Available online at www.sciencedirect.com
ScienceDirect
Gene 389 (2007) 196–203
GENE
SECTION
FUNCTIONAL GENOMICS
www.elsevier.com/locate/gene

Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes

Hiroaki Sakai [a,b,c], Kanako O. Koyanagi [d], Tadashi Imanishi [d,e], Takeshi Itoh [c,e], Takashi Gojobori [e,f,*]

[a] Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Bio-IT Research Bldg. 7F, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan
[b] Kyowa Hakko Kogyo Co. Ltd., 1-6-1 Ohtemachi, Chiyoda-ku, Tokyo 100-8185, Japan
[c] Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan
[d] Graduate School of Information Science and Technology, Hokkaido University, North 14, West 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan
[e] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Bio-IT Research Bldg. 7F, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan
[f] Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Received 6 October 2006; received in revised form 11 November 2006; accepted 14 November 2006
Available online 18 November 2006
Received by N. Okada

Abstract

Despite the wide distribution of processed pseudogenes in mammalian genomes, such as those of human and mouse, relatively little is known about their roles in genomic evolution. While gene duplications are recognized as one of the major driving forces in genome evolution, processed pseudogenes, which are retrotransposed copies of mRNAs, have been regarded as junk or selfish DNA for a long time. In order to elucidate the quantitative and qualitative contribution of processed pseudogenes to the mammalian genome evolution, we attempted to detect processed pseudogenes by extensively mapping the mRNAs to both the human and mouse genomes, and then we estimated the rate of their emergence. As a result, we revealed that the rate of pseudogene emergence was about 1–2% per gene per million years, which was as high as the rate (0.9%) of gene duplication in the human genome, although the rate of pseudogene emergence was found to drastically decrease in the hominid lineage. Furthermore, 1% of the processed pseudogenes seemed to be reinvigorated by post-retrotransposition transcription, many of them preserving the intact coding regions. Since the expression patterns of transcribed pseudogenes in various tissues were quite different between human and mouse, their emergence might have led to species-specific evolution. Our results indicate that the generation of processed pseudogenes was not wholly futile but instead has been an indispensable resource, driving dynamic evolution of the mammalian genomes.

Keywords: Processed pseudogene; Substitution rate; Expression profile

1. Introduction

Although intuition tells us that evolution should have compiled genome sequences in optimal forms by relinquishing unnecessary portions, controversies over nonfunctional genomic regions, which are generally referred to as junk or selfish DNA (Ohno, 1972; Orgel and Crick, 1980), suggested that there seemed to be selectively neutral regions of DNA that had a long residency period in a genome. In fact, it is widely recognized that nearly half of higher eukaryote genomes consist of transposon-derived repetitive sequences (Lander et al., 2001;

Abbreviations: LINEs, long interspersed elements; ORF, open reading frame; NG method, Nei–Gojobori method; mNG method, modified Nei–Gojobori method; ML method, maximum likelihood method; MYA, million years ago; TPP, transcribed processed pseudogene; UTR, untranslated region.

[*] Corresponding author. Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. Tel.: +81 55 981 6847; fax: +81 55 981 6848.

E-mail address: tgojobor@genes.nig.ac.jp (T. Gojobori).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.11.007

Waterston et al., 2002), while the repeats may play a role of biological significance (Nowak, 1994; Makalowski, 2000). Moreover, analyses of fully sequenced genomes have revealed that the genomes of human and other mammals possess a large number of pseudogenes whose corresponding protein functions are disabled (Goncalves et al., 2000; Harrison et al., 2003; Torrents et al., 2003).

In particular, processed pseudogenes, which are generated by reverse transcription of mRNAs and subsequent reintegration of their DNAs into the genome with the machinery of long interspersed elements (LINEs) (Vanin, 1985; Weiner et al., 1986; Mighell et al., 2000), appear to be almost “dead on arrival,” because they lack functional promoters. The fact that their protein-coding regions are in many cases disrupted by nonsense or frameshift substitutions supports this idea. Thus, on one hand, one can believe that processed pseudogenes are an oddity that makes up a proportion of ‘junk DNA’ and that they have not made any contributions to molecular evolution of the species. On the other hand, however, duplicated genes are thought to be indispensable evolutionary resources. Under the framework of the neutral theory of molecular evolution a gene cannot alter its function, but duplication events allow to relax purifying selection on copies of genes, so that these genes may later acquire novel functions (Nei, 1969; Hughes, 1994; Lynch and Force, 2000; Moore and Purugganan, 2003). Genome-wide or segmental duplication especially may be an important driving force behind genome evolution because it gives rise to a number of new functions concurrently (Ohno, 1970; Wolfe, 2001). It was reported that ~5% of the human genome had been generated by segmental duplications (Bailey et al., 2002). Moreover, the rate of gene duplication was estimated to be as high as that of nucleotide substitutions (Lynch and Conery, 2000). Therefore, when compared to pseudogenes, duplicated genes seem to play a more important role in evolution. However, it cannot be concluded decisively because the extent of mRNA retrotransposition and evolutionary contribution of processed pseudogenes have not yet been evaluated either quantitatively or qualitatively.

Here we have attempted to demonstrate the tempo and mode of processed pseudogenes in molecular evolution by examining the complete genomes and mRNA sequences of humans and mice. Several efforts had been made to detect pseudogenes by correlating sequence similarity to known proteins in combination with reported characteristics of pseudogenes such as, higher ratio of nonsynonymous-to-synonymous substitutions and genomic polyadenylation tail (Harrison et al., 2003; Torrents et al., 2003; Zhang et al., 2003; Zhang and Gerstein, 2004). However, as the number of extant processed pseudogenes had remained unclear, we thoroughly compared the mRNAs available in the DNA databanks with the human and mouse genomes (Imanishi et al., 2004), extensively searching for all intronless homologs that were convincing candidates of processed pseudogenes (Grimwood et al., 2004). The rates of emergence and loss of processed pseudogenes were estimated by using the processed pseudogenes detected. Moreover, the reinvigoration of processed pseudogenes by post-retrotransposition transcription is discussed.

2. Materials and methods

2.1. Mapping of mRNA sequences to the genomes

The mRNA sequence collection used in this study consisted of 113,196 human mRNA sequences and 101,096 mouse mRNA sequences deposited in DDBJ release 54. We masked all the repetitive and low complexity sequences in these mRNA sequences by RepeatMasker (Smit and Green, unpublished) with Repbase 7.5. We discarded those mRNAs that had short nonrepetitive nucleotide sequences (<30 bp in length) after they were repeat-masked. In order to perform the mapping, firstly we conducted BLASTN 2.2.6 (Altschul et al., 1997) searches for all the mRNA sequences against the genome sequences (NCBI build 34 for human and build 30 for mouse) and extracted corresponding genomic regions for each query sequence. Second, we used est2genome (EMBOSS package version 2.7.1) (Mott, 1997; Rice et al., 2000) to align the mRNA sequence to corresponding genomic regions with a threshold of ≥95% identity and ≥90% coverage. Finally, if an mRNA sequence could be mapped to multiple positions in the human and mouse genomes, we selected a single locus by examining the nucleotide identity, length coverage and number of exons of each hit.

2.2. Detection of processed pseudogenes

We detected processed pseudogene candidates by searching for loci that were homologous to mRNAs mapped to the genome by the aforementioned procedure. We conducted est2genome analysis for all these genomic portions, and all alignable regions were employed. The number of introns for each candidate region was compared with that of its parental locus. If one or more intron(s) was missing, this region was defined as a processed pseudogene. Gaps of less than 80 bp were allowed and not treated as introns because introns of less than 80 bp made up less than 0.5% of all the mapped mRNA sequences.

Transcribed processed pseudogene candidates were detected, when more than 70% of their genomic regions were covered by mRNAs and vice versa. These candidates were manually checked and selected if the mRNAs were virtually identical to the processed pseudogene regions in the genome and they could be unambiguously mapped to single positions in the genome.

2.3. Age distribution of processed pseudogenes

Each processed pseudogene sequence was aligned with the ORF sequence of a corresponding functional gene by the FASTY program (version 3.4t22) to construct a pairwise alignment of protein sequences (Pearson, 2000). The coding region in the processed pseudogene was then determined from the protein alignment and the codon-based alignment was constructed by FASTY and Clustal W (version 1.83) (Thompson et al., 1994). The numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions in the processed pseudogenes were calculated by using the Nei–Gojobori (NG) method (Nei and Gojobori, 1986).