Published in final edited form as:

Trends Genet. 2009 October ; 25(10): 429–434. doi:10.1016/j.tig.2009.09.002.

Processed pseudogenes: the ‘fossilized footprints’ of past gene expression

Ondrej Podlaha^{1,2} and Jianzhi Zhang^{1,*}

¹Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

²Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06518, USA

Abstract

Although our knowledge of the genes and genomes of extinct organisms is improving as a result of progress in sequencing ancient DNA, the transcriptomes of extinct organisms remain inaccessible, owing to the rapid degradation of messenger RNA after death. We provide empirical evidence that gene expression levels in the reproductive tissues of mice and during early mouse development correlate highly with the rate of inherited retroposition: the source of processed pseudogenes in the genome. Thus, processed pseudogenes might serve as fossilized footprints of the expression of their parent genes, shedding light on ancient transcriptomes that could provide significant insights into the evolution of gene expression.

Accessing the transcriptomes of ancient organisms

Sequencing the genomes of ancient extinct organisms is a ‘dream come true’ for evolutionary geneticists. With the unraveling of whole or partial mitochondrial and nuclear genome sequences of extinct species, including the woolly mammoth [1-3] and Neanderthal [4-6], many important evolutionary events, from human origins to past epidemics and crop domestications, can be dissected to an unprecedented resolution. For example, Neanderthal genome sequences have and will continue to inform us about the potential genetic contribution of Neanderthals to the contemporary human gene pool [4-7] and will enable the identification of human-unique genetic changes that occurred in the past few hundred thousand years of human evolution.

Research in evolutionary developmental biology repeatedly demonstrated the importance of gene expression changes in phenotypic evolution [8-10]. However, owing to the instability of message RNA (mRNA), its extraction from ancient organisms is beyond reach. Although prediction of gene expression levels and patterns from DNA sequences has been attempted in unicellular model organisms [11], it is not yet possible in complex multicellular organisms, particularly if the prediction is based on ancient DNA sequences that are incomplete or inaccurate. Although it is theoretically possible to infer the expression of a gene in an ancestral organism using the expression data from extant organisms [12], such inferences are likely to be unreliable, because gene expression evolves rapidly [13-15]. Here we explore the possibility

*Correspondence: Jianzhi Zhang, Department of Ecology and Evolutionary Biology, University of Michigan, 1075 Natural Science Building, 830 North University Avenue, Ann Arbor, MI 48109, USA, Tel: 734-763-0527, Fax: 734-763-0544, jianzhi@umich.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

that processed pseudogenes provide information about past gene expression levels and present an analytical framework for retrieving such information.

Are processed pseudogenes the fossilized footprints of past gene expression?

The genomes of many eukaryotes, including humans, are ‘littered’ with pseudogenes – defunct relatives of known genes that have lost their protein-coding ability. Many pseudogenes originate through germline retroposition, where the mRNA of a gene is reverse-transcribed to complementary DNA and reinserted into the genome at random [16]. Most of these re-inserted retrogenes lack promoters and are ‘dead-on-arrival’. They gradually accumulate open-reading-frame (ORF)-disrupting mutations, and are eventually recognizable as processed pseudogenes (Pψs). Pψs have been used for archiving splice variants [17] and for studying rates and patterns of neutral mutations [18]. In principle, the number of Pψs of a given parent gene depends on the abundance of the mRNA of the parent gene in the germline, which is determined jointly by the rates of mRNA production and degradation [19,20]. Indeed, a large proportion of Pψs are from housekeeping genes, which generally are highly transcribed and have stable mRNAs [21]. However, a recent comparison between the expression level of a gene with the number of its ‘Pψ offspring’ failed to find gene expression level to be a dominant determinant of the rate of retroposition [21]. Here we hypothesize that the present-day expression level of a gene should be best correlated with the number of its young Pψ offspring, not all Pψ offspring, because a gene's expression in the germline might not have been constant during evolution. Because retroposition occurring at earlier stages of development (i.e. embryogenesis) are more likely than that in later stages to impact the germline, we examined mouse expression data from a variety of tissues at different developmental stages (see supplementary material online). Our datasets include the comprehensive catalog of mouse pseudogenes (www.pseudogene.org) and 23 mouse microarray expression datasets from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) (Table S1).

As a proxy for Pψ age, we calculated the nucleotide sequence divergence (d) between each of 5,315 Pψs and their respective parent genes in mouse, at third codon positions, using Kimura's 2-parameter distance [22]. We then binned the pseudogenes in increments of $d = 0.05$. The age distribution of Pψs in our dataset (Figure 1a) reflects the combined effects of an exponential decay of nonfunctional sequences in the genome and an accrual of pseudogene characteristics over time. The former effect is due to the gradual deletion of nonfunctional sequences from the genome and gradual divergence of pseudogenes from their parent genes beyond recognition. The latter effect occurs because very recently formed Pψs are under-detected due to the lack of ORF-disrupting substitutions (also below). The age distribution of Pψs is also highly influenced by the cellular reverse transcriptase activity in the past, which in mammals is determined by the activity of *long interspersed element 1* (LINE1) [20]. We correlated parent gene expression in various wild-type tissues at various developmental stages with the number of Pψ offspring of different age groups. We found the correlations to be greater for expression levels in reproductive and embryonic tissues than for those in somatic tissues (e.g. lung, kidney, liver, and neural tissues) across all Pψ age categories (Figure 1b; see also Figure S1 and Table S1 in the online supplementary material) When considering all Pψs younger than d (see earlier definition of d), we observed the correlation to reach its maximum at $d = 0.25$ (Figure S1). In theory, the lower the d value, the higher the correlation. In reality, however, very recently formed Pψs are under-detected (see above). Based on the rates of point and insertion/deletion mutations in mammals, we calculated that retropseudogenes should be detectable when $d = 0.1$ (see Supplementary Methods). The likely reason that the correlation is not higher for $d = 0.1$ than $d = 0.25$ is that the former group contains much fewer genes than the latter group. We divided all Pψs into three broad age groups that may be referred to as young ($d < 0.25$), middle ($d = 0.25-0.65$), and old ($d > 0.65$) Pψs. It is apparent that the correlation is much higher for the

young group than for the middle and old groups (Figure 1b). The highest correlation was found between the parent gene expression levels during embryonic gonad development and the number of Pψ offspring with $d < 0.25$ (Pearson's $R = 0.54$, $P < 10^{-115}$; Spearman's $\rho = 0.38$, $P < 10^{-50}$; Figure 1c). For this dataset, one can indeed use the number of Pψs to predict the expression levels of their parent genes to some extent. For example, expressions of parent genes with 5 Pψs are significantly higher than those with 4 Pψs ($P = 0.003$, one-tail Mann-Whitney U test; see additional results in Table S2). Note that because microarray-based measurement of gene expression is imprecise, particularly for comparisons among genes as in the present case [14], the actual correlation between parent gene expression and Pψ offspring number is expected to be even greater. These results confirm that expression levels of parent genes affect the rate of retroposition and is consistent with the understanding that inheritable retroposition occurs in germlines or in early embryos before the separation of germlines. The relatively high correlation ($R \approx 0.3$) for some somatic tissues is probably caused by the presence of genes in the data set (especially housing-keeping genes) with correlated expression between somatic tissues and germlines.

Given the strong correlation between the number of recent Pψs and the current expression levels of their parent genes, Pψs of different ages can be treated as fossilized footprints to infer past expression levels of parent genes in embryonic or reproductive tissues, which might provide insights into the evolution of gene expression. For example, overall, 62%, 28%, and 10% of mouse Pψs fall into young ($d < 0.25$), middle ($d = 0.25-0.65$), and old ($d > 0.65$) age groups, respectively (Figure 2a). By contrast, mouse *Hmgb2* (encoding the high mobility group B 2 protein), which is important for male gonad development and spermatogenesis [23], gave rise to 65 Pψs, of which 16 (25%), 8 (12%), and 41 (63%) belong to the three age groups, respectively. This age distribution is significantly different from the genomic distribution (χ^2 test, $P < 10^{-45}$) and shows a decrease in the expression of *Hmgb2* (compared with the genomic average) in the evolution leading to mouse (Figure 2a). By examining all parent genes in a likewise fashion, we observed that 59 (3.2%) mouse and 37 (1.9%) human parent genes differ significantly from their respective genomic averages in age distribution at a false discovery rate of 1%, presumably resulting from evolutionary changes in expression level in embryonic tissues or germlines (Figure 2b; Table S3). A similar approach can be used to study the evolutionary trajectory of expression divergence between duplicate genes.

Caveats to this approach

Here we showed strong correlation between the number of recent Pψs and the current expression levels of their parent genes in mouse. Because this phenomenon is a result of the retroposition rate being determined by the abundance of mRNA, we predict that similarly strong correlations also exist in other organisms, although it is not feasible to validate this prediction at this time, owing to the lack of appropriate gene expression data from germlines or during early development. We illustrated the utility of Pψs in analyzing gene expression evolution. The overall accuracy of similar analyses in any extant or ancient genome depends on three important factors: identification of Pψs, identification of the parent genes of Pψs, and the estimation of Pψ ages. Identification of Pψs is not a trivial task and it is usually executed by searching in a sequenced genome for protein-coding-gene homologs that contain ORF-disrupting substitutions and lack introns. By design, this approach is limited to Pψs of intron-containing parent genes, due to the difficulty in distinguishing between processed and non-processed pseudogenes of intron-less parent genes (~10% in mouse [24]). Furthermore, because it takes on average a few million years for ORF-disrupting mutations to occur and get fixed in Pψs (of humans and related primates) [25], very young Pψs will be under-detected, hampering the study of parent gene expression in the very recent past. Other factors, such as the *E*-value cutoff used in the BLAST search of homologous sequences, also affect the identification of Pψs. For example, very old Pψs are hard to detect even with loose *E*-value

cutoffs, simply because these P Ψ s are too different in DNA sequence from their parent genes. But, as long as the same criteria are applied to all P Ψ s, the bias for individual parent genes is expected to be minimal.

Currently, P Ψ s are assigned to their functional parent genes by BLAST searches and are assumed to have arisen from their parent genes through retroposition. However, it is possible that a P Ψ is secondarily generated by simple duplication of another P Ψ , rather than by retroposition of its functional parent gene. In theory, the number of duplication-generated P Ψ s is expected to be proportional to the number of retroposition-generated P Ψ s. Thus, treating all P Ψ s as retroposition-derived should not bias the age distribution of a parent gene relative to that of the genomic average. In reality, however, segmental duplication rate varies widely across the genome and it might be necessary to differentiate between duplication-generated and retroposition-generated P Ψ s. For this purpose, phylogenetic analysis will be useful, as the sister clade of each duplication-generated P Ψ would not include the functional parent gene. Another complication in identifying the parent gene of a P Ψ is that the parent gene might no longer exist, for two reasons. First, the parent gene might have been lost during evolution, which will lead to the identification of a wrong parent gene or no parent gene, depending on whether the true parent gene has a paralog in the genome. Second, a parent gene might have been duplicated to generate two (or more) daughter genes since the birth of the P Ψ that is under investigation. In this case, BLAST can assign the P Ψ to either daughter gene, depending on which one evolves more slowly, although the true parent gene should be the common ancestor of these daughter genes and can be identified through phylogenetic analysis.

We measured the age of a P Ψ by its nucleotide sequence divergence d from its functional parent gene at the parent gene's third codon positions. This approach is applicable to both retroposition-generated and duplication-generated P Ψ s. However, it could lead to an underestimation of the ages of P Ψ s of more conserved parent genes than those of less conserved parent genes. This is because, even at third codon positions, a substantial fraction of mutations are nonsynonymous (~31% if there is no transition/transversion bias), which are less likely to be fixed in a conserved parent gene than in a less conserved one. Thus, for two P Ψ s of the same true age, the one derived from the more conserved parent gene tends to look younger than the one derived from the less conserved parent gene. This bias could be alleviated if one calculates d using only four-fold degenerate sites of parent genes, but such d estimates might have larger variances owing to smaller numbers of usable sites. Because transcribed but untranslated regions (UTRs) of parent genes are also frequently present in P Ψ s, they may be used for estimating d as well. However, because UTRs of parent genes are subject to stronger functional constraints than four-fold degenerate sites [26], they are less ideal for P Ψ age estimation.

We used the P Ψ information to infer relative changes of gene expression during evolution. Is it possible to infer the absolute gene expression level in the past? We found that even among parent genes with the same number of young P Ψ s, their log₂-transformed microarray expression levels still vary substantially, with a standard deviation that often reaches a quarter of the mean (Table S2). Thus, it is probably difficult to infer precisely the absolute expression level of a gene in the past from the number of its P Ψ s.

These caveats notwithstanding, our results point to a fascinating observation that the P Ψ record embedded in the genome of an organism provides a uniquely informative glimpse of past gene expression levels, which complements the knowledge gained from sequencing ancient extinct genomes to provide a more comprehensive picture of evolution. A recent study suggested that gene expression differences between species are generally smaller at the protein level than at the mRNA level [27]. Our approach can infer mRNA expression changes, but not protein expression changes. Future research should improve the reliability of the retrieval of ancient transcriptome information and apply it to the study of evolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Zhang laboratory and three anonymous reviewers for valuable comments. This work was supported by research grants from the National Institutes of Health to J.Z.

References

1. Miller W, et al. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 2008;456:387–390. [PubMed: 19020620]
2. Krause J, et al. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 2006;439:724–727. [PubMed: 16362058]
3. Rogaev EI, et al. Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *PLoS Biol* 2006;4:e73. [PubMed: 16448217]
4. Green RE, et al. Analysis of one million base pairs of Neanderthal DNA. *Nature* 2006;444:330–336. [PubMed: 17108958]
5. Noonan JP, et al. Sequencing and analysis of Neanderthal genomic DNA. *Science* 2006;314:1113–1118. [PubMed: 17110569]
6. Green RE, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 2008;134:416–426. [PubMed: 18692465]
7. Krings M, et al. Neandertal DNA sequences and the origin of modern humans. *Cell* 1997;90:19–30. [PubMed: 9230299]
8. Stern DL, Orgogozo V. The loci of evolution: how predictable is genetic evolution? *Evolution* 2008;62:2155–2177. [PubMed: 18616572]
9. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 2008;134:25–36. [PubMed: 18614008]
10. Wray GA. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 2007;8:206–216. [PubMed: 17304246]
11. Beer MA, Tavazoie S. Predicting gene expression from sequence. *Cell* 2004;117:185–198. [PubMed: 15084257]
12. Gu X. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics* 2004;167:531–542. [PubMed: 15166175]
13. Khaitovich P, et al. A neutral model of transcriptome evolution. *PLoS Biol* 2004;2:E132. [PubMed: 15138501]
14. Liao BY, Zhang J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* 2006;23:530–540. [PubMed: 16280543]
15. Gu Z, et al. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 2002;18:609–613. [PubMed: 12446139]
16. Kaessmann H, et al. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* 2009;10:19–31. [PubMed: 19030023]
17. Shemesh R, et al. Genomic fossils as a snapshot of the human transcriptome. *Proc Natl Acad Sci U S A* 2006;103:1364–1369. [PubMed: 16432206]
18. Graur D, et al. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol* 1989;28:279–285. [PubMed: 2499684]
19. Pavlicek A, et al. Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends Genet* 2006;22:69–73. [PubMed: 16356584]
20. Zhang Z, et al. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet* 2004;20:62–67. [PubMed: 14746985]
21. Balasubramanian S, et al. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* 2009;10:R2. [PubMed: 19123937]

22. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;16:111–120. [PubMed: 7463489]
23. Ronfani L, et al. Reduced fertility and spermatogenesis defects in mice lacking chromosomal protein Hmgb2. *Development* 2001;128:1265–1273. [PubMed: 11262228]
24. Sakharkar KR, et al. Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Lett* 2006;580:1472–1478. [PubMed: 16469316]
25. Zhang J, Webb DM. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc Natl Acad Sci U S A* 2003;100:8337–8341. [PubMed: 12826614]
26. Li, WH. *Molecular Evolution*. Sunderland, Mass: Sinauer; 1997. p. p184
27. Schrimpf SP, et al. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol* 2009;7:e48. [PubMed: 19260763]

Figure 1.

Number of young processed pseudogenes correlates with parent gene expression in mouse.

(a) The age distribution of mouse processed pseudogenes. Age is measured by Kimura's 2-parameter nucleotide distance between a pseudogene and its parent gene at third codon positions. (b) Mean Fisher's Z transformed correlations of parent gene expression and the number of processed pseudogene offspring of different age groups. $Z = 0.5 \ln[(1+R)/(1-R)]$, where R is Pearson's correlation coefficient. Unlike R , Z is normally distributed. Error bars represent one standard error. Note that 'reproductive' includes embryonic expressions that potentially predate the separation of germlines. (c) Pearson's correlation between the number of young processed pseudogenes ($d < 0.25$) and their parent gene expression in mouse embryonic testis 14.5 days past conception ($R = 0.54$, $P < 10^{-115}$). For the same data set, Spearman's rank correlation is $\rho = 0.38$ ($P < 10^{-50}$). When the apparent outlier that has the highest number of pseudogene offspring is removed, Pearson's correlation becomes 0.58 ($P < 10^{-124}$) and Spearman's correlation becomes 0.37 ($P < 10^{-50}$).

<u>Mouse</u>			<u>Human</u>		
Gene Symbol	P - Value	Q - Value	Gene Symbol	P - Value	Q - Value
<i>0610038D11Rik</i>	1.04E-59	1.92E-56	<i>C19orf48</i>	4.29E-174	8.25E-171
<i>H3f3b</i>	3.43E-56	3.18E-53	<i>NRG4</i>	4.93E-119	4.75E-116
<i>Hmgb2</i>	8.24E-46	5.09E-43	<i>AMOTL2</i>	2.41E-35	1.54E-32
<i>Ldhb</i>	3.39E-42	1.57E-39	<i>AL139099.3</i>	9.08E-28	4.37E-25
<i>Pcdhgc5</i>	1.35E-24	4.99E-22	<i>FAM95B1</i>	1.86E-24	7.14E-22
<i>Rps6</i>	1.30E-22	3.48E-20	<i>H3F3B</i>	1.87E-15	6.00E-13
<i>Olfir1317</i>	1.31E-22	3.48E-20	<i>HMGB2</i>	1.82E-12	4.99E-10
<i>Eef1a2</i>	1.16E-12	1.79E-10	<i>AC005253.1</i>	1.39E-10	3.34E-08
<i>Olfir1255</i>	1.16E-12	1.79E-10	<i>AC093582.3</i>	4.48E-10	9.59E-08
<i>Myl3</i>	1.16E-12	1.79E-10	<i>AL645568.9</i>	1.54E-09	2.97E-07
<i>Mrgpra6</i>	1.16E-12	1.79E-10	<i>VNIR4</i>	3.04E-09	5.32E-07
<i>Olfir1194-ps1</i>	1.16E-12	1.79E-10	<i>GAPDH</i>	6.72E-09	1.08E-06
<i>Mark2</i>	1.13E-10	1.50E-08	<i>TCEB1</i>	1.05E-08	1.55E-06
<i>Olfir1052</i>	1.13E-10	1.50E-08	<i>PTS</i>	1.47E-08	2.02E-06
<i>9030619P08Rik</i>	1.33E-10	1.54E-08	<i>EEF1A2</i>	2.15E-08	2.59E-06
<i>Olfir1020</i>	1.33E-10	1.54E-08	<i>AC018865.4</i>	2.15E-08	2.59E-06
<i>Olfir1212</i>	1.42E-10	1.55E-08	<i>RPL12</i>	4.16E-08	4.71E-06
<i>Mageb1</i>	3.66E-09	3.77E-07	<i>CIC</i>	1.40E-07	1.50E-05
<i>Tmem149</i>	1.10E-08	9.30E-07	<i>PCDHGC5</i>	1.96E-07	1.99E-05
<i>Gapdhs</i>	1.10E-08	9.30E-07	<i>AL138795.21</i>	3.34E-06	3.21E-04

Figure 2.

Variation of gene expression level over evolutionary time, probed from biased age distributions of processed pseudogenes. (a) Examples of parent genes whose processed pseudogenes have age distributions significantly different from the genomic distribution (eukaryotic translation initiation factor 1 [*Eif1*], high mobility group box 2 [*Hmgb2*], and ribosomal protein S6 [*Rps6*]) or similar to the genomic distribution (ribosomal protein L7 [*Rpl7*]). (b) The top 20 mouse and human parent genes whose processed pseudogenes have age distributions significantly different from the genomic distribution (ranked by *Q*-values). *P*-values are from χ^2 tests and *Q*-values are false discovery rates.