all animal genomes so far studied, and their abundance has been shown to be proportional to the length of oogenesis (Weiner et al. 1986; Drouin 2006). However, it has long been known that all genes do not generate the same number of processed pseudogenes. A well known example is the number of processed pseudogenes generated by the 10 enzymes of the glycolytic cycle where the number of processed pseudogenes generated by the glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene greatly outnumbers that of the other glycolytic enzymes (Weiner et al. 1986; Graur and Li 2000). Liu et al. (2009) also recently characterized the relative abundance of processed pseudogenes generated from glycolytic

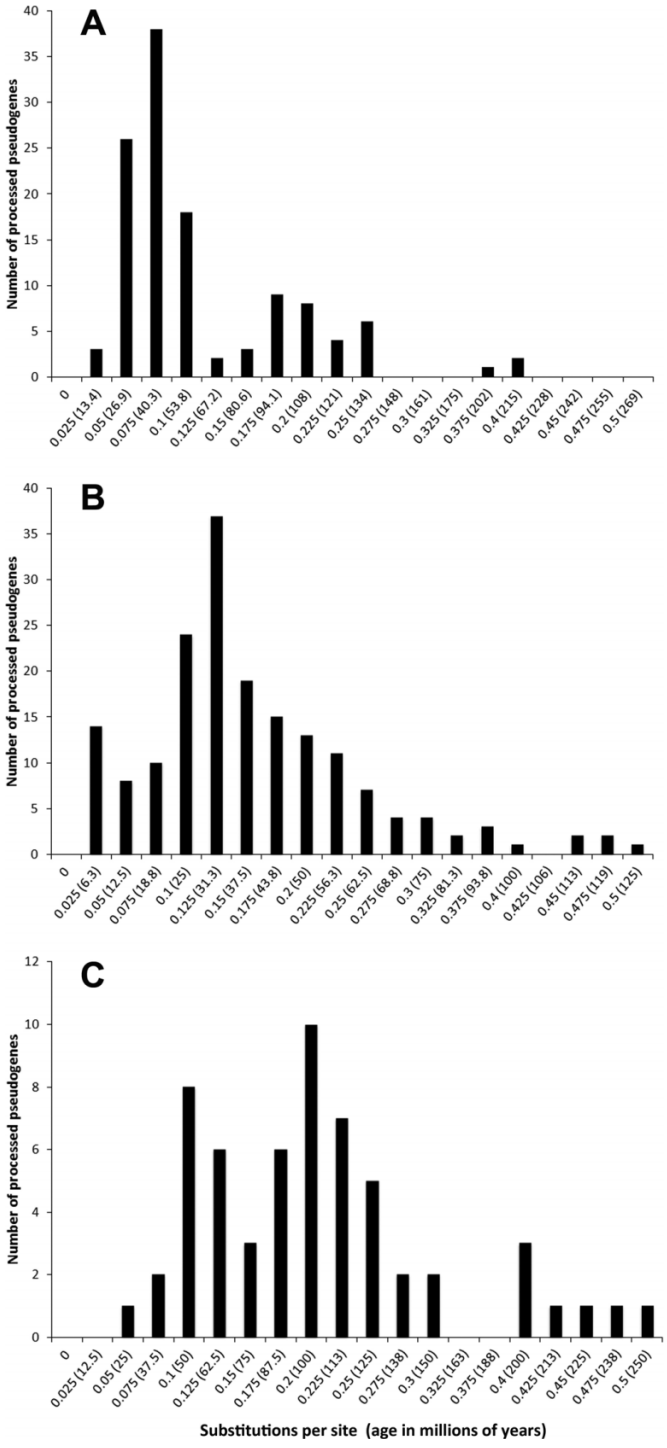
Received 7 October 2011. Accepted 3 January 2012. Published at www.nrcresearchpress.com/gen on 6 February 2012.

Paper handled by Associate Editor T.E. Bureau.

L. McDonell and G. Drouin. Département de biologie et Centre de recherche avancée en génomique environnementale, Université d'Ottawa, Ottawa, ON K1N 6N5, Canada.

Corresponding author: Guy Drouin (e-mail: gdrouin@science.uottawa.ca).

Fig. 1. Distribution of the *GAPDH*-processed pseudogenes grouped according to their number of substitutions per site (and age in millions of years) in the (A) dog, (B) mouse, and (C) human genomes. Note that 0.025 substitutions per site represent roughly 6.25 million years in mice, 12.5 million years in humans, and 13.4 million years in dogs.



genes and suggested that the overabundance of *GAPDH* processed pseudogenes was due to the fact that this enzyme had many more biological roles outside glycolysis. However, they did not address whether there were correlations between the abundance of processed pseudogenes derived from glycolytic genes and their expression.

Table 1. Number of glycolytic-processed pseudogenes in the human, mouse, and dog genomes.

	Human	Mouse	Dog
HK1	0	0	0
HK2	0	0	0
HK3	1	0	0
HK4	0	0	0
GPI1	0	1	3
PFKL	0	0	0
PFKM	0	0	0
PFKP	0	0	0
ALDOA	1	9	2
ALDOB	0	0	0
ALDOC	0	0	0
TPI	3	4	0
GAPDH	56	166	120
PGK1	0	1	2
PGK2	1	0	2
PGM1	14	5	2
PGM2	1	1	1
ENO1	2	7	3
ENO2	0	1	1
ENO3	0	0	0
PKM1	1	0	1
PKM2	1	7	0

Note: Number of glycolytic-processed pseudogenes for the genes for which gene expression data was available. For example, even though there are four genes coding for phosphoglycerate mutase, expression data was available only for *PGM1* and *PGM2*. See supplemental Table S2 for the total number of processed pseudogenes and functional genes for each of the 10 glycolytic genes. Abbreviations: *HK*, hexokinase; *GPI*, glucose-6-phosphate isomerase; *PFK*, 6-phosphofructokinase; *ALDO*, aldolase; *TPI*, triose-phosphate isomerase; *GAPDH*, glyceraldehyde-3-phosphate dehydrogenase; *PGK*, phosphoglycerate kinase; *PGM*, phosphoglycerate mutase; *ENO*, enolase; *PK*, pyruvate kinase.

Here, we show that there are significant correlations between the abundance of glycolytic-processed pseudogenes and gene expression. We also show that the age distribution of mouse and dog *GAPDH*-processed pseudogenes corresponds to the age distribution of LINE1 elements in the mouse genome. Our results therefore support the hypothesis that the expression level of glycolytic genes affects their level of processed pseudogene production.

Materials and methods

A catalogue of the mouse, human, and dog glycolytic cycle processed pseudogenes sequences was assembled from the databases found on Pseudogene.org (http://www.pseudogene.org/). Databases of total mouse- and human-processed pseudogenes were already in place. However, we had to build a processed pseudogene catalogue for the dog from a bank of processed, duplicated, and ambiguous pseudogene sequences. For this species, we choose to extract only the 6126 sequences that had been identified as being processed pseudogenes and did not consider the remaining 5198 ambiguous pseudogene sequences.

ClustalW was used to align the processed pseudogene sequences with their parent gene using default parameters (Thompson et al. 1994). MEGA 4.0 (Tamura et al. 2007) was used to calculate the pairwise distances between the

Table 2. Spearman rank correlation tests between the number of processed pseudogenes and gene expression.

	Human				Mouse				Dog	
	ESC	Liver	Brain	Muscle	ESC	Liver	Brain	Muscle	Heart 1	Heart 2
ρ	0.421	0.220	0.455	0.238	0.702	0.449	0.698	0.575	0.535	0.041
P value	0.065	0.352	0.044	0.313	0.0004	0.041	0.0004	0.008	0.022	0.875

Note: ρ, rho (Spearman rank correlation test); ESC, embryonic stem cells.

functional *GAPDH* gene and its processed pseudogenes using the Tajima–Nei model. These distances where used to create Fig. 1.

The gene expression values were extracted from NCBI's GEO datasets (http://www.ncbi.nlm.nih.gov/geo/). We used only data obtained from Affymetrix chips and, in tissues where standard gene expression was not studied, we used the values for the negative controls. The datasets used and the level of gene expressions are shown in the Supplementary data,[1] Table S2. Unfortunately, gene expression data for germ line cells were not available. For mouse and human, we therefore used gene expression data for embryonic stem cells and three somatic tissues. For the dog, gene expression data were only available for two heart samples. Spearman rank correlation tests were used to assess the relationship between the number of processed pseudogenes and their expression levels in the different tissues (Cureton 1958). Each gene had three or more expression values for a given tissue and their individual means were calculated (supplemental Table S2). These means were then scaled to that of the *GAPDH* expression mean, and the Spearman rank correlation test was performed from the resulting values. Each gene was considered separately in the ranking test; for example, *AldoA*, *AldoB*, and *AldoC* where ranked as three entries such that each gene was ranked according to its respective number of processed pseudogene.

Results

Correlations between abundance and the level of gene expression

Table 1 shows that the number of processed pseudogenes generated by the different glycolytic genes are highly variable. This is true not only for genes coding for different enzymatic functions but also for genes coding for similar enzymatic functions. For example, *GAPDH* generates much more processed pseudogenes than all other glycolytic genes and *PGM1* generates more processed pseudogenes than *PGM2*. Also note that all the human and mouse genes coding for different isoforms of the same enzyme, apart from the *PGK2* gene mentioned above and the human *PGAM4* gene, all contain numerous introns (results not shown). Most of these duplicates therefore did not arise through an RNA intermediate.

There are significant correlations between the number of processed pseudogenes and the levels of glycolytic gene expression in all four mouse tissues we examined (Table 2). The fact that this correlation is highest for embryonic stem cells is not particularly surprising because the gene expression in these pluripotent tissues is anticipated to better mirror the gene expression levels expected for germ line cells. What is surprising is that the correlation with gene expression levels in the brain are equivalent to those in stem cells, and significant correlations are also observed with gene expression levels in the liver and in muscle. This suggests that the expression levels of genes in different tissues are not independent. The lower correlations, or nonsignificant correlations, observed in the human and dog genomes likely reflects the lower numbers of glycolytic-processed pseudogenes in these two genomes which have, respectively (roughly only one and two-third as many glycolytic processed pseudogenes compared with the mouse genome; Table 1). These results therefore support the hypothesis that the level of gene expression in the germ line affects the abundance of processed pseudogeneds (Weiner et al. 1986).

Age distribution of processed pseudogenes

Another way to look at the relationship between gene expression and processed pseudogenes is to estimate when processed pseudogenes were generated. If gene expression is correlated with processed pseudogene production, and that the reverse transcriptase responsible for their reverse transcription originates from LINEs, then more processed pseudogenes should be produced when more LINEs are active. In mammals, the LINE1 subfamily is the only active LINE subfamily and is responsible for the generation of *Alu* elements and processed pseudogenes (Feng et al. 1996; Jurka 1997; Weiner 2000; Esnault et al. 2000).

Figure 1 shows the relative age distribution of *GAPDH*-processed pseudogenes in the mouse, human, and dog genomes. Given that the substitution rate of mouse genes (4×10^{-9} substitutions/site/year; Mouse Genome Sequencing Consortium 2002) is approximately twofold higher than those of humans (2×10^{-9} substitutions/site/year; Mouse Genome Sequencing Consortium 2002) and dogs (1.86×10^{-9} substitutions/site/year; Liu et al. 2006), the age distribution of the *GAPDH*-processed pseudogenes shows that there was a peak of mouse and dog *GAPDH* retrotransposition roughly 34 million years ago.

Discussion

The genes that give rise to processed pseudogenes have to be expressed in the germ line, because only the genes retrotransposed in the early stages of development can be fixed in the genome and inherited in subsequent generations (Gonçalves et al. 2000). Furthermore, a high level of mRNA expression in the germ line would be expected to increase the likelihood of retrotransposition by increasing the chance of the transcript's association with the reverse transcriptase produced by a LINE (Esnault et al. 2000). We therefore analyzed the levels of gene expression in different tissues to

[1]Supplementary data are available with the article through the journal Web site (http://nrcresearchpress.com/doi/suppl/10.1139/g2012-002).

determine whether there is a correlation between the number of processed pseudogenes and the levels of glycolytic gene expression. We found several significant correlations between the number of processed pseudogenes and the levels of glycolytic gene expression (Table 2). This is in agreement with the recent results of Podlaha and Zhang (2009) and supports the hypothesis that the abundance of processed pseudogenes is proportional to the expression level of the genes giving rise to them.

Interestingly, the peak of dog and mouse *GAPDH* retrotransposition roughly 34 million years ago we observed (Figs. 1A and 1B, respectively) corresponds to the age distribution of LINE1 elements in the mouse genome (Fig. 2*a* of Zhang et al. 2004). The fact that an excess of *GAPDH*-processed pseudogenes was produced when LINE1 elements were the most active also supports the relationship between gene expression and processed pseudogene production. Thus, it may be possible to study ancient transcriptomes by analyzing processed pseudogenes because they are the "fossilized footprints" of past gene expression (Podlaha and Zhang 2009; Zhang et al. 2004).

From the above, it is clear that the overabundance of *GAPDH*-processed pseudogenes is due to their higher level of gene expression. The question that remains is why are some glycolytic genes, such as *GAPDH*, more highly expressed than others? The answer lays in GAPDH's role as a housekeeping enzyme which means that this enzyme has many other functions other than those involved in glycolysis (Warrington et al. 1999). For instance, GAPDH is involved in numerous subcellular processes unrelated to those of energy production. The multiple functions of GAPDH include, playing a role in membrane fusion, in microtubule bundling, and phosphotranferase activity (Sirover 1999). Furthermore, GAPDH is relocalized to the nucleus where it participates in a variety of nuclear pathways such as the transcriptional regulation of histone gene expression, the recognition of fraudulently incorporated nucleotides in DNA, and in telomere maintenance (Sirover 2005). Therefore, GAPDH is not simply a metabolic enzyme. Its multiple membrane, cytosolic and nuclear functions ensure that it is more highly expressed than the other genes involved in glycolysis, thus increasing the likelihood of its mRNA being reversed transcribed by a LINE1 reverse transcriptase; hence the overabundance of its processed pseudogenes. Finally, the correlations we observed between gene expression and processed pseudogene abundance are also due to the fact that six other glycolytic genes (hexokinase 1, 6-phosphofructokinase, aldolase A, triosephosphate isomerase, phosphoglycerate kinase, phosphoglycerate mutase, and enolase A) are also housekeeping genes which have functions other than those involved in glycolysis (Warrington et al. 1999).

In conclusion, the correlation between processed pseudogene abundance and gene expression levels, and the fact that more processed pseudogenes are generated when larger numbers of LINE 1 retrotransposons are active, support the hypothesis that gene expression levels affect the level of processed pseudogene production because the stochastic probability of a mRNA being reversed transcribed into a processed pseudogene increases with the level of gene expression of each gene and the amount of reverse transcriptase present in the cells.

Acknowledgements

We thank the two anonymous referees for their useful and constructive comments. This work was supported by a Discovery Grant from the Natural Science and Engineering Research Council of Canada to G.D.

References

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene, **238**(1): 115–134. doi:10.1016/S0378-1119(99)00227-9. PMID: 10570990.

Cureton, E.E. 1958. The average spearman rank criterion correlation when ties are present. Psychometrika, **23**(3): 271–272. doi:10.1007/BF02289240.

Drouin, G. 2006. Processed pseudogenes are more abundant in human and mouse X chromosomes than in autosomes. Mol. Biol. Evol. **23**(9): 1652–1655. doi:10.1093/molbev/msl048. PMID: 16809623.

Esnault, C., Maestre, J., and Heidmann, T. 2000. Human LINE retrotransposons generate processed pseudogene. Nat. Genet. **24** (4): 363–367. doi:10.1038/74184. PMID:10742098.

Feng, Q., Moran, J.V., Kazazian, H.H., Jr, and Boeke, J.D. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell, **87**(5): 905–916. doi:10.1016/S0092-8674(00)81997-2. PMID:8945517.

Gojobori, T., Li, W.-L., and Graur, D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18** (5): 360–369. doi:10.1007/BF01733904. PMID:7120431.

Gonçalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. Genome Res. **10**(5): 672–678. doi:10.1101/gr.10.5.672. PMID:10810090.

Graur, D., and Li, W.-H. 2000. Fundamentals of molecular evolution. 2nd ed. Sinauer Associates, Sunderland, Mass, USA.

Graur, D., Shuali, Y., and Li, W.-H. 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. **28**(4): 279–285. doi:10.1007/BF02103423. PMID:2499684.

Gu, X., and Li, W.-H. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. **40** (4): 464–473. doi:10.1007/BF00164032. PMID:7769622.

Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. U.S.A. **94**(5): 1872–1877. doi:10.1073/pnas.94.5.1872. PMID: 9050872.

Liu, G.E., Matukumalli, L.K., Sonstegard, T.S., Shade, L.L., and Van Tassell, C.P. 2006. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. BMC Genomics, **7**: 140. doi:10.1186/1471-2164-7-140. PMID:16759380.

Liu, Y.-J., Zheng, D., Balasubramanian, S., Carriero, N., Khurana, E., Robilotto, R., and Gerstein, M.B. 2009. Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity. BMC Genomics, **10**: 480. doi:10.1186/1471-2164-10-480. PMID:19835609.

McCarrey, J.R., Kumari, M., Aivaliotis, M.J., Wang, Z., Zhang, P., Marshall, F., and Vandeberg, J.L. 1996. Analysis of the cDNA and encoded protein of the human testis-specific *PGK-2* gene. Dev. Genet. **19**(4): 321–332. PMID:9023984.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, **420** (6915): 520–562. doi:10.1038/nature01262. PMID:12466850.

Ophir, R., Itoh, T., Graur, D., and Gojobori, T. 1999. A simple

method for estimating the intensity of purifying selection in protein-coding genes. Mol. Biol. Evol. **16**(1): 49–53. PMID: 10331251.

Podlaha, O., and Zhang, J. 2009. Processed pseudogenes: the "fossilized footprints" of past gene expression. Trends Genet. **25** (10): 429–434. doi:10.1016/j.tig.2009.09.002. PMID:19796837.

Sirover, M.A. 1999. New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. Biochim. Biophys. Acta, **1432**(2): 159–184. doi:10.1016/S0167-4838(99)00119-3. PMID:10407139.

Sirover, M.A. 2005. New nuclear functions of the glycolytic protein, glyceraldehyde-3-phosphate dehydrogenase, in mammalian cells. J. Cell. Biochem. **95**(1): 45–52. doi:10.1002/jcb.20399. PMID: 15770658.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**(8): 1596–1599. doi:10.1093/molbev/msm092. PMID:17488738.

Thompson, J.D., Higgins, D.J., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**(22): 4673–4680. doi:10.1093/nar/22.22.4673. PMID:7984417.

Vanin, E.F. 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. **19**(1): 253–272. doi:10.1146/annurev.ge.19.120185.001345. PMID:3909943.

Warrington, J.A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. 1999. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol. Genomics, **2**(3): 143–147. PMID:11015593.

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**(1): 631–661. doi:10.1146/annurev.bi.55.070186.003215. PMID:2427017.

Weiner, A.M. 2000. Do all SINEs lead to LINEs? Nat. Genet. **24**(4): 332–333. doi:10.1038/74135. PMID:10742088.

Zhang, Z., Carriero, N., and Gerstein, M. 2004. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. **20**(2): 62–67. doi:10.1016/j.tig.2003.12.005. PMID:14746985.