Pakistan J. Zool., vol. 48(6), pp. 1805-1810, 2016.

Possible Roles of the Pseudogenes of *Salmonella typhimurium*

Fizza Tariq, Qanita Khalid, Sheikh Arslan Sehgal, Shazia Mannan and Farrukh Jamil*
Department of Biosciences, COMSATS Institute of Information Technology, Sahiwal

ABSTRACT

Pseudogenes have long been considered as non-functional genes due to specific features in their sequences such as frame shift mutations and premature stop codons. However, several recent studies have shown that some pseudogenes are active and play vital roles in other gene's regulation. In this study, we have analyzed all the 54 pseudogenes of human pathogen *Salmonella enterica subspecie enterica serovar* Typhimurium by using various well-established computational tools. Our analysis showed that of the 16 pseudogenes pose highly stable mRNA structures, suggesting that these genes might be regulators of other homologous genes in the *Salmonella*. Moreover, predicted functions of these pseudogene derived proteins suggested their involvement in vital metabolic pathways of the microorganism, such as transport, amino acid metabolism and energy metabolism. Most importantly, our study has identified six candidate pseudogenes that can potentially be translated into thermodynamically stable proteins/enzymes, and these proteins can play some important functions for the microorganism.

Article information
Received 28 August 2015
Revised 15 October 2015
Accepted 31 October 2015
Available online 25 September 2016

Authors' Contributions:
FJ designed the study. FT and QK performed the experiments. SAS and SM collected and analyzed the data. FJ gathered the data, supervised and prepared the manuscript.

Key words:
Pseudogenes, Salmonella, Protein structure, mRNA stability

INTRODUCTION

Pseudogenes are DNA sequences that, show high sequence homology with some functional genes but, fail to transcribe or translate due to frameshift mutations and premature stop codons (Welch *et al*., 2015). Therefore, they were considered as genomic fossils (Pink *et al*., 2015). However, recent studies have shown that pseudogenes are functionally active and control other genes by their regulatory sequences (Balakirev and Ayala, 2003; Zheng and Gerstein, 2007). For example, a non-coding RNA of MYLKP1 gene inhibited the expression of its parent MYLK gene (Pei *et al*., 2012). Additionally, some pseudogenes have shown protein expression though their physiological significances have not been established (Xio-Jie *et al*., 2014).

Pseudogenes have been classified into three classes: processed, duplicated and unitary pseudogenes (Rouchka and Cha, 2009). Processed pseudogenes are derived from retro-transposition events or integration of cDNA into the genomes (Zou *et al*., 2009). On the other hand, unprocessed pseudogenes formed by segmental duplications and certain mutations (Milligan and Lipovich, 2015), while unitary pseudogenes are a subclass of unprocessed pseudogenes with no functional counterparts (Zhang *et al*., 2015).

Salmonella enterica subspecie *enterica serovar* Typhimurium (*S. typhimurium*) is a bacterial gram-negative pathogen that is considered a leading cause of human gastroenteritis (Vogels *et al*., 2011). Sequencing it's genome sequence reveals 54 pseudogenes (accession number; GCA_000006945). These pseudogenes might be significant for regulation of certain other genes of the microorganism or may play an important role in causing infection. This study was aimed to understand the possible roles of the pseudogenes in the microorganism as they can produce very different proteins with some unique functions.

METHODS

The genome of *S. typhimurium* strain LT2 hosts 54 pseudogenes. (accession number; GCA_000006945). The sequences of these pseudogenes were retrieved from NCBI and translated into protein sequences by using Transeq tool of European Bioinformatics Institute (EBI) (Goujon *et al*., 2010; Hoefman *et al*., 2014). The 18 pseudogene-derived proteins showed a reasonable sequence homology with other functional proteins (here referred as functional parents) and these were selected for further downstream analyses.

Sequence based functional prediction

Protein BLAST analysis was used to identify functional parents of the pseudogene-derived proteins (Altschul *et al*., 1990). The promoters of the pseudogenes were recognized by BPROM program (Solovyev and Salamov, 2011). Strengths of the pseudogene's promoters and their functional parents were expressed in linear

* Corresponding author: farrukhccb@gmail.com
0030-9923/2016/0006-1805 $ 8.00/0

Abbreviations: mRNA, messenger RNA; GRAVY, Grand average of hydropathy; MFE, Minimum free energy; LDF, linear discriminant function.

discriminant function (LDF) value. Their messenger RNA (mRNA) stability and folding patterns were predicted by using RNA fold web server (Zuker and Stiegler, 1981); and sub-cellular localization was predicted by using ProtCompB program. The possible functions of the pseudogene-derived proteins were obtained by ProtFun tool (Jensen *et al*., 2002, 2003); and their physiochemical properties: molecular weight, theoretical pI, aliphatic index and GRAVY (average of hydropathy) were identified by ExPASy ProtParam server (Ikai, 1980; Kyte and Doolittle 1982; Gasteiger *et al*., 2003). The tertiary structures of the pseudogene-derived proteins were predicted by SWISS-MODEL and I-TASSER server (Zhang, 2008; Biasini *et al.*, 2014). Similar tools were also applied to the corresponding functional parents of the pseudogenes.

Stability of the proteins

The stabilization centers of the pseudogene-derived proteins were evaluated by SCIde (Dosztanyi *et al*., 2003), and several parameters were obtained such as total free energy, non-bonded and electrostatic constraints in the proteins. The CaPTURE program was used to figure out the number of cation-π interactions and their energies (Gallivan and Dougherty, 1999). Moreover, the instability index was calculated by using ExPASy ProtParam server.

RESULTS

Sequence-based function prediction

Promoter analysis of the pseudogenes and their functional parents

Analyses of the 5′ UTR of the pseudogenes and their functional parents showed that 7 pseudogenes (dcoA, hutU, treB, STM 1860, STM 2764, STM 555 and STM1052) host stronger promoter as their LDF values are more (> 1.8) than their functional parents, while 5 pseudogenes host a weak promoter reign (Table S1). There are three genes (STM0326, STM3191 and STM4431) that showed 100% sequence identity with the functional parents of other species and their promoter are also very similar (Table S1).

mRNA stability prediction for the pseudogenes and their functional parents

The pseudogenes and their functional parents were evaluated for mRNA stability on basis of their minimum free energy (MFE) values. The MFE values of the pseudogenes vary from -518 to -0.7 kcal/mol, which corresponds well to MFE values of their functional parents (from -515.5 to -23.7 kcal/mol) (Table S1). Analyses showed that most of the pseudogenes host comparable mRNA secondary structure stability to those of their parent genes (Table S1). There are three pseudogenes namely ycdF, STM2664 and STM3460 that transcribed into mRNA of 15, 36 and 42 nucleotides. Such small size mRNA may act like a siRNA. The 9 pseudogenes showed 100% identical protein sequence (with 100% query coverage) to those of their parent proteins but they differ in their mRNA stabilities (Table S1). This may be attributed to presence of different codons between parent and these pseudogenes.

Functional prediction of the pseudogenes derived proteins

The functions of 18 pseudogenes derived proteins were determined by ProtFun tool. Analyses of the data showed that they are potentially involved in different physiological functions such as translation (5 enzymes), energy metabolism (4), amino acid metabolism (3), central intermediary metabolism (1), transport and binding (2) and cell envelop (1), while the functions of 2 proteins could not be determined satisfactorily (Fig. 1).

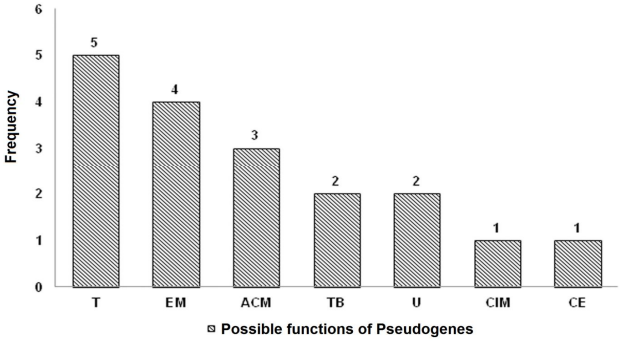

Fig. 1. Function prediction of pseudogene derived proteins. These proteins were predicted to be involved in translation (T), energy metabolism (EM), amino acid metabolism (ACM), transport and binding (TB), central intermediary metabolism (CIM) and cell envelop (CE). The functions of two proteins could not be determined satisfactorily (U).

Sub-cellular localization of the pseudogenes derived proteins

Localization of proteins indicates their possible different roles in the cell. The ProtCompB predicted that most of the pseudogene derived proteins are located in inner membrane and cytoplasm while others would be in periplasm and outer membrane (Fig. 2). The gene STM3654 showed considerable potential of protein production and secretion which may be significant for further studies as it might be involve in pathogenicity of the microorganism.

Table I.- Summary of the stability parameters of selected pseudogenes.

Sr. No	Sequence ID	Stability centers	Instability index	Total energy (kcal/mol)	Cation-π Interactions
1	STM4431 (2673766)	58	36.34	-640.26	2
2	STM3191 (1254714)	147	31.75	-220.14	4
3	STM3828.1N (2673742)	210	29.96	-3615.71	3
4	fucP (1254498)	38	38.85	-2160.06	2
5	STM0326 (1251845)	10	39.63	-392.22	0
6	STM1553 (1253071)	2	30.04	-569.19	0

Table II.- Physiochemical properties and sub-cellular localization of the pseudogenes. This table shows the molecular mass, pI, aliphatic index and sub-cellular localization of the pseudogenes.

Sr. No	Sequence ID	Molecular mass (K Da)	Theoretical pI	Aliphatic index	GRAVY	Sub-cellular localization
1	STM4431 (2673766)	38.36	6.15	89.23	-0.134	Inner membrane
2	STM3191 (1254714)	35.42	6.2	82.02	-0.182	Periplasm
3	STM3828.1N (2673742)	16.12	5.23	106.41	0.453	Inner membrane
4	fucP (1254498)	31.21	9.1	117.41	0.565	Inner membrane
5	STM0326 (1251845)	11.44	4.38	101.93	0.237	Cytoplasm
6	STM1553 (1253071)	13.78	10.83	108.92	-0.292	Cytoplasm

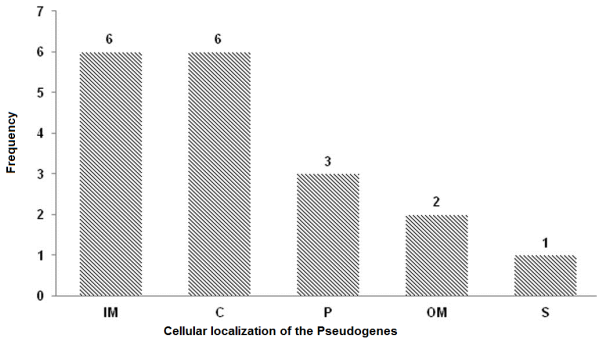

Fig. 2 Sub-cellular localization of pseudogene-derived proteins.
The proteins were detected to be localized in inner membrane (IM), cytoplasm (C), periplasm (P), outer membrane (OM) and 1 protein appears to be secreted (S).

Tertiary structure prediction of the pseudogenes and their functional parents

We generated the 3D models of the 18 pseudogene-derived proteins that showed considerable protein sequence homologies with already known structures of different enzymes of other species (Fig. 3, Table S2). The models of the 13 genes are similar as those of their parents while 5 are different.

The six models of the genes STM3828.1N, STM3191, STM4431, fucP, STM0326 and STM1553 were determined to be thermodynamically stable (Fig. 3, Table I). This stability might be attributed to the presence of several stabilization centers in these models (Shidhi *et al*., 2014**)**. The cation-π interactions were detected only in first four above mentioned models, however they were absent in the models of STM0326 and STM1553 (Table I, Fig. 4). The total free energies of these models vary from -3615.71 to -220.14 kcal/mol and analyses suggested a correlation between the numbers of stabilization centers with overall structural stability (Table I).

Physiochemical properties of the six selected pseudogenes derived proteins

The molecular masses of the pseudogene-derived proteins were found to vary from 11.14 to 38.36 KDa, indicating presence of different sizes of the proteins (Table II). The isoelectric points (pI) of the proteins vary from 4.38 to 10.83. Analyses of data showed that the 4 proteins will be acidic while 2 will basic at pH 7. Moreover, the aliphatic index values (82.02 to 117.41) of these proteins are comparable to their functional parents suggesting a similar interaction with the aqueous medium. Hydropathicity analyses showed that three proteins are hydrophobic in nature, while three are hydrophilic in nature (Table II).

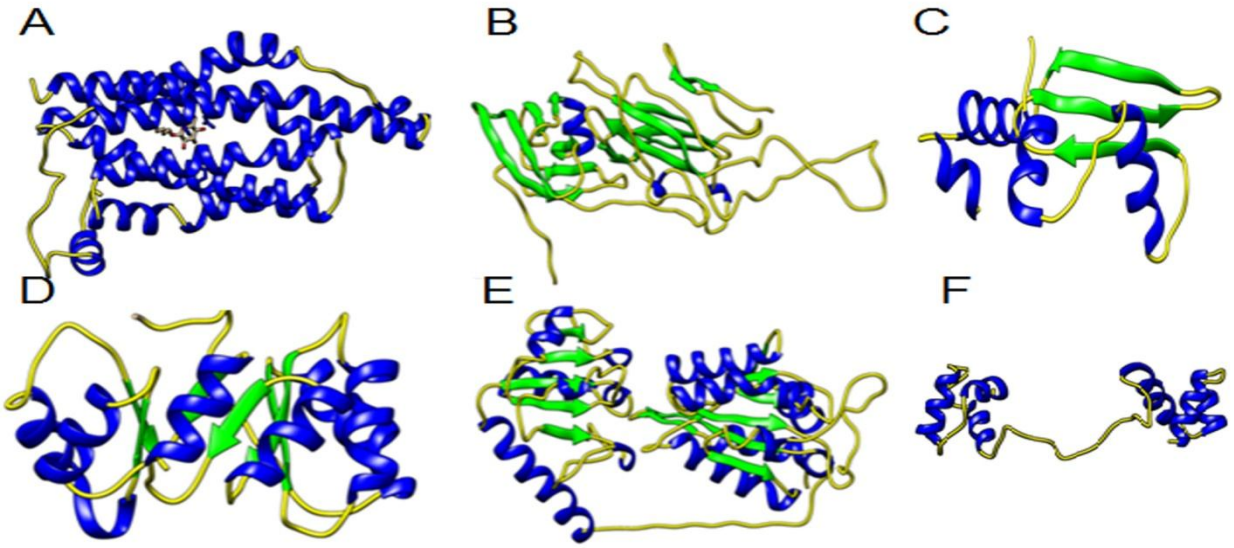

Fig. 3. Predicted 3D models of the proteins. The 3D models of fucP. A, fuP; B, STM3191; C, STM0326; D, STM3828.1N; E, STM4431 and F, STM1553.

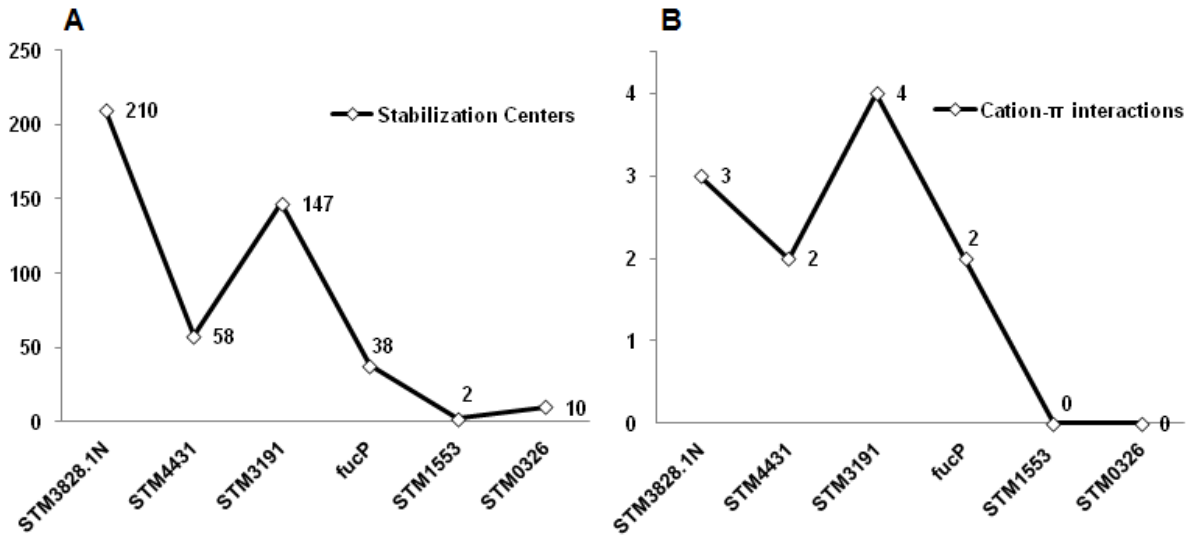

Fig. 4. Number of stabilization centers (A) and cation-π interactions (B) in the six pseudogene derived proteins models.

DISCUSSION

This study presents an innovative way to understand the possible roles of the pseudogenes of *S. typhimurium* by using different bioinformatics tools. Artificial transcription and translation of the pseudogenes indicate a strong potential of these genes to be active. Analyses showed a comparable mRNA stability of most of the pseudogenes with parent genes but only 6 of the pseudogenes could produce a stable protein. Here, we are focusing on the possible physiological functions of these stable mRNAs and proteins in light of previous research on pseudogenes of other species.

The primary structures of the pseudogenes derived

proteins showed very high sequence identity (<83%) with other bacterial enzymes such as aldolases, synthase, and transposases (Table S1), suggesting that these genes can produce active proteins if transcribe. Previous studies have shown that highly expressed genes may show less stable secondary structure of mRNA, while low expressed genes may show more stable secondary structure of mRNA (Mukund *et al*., 1999; Drummond *et al*., 2005; Dhar *et al*., 2009). We artificially transcribed these pseudogenes and free energies of their mRNAs were comparable to their parent's mRNAs (Table S1), suggesting that the pseudogenes can be transcribe into stable mRNA. Even, the mRNA products of the pseudogenes fail to translate into proteins; they may act as a regulator for other genes as some previous studies have established regulatory roles of some other pseudogenes (Dhar *et al*., 2009). For example, the mRNA of a human pseudogene PTENP1 regulates its functional parent, human tumor suppressor, PTEN gene and exert growth suppressive role (Poliseno *et al*., 2002).

The predicted tertiary structures of the six pseudogenes (STM3828.1N, STM3191, STM4431, fucP, STM0326 and STM1553) showed considerable structural homology with bacterial 2-dehydro-3-deoxy-6-phosphogalactonate aldolase, aryl sulfate sulfotransferase, acetohydroxy-acid synthase, L-fucose transporter, glycerol dehydratase activator and paired box proteins respectively (Fig. 3). These models were found to be thermodynamically stable and total free energy of these structures vary between (-3615 to -220 kcal/mol). We also checked the instability of these proteins. The stability index below 40 is considered as a good evidence of stability and it shows that proteins are stable *in vivo* (Guruprasad *et al*., 1990). All the 6 models showed stability index below 40, suggesting *in vivo* stability of these proteins (Table I). Structurally, the 3D models of these six, pseudogene, derived proteins are similar to their functional parents (Table S2), suggesting their potential to work as an active proteins and play a significant role in cellular pathways. This assumption is based on previous studies that have shown the expression of pseudogenes. For instance, a protein PGAM3 was coded by processed pseudogene in primate white blood cells (Xio-Jie *et al*., 2014). We assume that the pseudogenes of *S. typhimurium* might be expressed under certain specific conditions. However, further studies are required to establish exact roles of these genes that are made silent by the nature.

CONCLUSION

In conclusion, our study identifies 16 pseudogenes that may act as regulator of other functionally alive genes in *S. typhimurium*, and foresee its 6 genes as promising candidates for making novel proteins with some unique features.

ACKNOWLEDGEMENT

We gratefully acknowledge Higher Education Commission (HEC) of Pakistan for grants to establish Bioinformatics research laboratory at COMSATS, Sahiwal.

Statement of conflict of interest

Authors have declared no conflict of interest.

Supplementary Material Comprises two Tables. S1 and S2, which appear on weblink http://www.zsp.com.pk/pdf48/QPJZ-0453-2015%20(Supplementary%20Tables%20I%20and%20II).pdf

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.

Balakirev, E.S. and Ayala, F.J., 2003. Pseudogenes: are they "junk" or functional DNA?, *Annu. Rev. Genet.*, **37**: 123–151.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L. and Schwede, T., 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucl. Acids Res*, **42**: W252–258.

Dhar, P.K., Thwin, C.S., Tun, K., Tsumoto, Y., Maurer-Stroh, S., Eisenhaber, F. and Surana, U., 2009. Synthesizing non-natural parts from natural genomic template. *J. Biol. Eng.*, **3**: 2.

Dosztanyi, Z., Magyar, C., Tusnady, G. and Simon, I., 2003. SCide: identification of stabilization centers in proteins. *Bioinformatics*, **19**: 899–900.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. and Arnold, F.H., 2005. Why highly expressed proteins evolve slowly. *Proc. natl. Acad. Sci. USA*, **102**: 14338–14343.

Gallivan, J.P. and Dougherty, D.A., 1999. Cation-pi interactions in structural biology. *Proc. natl. Acad. Sci. USA*, **96**: 9459–9464.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A., 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucl. Acids Res.*, **31**: 3784–3788.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL–EBI. *Nucl. Acids*

Res., **38**: W695–W699.

Guruprasad, K., Reddy, B.V. and Pandit, M.W., 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, **4**: 155**–**161.

Hoefman, S., Van-Der-Ha, D., Boon, N., Vandamme, P., De-Vos, P. and Heylen, K., 2014. Niche differentiation in nitrogen metabolism among methanotrophs within an operational taxonomic unit. *BMC Microbiol.*, **14**: 83.

Ikai, A., 1980. Thermostability and aliphatic Index of globular proteins. *J. Biochem.*, **88**: 1895–1898.

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A., Knudsen, S., Krogh, A., Valencia, A. and Brunak, S., 2002. Prediction of human protein function from post-translational modifications and localization features. *J. mol. Biol.*, **319**: 1257–1265.

Jensen, L.J., Gupta, R., Staerfeldt, H.H. and Brunak, S., 2003. Prediction of human protein function according to gene ontology categories. *Bioinformatics*, **19**: 635–642.

Kyte, J. and Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *J. mol. Biol.*, **157**: 105–132.

Milligan, M.J. and Lipovich, L., 2015. Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front. Genet.*, **5**: 476.

Mukund, M.A., Bannerjee, T., Ghosh, I. and Datta, S., 1999. Effect of mRNA secondary structure in the regulation of gene expression: unfolding of stable loop causes the expression of Taq polymerase in *E. coli*. *Curr. Sci.*, **76**: 1486–1490.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L. and Mu, X.J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J. and Gerstein, M. B., 2012. The GENCODE pseudogene resource. *Genome Biol.*, **13**: R51.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L. and Carter, D.R.F., 2015. Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, **17**: 792–798.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J. and Pandolfi, P.P., 2002. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**: 1033–1038.

Rouchka, E.C. and Cha, I.E., 2009. Current trends in pseudogene detection and characterization. *Curr. Bioinform.*, **4**: 112–119.

Solovyev, V. and Salamov, A., 2011. Automatic annotation of microbial genomes and metagenomic sequences. In: *Metagenomics and its applications in agriculture. biomedicine and environmental studies* (ed. R.W. Li.), Nova Science Publishers, pp. 61–78.

Shidhi, P.R., Suravajhala, P., Nayeema, A., Nair, A.S., Singh, S. and Dhar, P.K., 2014. Making novel proteins from pseudogenes. *Bioinformatics*, **31**: 33–39.

Welch, J.D., Baran-Gale, J., Perou, C.M., Sethupathy, P. and Prins, J.F., 2015. Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. *BMC Genomics*, **16**: 113.

Xio-Jie, L., Ai-Mei, G., Li-Jaun, J. and Jiang, X., 2014. Pseudogene in cancer: real functions and promising signature. *Br. med. J.*, **52**: 17–24.

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinform.*, **9**: 40.

Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J. and Gerstein, M.B., 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.*, **11**: R26.

Zheng, D. and Gerstein, M.B., 2007. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.*, **23**: 219–224.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.H., 2009. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Pl. Physiol.*, **151**: 3–15.

Zuker, M. and Stiegler, P., 1981. Optimal computer folding of relarge RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**: 133–148.