Ilenia D’Errico is a postdoctoral fellow in the Department of Biochemistry and Molecular Biology, University of Bari, Italy.

Gemma Gadaleta is Professor of Molecular Biology at the Faculty of Science, University of Bari.

Cecilia Saccone is Professor of Molecular Biology at the University of Bari and Head of Division of Genomics and Bioinformatics at Istituto Tecnologie Biomediche-Consiglio Nazionale delle Ricerche, Bari.

Keywords: processed pseudogenes, non-processed pseudogenes, mitochondrial pseudogenes, mitochondrial transcription factor A (mtTFA)

INTRODUCTION
Pseudogenes were originally defined as DNA sequences structurally similar to functional genes but containing important defects, which make them unable to produce functional proteins. Such defects include, for example, the loss of the start codon, the presence of additional stop signals and the lack or abnormality of flanking regulatory regions. Nevertheless, this definition has to be revised in the light of data showing the possibility that pseudogenes can acquire novel functions during evolution (see below).

Pseudogenes can be distinguished into three categories: processed, non-processed and mitochondrial pseudogenes. The presence of pseudogenes is well known in prokaryotes, where they represent genes dying and disappearing from the genome in response to a fundamental niche change. For instance, in Mycobacterium tuberculosis, the origin of pseudogenes is ascribed to the loss of different sets of sigma factors in different moments of the evolution of this species. In eukaryotes, pseudogenes have been identified in different living organisms including plants, insects and vertebrates.

In this paper, metazoan pseudogenes, particularly those present in the completely sequenced genomes, will be reviewed. In addition, some experimental data obtained by the authors’ group on the processed pseudogenes of the mitochondrial transcription factor A (mtTFA) in different mammals will be reported.

NON-PROCESSED PSEUDOGENES
Non-processed pseudogenes are usually found on the same chromosome inside clusters of similar functional sequences; they may possess introns and flanking regulatory sequences like the functional gene. They usually originate from a gene duplication mechanism producing an extra copy of the gene which, being unnecessary, can accumulate mutations without damaging the organism, but they can also be generated by unequal crossing-over mechanisms. Premature stop codons, frameshift
mutations, disablement of regulatory regions and alterations in splice sites are the most obvious characteristics of pseudogenes. Thus, the copies of most duplicated genes are expected to become non-functional. Alternatively, pseudogenes may continue to drift until they are either deleted or become unrecognisable as a genetic copy. Depending on the order in which mutations accumulate over evolutionary time, a duplicated pseudogene may still be transcribed and, even if rarely, could become a functional unit acquiring a novel function or mode of expression and become fixed within a population. An example of this phenomenon is the human α-globin cluster of genes on chromosome 16 that has arisen by gene duplication and divergence. This cluster includes ζ2, which is expressed in the embryonic yolk sac, and its non-processed pseudogene Ψζ1. The latter has a non-functional promoter but, in some individuals, its gene conversion by ζ2 has resulted in restoration of a functional promoter and the generation of ζ1 from Ψζ1. Another reported case is that of bovine seminal ribonuclease, which has lain dormant for about 20 million years and which then appears to have been resurrected to form a functioning gene — probably via a gene conversion event. Most of the known processed pseudogenes produced by a retrotransposition process lose their functionality as a consequence of defects in the mechanism generating them. In fact, reverse transcription is a process producing errors, and a lot of changes between the template RNA and the complementary DNA (cDNA) can be accumulated. Moreover, unless the processed gene is transcribed by RNA polymerase III, it should not contain the promoter, which usually lies in non-transcribed regions and, so is quite likely to be inactive even though its coding region is intact.

Finally, a processed gene can be inserted in a genomic localisation inappropriate for its expression which can also be a different chromosome compared with its functional counterpart. For this reason, a processed gene is ‘dead on arrival’ in most cases. Nevertheless, reverse transcription polymerase chain reaction experiments have shown, through transcript identification, that some pseudogenes can be transcribed and, in some cases, they can have a different role from that of the original gene. Even if they do not possess all the transcriptional control regions present in the functional gene, they can use other transcriptional elements. For example, pseudogene transcription can be directed by a promoter apparently near a non-correlated sequence.
Due to the ubiquity of reverse transcription, mammalian genomes are literally bombarded by copies of retrotranscribed sequences, and most of these copies become non-functional as soon as they integrate in the genome. Moreover, these sequences cannot be easily repaired through the gene conversion process, because they are mostly placed at long chromosomal distances from the parent functional gene. The generation of defective copies of a functional locus and their scattering throughout the genome has been compared to a volcano producing lava — that is why the process is called ‘Vesuvian mode’ evolution.22

As soon as a retropseudogene settles within a chromosome, it undergoes two different evolutionary processes.23 The first process involves a rapid accumulation of point mutations which can hide the similarity between the pseudogene sequence and its functional homologue, which evolves much more slowly. The processed pseudogene nucleotide composition will tend to resemble more and more the surrounding non-functional region, enabling the pseudogene to blend with it. This process is called ‘compositional assimilation’.

The second evolutionary process involves the reduction of pseudogene size compared with the functional gene. This shrinkage is caused by an excess of deletions over insertions. It has been estimated that a processed pseudogene loses about one-half of its DNA in nearly 400 million years. This process is so slow that the human genome, for instance, still contains a large quantity of pseudogene DNA related to very distant ancestors.17

Obviously, these ancient pseudogenes have often lost almost all their similarity with the functional genes. The shrinkage is too slow a process to counterbalance the increase in genome dimensions which results from the continuous Vesuvian eruptionism. So, the restriction in pseudogene number in the genome is probably due to other factors, such as natural selection (Figure 1).17

MITOCHONDRIAL PSEUODGENES

A few decades ago, the presence of sequences of many animal species having significant homology to mitochondrial DNA inside the nucleus was ascertained. These nuclear insertions of mitochondrial DNA are called pseudogenes because, unlike their homologous counterparts, they are not transcribed or translated into functional proteins, owing to the different mitochondrial genetic code.24

The integration process of the mitochondrial fragments in the nucleus is probably very ancient; indeed, it is thought to have started soon after the settling of the first endosymbiont as an organelle.25 In fact, at least as far as the animal line is concerned, there has been a progressive thinning of the mitochondrial genome as a consequence of the transfer of genes coding for mitochondrial components to the nucleus. Thus, the unsuccessful transfers could have given rise to pseudogenes. There are essentially two mechanisms which could explain the integration of these mitochondrial fragments in the nucleus: direct DNA transfer26 and RNA-mediated transfer.27 Most of the experimental data available support the hypothesis that transfer is by DNA, although an origin of mitochondrial pseudogenes from RNA cannot be excluded.28

From the available literature, it seems that mitochondrial pseudogenes are not equally distributed in all species — they are abundant in mammals and birds, but seem to be almost completely absent in fish.29 Very few are found in Caenorhabditis elegans (two pseudogenes) and Drosophila melanogaster (three pseudogenes) compared with Homo sapiens (354 pseudogenes).

WHY ARE PSEUODGENES INTERESTING?

Pseudogenes are interesting in many aspects. Certainly, pseudogenes are important in molecular evolution and in assessing how organisms have adapted to ensure their survival. Moreover, if
Pseudogenes prove to be devoid of function, these sequences turn out to be ideal for the study of spontaneous mutation patterns as any mutation arising in them can settle in a population. For example, since the absolute rate of mutation which occurs in the nucleus proves to be lower than the one taking place in mitochondria, mitochondrial pseudogenes can be considered as 'photographs' of the mitochondrial DNA at the time of transfer, since they evolve more slowly than their functional counterparts in the mitochondrion.

In the case of processed pseudogenes deriving from mRNA, the comparison between the two types of sequences (original or retrotranscribed copy) may bring to light some features of the mRNA expressed only in particular conditions, or now extinct or difficult to clone and analyse. An example of this kind of information, obtained through the analysis of processed pseudogenes, comes from the retrotransposed copies of high mobility group (HMG) genes, where the alignment between different retropseudogenes of the HMGB1 gene coming from different chromosomes and the functional sequence has proven the existence of longer transcripts than those already known. The increase in the size of pseudogenes seems to be due in some cases to an extension of the 3′ extremity and, in others, to the 5′ extremity. As shown for HMG pseudogenes, a longer 3′ end could have originated using alternative polyadenylation signals; the extension of the 5′ end could be due to an inaccurate identification of the correct start site of the gene transcription.

As far as the transcribed pseudogenes are concerned, an example of a transcript from a pseudogene 'provided with function' has been described for the neural nitric oxide synthase in the snake.

---

### Pseudogene origin and features

<table>
<thead>
<tr>
<th>Pseudogene origin</th>
<th>Pseudogene features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processed pseudogenes</strong></td>
<td></td>
</tr>
<tr>
<td>[transcription]</td>
<td>• Absence of both 5′ promoter sequence and introns</td>
</tr>
<tr>
<td>[retrotranscription]</td>
<td>• Presence of flanking direct repeats</td>
</tr>
<tr>
<td>[integration]</td>
<td>• Presence of a 3′ polyadenylation tract</td>
</tr>
<tr>
<td>[integration]</td>
<td>• Random integration into the genome</td>
</tr>
<tr>
<td><strong>Non-processed pseudogenes</strong></td>
<td></td>
</tr>
<tr>
<td>[gene duplication]</td>
<td>• Presence of introns</td>
</tr>
<tr>
<td></td>
<td>• Often adjacent to their paralogous functional gene</td>
</tr>
</tbody>
</table>

---

**Figure 1**: Pseudogenes origin and features are described. Genes are represented by black (exons) and white (introns) blocks. Single and double zigzag lines represent poly A (+)mRNA and cDNA produced by retrotranscription, respectively.
Here, the pseudogene has a segment which proves to be complementary and inverted compared with the normal gene and, through the formation of an RNA duplex, it interferes with the expression of the neural nitric oxide synthase.

The expression of transcribed pseudogenes can vary considerably compared with the expression of homologous living genes. For instance, for the 5-HT7 receptor, the transcripts of a pseudogene are expressed in various tissues while the transcripts of the corresponding functional gene are absent. Transcripts of pseudogenes can increase their expression in tumour cells such as those of laryngeal carcinoma or of glioblastoma.

In the literature, cases of pseudogenes which are transcribed and translated can also be found — namely, the retropseudogene of the rat preproinsulin I; one of the mouse pseudogenes for the phosphoglycerate kinase; some human genes, such as the homoprotein pseudogene HPX42B coding for a new tumoural antigen recognised by T CD8(+) cells; and the testicle-specific form of the ß-subunit of the pyruvate dehydrogenase E1 coded by a gene without introns on chromosome four. ABUNDANCE OF PSEUDOGENES IN DIFFERENT ORGANISMS

The number of pseudogenes in a genome obviously depends on the relative rates of gene duplication and pseudogene loss. Data about the relationship between the size of a pseudogene population and the size of a proteome, and the amount of coding and non-coding DNA in genomes as whole entities, are still unclear. It has been proposed that C. elegans has a very high rate of gene duplication, generating 383 duplicated genes every million years, compared with, for example, 31 for D. melanogaster. A large proportion of these duplicated genes are pseudogenes; at least 4 per cent of the annotated genes in the C. elegans genome can be recognised as pseudogenes.

Processed pseudogenes deriving from mRNA are abundant in mammals, although there is a tendency to underestimate their number because it may be difficult to distinguish non-processed pseudogenes from functional genes. In some cases, the number of processed pseudogenes is very high; for instance, in the mouse genome it has been found that about 200 processed pseudogenes are derived from a single gene for glyceraldehyde-3-phosphate dehydrogenase. Although the processed pseudogenes are abundant in mammals, they are relatively rare in other organisms such as chickens, amphibians and Drosophila. A hypothesis is that this is due to differences in gametogenesis between mammals and other organisms; indeed, pseudogenes are fixed in a population through the germinal line.

Initial surveys of processed pseudogenes suggest that their occurrence is largely based on two factors: (1) the expression levels and (2) the amount of intergenic DNA available for insertion. The first factor could explain the large number of processed pseudogenes for ribosomal proteins found in the human genome. The second factor explains the large number of processed pseudogenes in mammals by comparison with the worm and fruitfly (D. melanogaster). There is evidence that the latter has few pseudogenes, and this observation seems to be linked to a high deletion rate of genomic DNA. Approximately 100 pseudogenes have been found in D. melanogaster (with at least one-sixth of these as candidate processed pseudogenes), which is about one pseudogene for each of the 130 proteins encoded in the genome (Table 1).

Recent surveys have shown the presence of about 5,000 processed pseudogenes in the mouse genome and about 8,000 processed pseudogenes in the human genome. So, although the sizes of the mouse and human genomes are very similar, the number of processed...
pseudogenes in mouse seems to be only one-half of that observed in human. This phenomenon seems not to be due to a lower retrotransposition activity in the mouse, because it is known that the mouse genome has higher nucleotide substitution, insertion and deletion rates than the human genome.23,50 The reason why genomes that are similar in size show a different number of pseudogenes is still unknown. In both human and mouse, the number of processed pseudogenes on each chromosome is proportional to chromosome length.50

LOCALISATION OF PSEUDOGENES IN THE GENOME

Today, it is widely accepted that transposable elements and pseudogenes represent two categories of that pool of sequences creating great amounts of ‘rubbish’ (‘junk’) DNA which contributes to the increase of genome size. The position of these sequences in the sequenced genomes seems not to be equally distributed. Eukaryotic pseudogenes tend to occur least frequently near the ‘heart’ of the chromosomes — that is, between centromere and telomere.47,51 In particular, in the worm, they occur markedly near the ends. An exception seems to be D. melanogaster, where processed pseudogenes appear to be dispersed randomly along the chromosome.42

The greatest amount of information about pseudogene localisation is available for the C. elegans genome,31,52 where it has been revealed that, on the chromosomal arms, conserved genes tend to be less numerous, whereas pseudogenes appear to be more abundant, which indicates that genomic DNA could evolve more rapidly towards the ends of the chromosomes. Moreover, the distribution of pseudogenes between the chromosomes is also uneven — eg chromosome IV appears to have more ‘dead’ genes than chromosome II.31 Only a few pseudogenes are processed in the worm genome, which is in marked contrast with what happens in the human genome, in which 80 per cent of the pseudogenes are thought to be processed.53,54

In human and mouse genomes, repeats and pseudogenes are present both in the euchromatin and in the heterochromatin with a general uniformity, but with variations in local density. Processed pseudogene distribution is uneven among regions of different G+C composition. While in the mouse genome the processed pseudogenes have the highest density in the GC-poor regions and are depleted in the GC-rich regions, in the human genome processed pseudogenes are most numerous in the regions of intermediate G+C content.56 In spite of a general uniformity in the distribution of processed pseudogenes in the entire euchromatin of the human genome, there are regions where they tend to accumulate, some of which are close to telomeres.54

In contrast to the human genome, in

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size</th>
<th>No. of ORFs</th>
<th>No. of gene transcripts</th>
<th>No. of processed pseudogenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans</td>
<td>100 Mb</td>
<td>20,598</td>
<td>22,941</td>
<td>104</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>128 Mb</td>
<td>13,525</td>
<td>18,289</td>
<td>34</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>2.6 Gb</td>
<td>26,762</td>
<td>34,076</td>
<td>5,000</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>2.8 Gb</td>
<td>23,531</td>
<td>31,609</td>
<td>8,000</td>
</tr>
</tbody>
</table>
the genome of another vertebrate, the pufferfish *Tetradon nigroviridis*, which is among the smallest known vertebrate genomes, transposable elements and pseudogenes are found almost exclusively in the heterochromatin, in specific regions corresponding to the short arms of subtelocentric chromosomes.

### NATURE AND STRUCTURE OF THE GENES GENERATING RETROPSEUDOGENES

The features of the genes generating pseudogenes have emerged from the analysis of 181 human genes associated with retropseudogenes. The main features are that retrotranscribed genes must be:
1. widely expressed;
2. highly conserved;
3. short; and

These features are not a rule among metazoa. Indeed, in *Drosophila*, coding sequences that give rise to pseudogenes tend to be rather longer than the average coding sequence. It has been found that the closest matching proteins for pseudogenes tend to be about 60 per cent longer than the average *Drosophila* protein. It has been suggested, in the human, that the first two properties (1 and 2) identified for genes are probably connected to the fact that the genes which originate retropseudogenes are expressed in the germinal line. The properties 3 and 4 suggest that reverse transcription and transposition are more efficient in GC-poor and short mRNAs. The genes widely expressed have short coding sequences (CDS) and a lower GC content compared with tissue-specific genes. The reverse transcriptase enzyme involved in the reverse transcription of genes originating as retropseudogenes is probably a 'LINE reverse transcriptase'. Thus, the insertion process of retropseudogenes could be similar to the one of L1 retrotransposition. Much like the genes producing retropseudogenes, the LINE elements of mammals are poor in G+C. Among the genes producing retropseudogenes, it has been reported that the pseudogene number (copy number) increases with the decrease in CDS size. Genes with a single copy and widely expressed in the human genome, such as α-actin, laminine, arginine succinate synthetase and lactate dehydrogenase A genes, have a number of pseudogenes varying from ten to 30.

An example of genes with a high retrotranscription probability according to the features described above is the gene encoding the mtTFA protein, which the authors have studied in their laboratory. mtTFA is a single copy nuclear gene which codes for an activator of mitochondrial transcription in mammals having two HMG boxes and so belong to the HMG protein family. mtTFA enhances mtDNA transcription by mitochondrial RNA polymerase in a promoter-specific fashion and in the presence of mitochondrial transcription factor B. Since mitochondrial replication and transcription seem to be coupled, mtTFA could be essential for mtDNA replication. Recently, it has been reported that human mtTFA is abundant enough to wrap entire mtDNA and that most mtTFA molecules associate with mtDNA. mtTFA mRNA is widely distributed in human tissues, it is highly conserved, short (741 base pairs) and GC poor. The analysis of whole CDS of the h-mtTFA gene revealed a value of 40.01 per cent GC and 33.20 per cent of GC, which corresponds to the features observed in genes located in the L isochore.

According to these data, the authors have demonstrated the presence of mtTFA pseudogenes in rat and human. From Southern blotting hybridisation and *in silico* mapping obtained launching mtTFA CDS as a query in Blast analysis, it was determined that about 20 mtTFA retrotransposed copies exist in human genomic DNA. Three processed pseudogenes have been 'in silico' mapped in chromosomes 7, 11 and X, which are all locations different from those of the normal gene that maps on chromosome 10. Because of the unavailability of the complete rat...
mTFA pseudogenes derive either from the full-length mRNA or from the alternatively spliced mRNA

genome, the number of rat mtTFA retrotransposed copies has been inferred by the in vitro hybridisation technique, which indicates that there are at least 12 copies.

The authors have also searched for the presence of mtTFA retropseudogenes in the genomes of different primates. In Homo sapiens and Presbytis cristata, shorter pseudogenes of 383 and 377 base pairs, respectively, have been revealed, showing a high similarity (98.9 per cent and 81.4 per cent, respectively) with mtTFA cDNAs but lacking exon 5 in H. sapiens and part of exon 4 in P. cristata. Both deletions involve part of the HMG boxes.

In both human

and rat

mtTFA mRNAs populations, the presence of at least one splicing isoform lacking exon 5 is known, which represents almost 30 per cent and 10 per cent, respectively, of all mtTFA transcripts. Hence, it is supposed that the shorter fragment (of 383 base pairs) amplified in the human genome could result from the retrotranscription and integration of this splicing isoform. This could mean that mtTFA pseudogenes derive either from the full-length mRNA or from the alternatively spliced mRNA. A similar hypothesis could be proposed for the shorter fragment obtained in P. cristata, even if the presence of a splicing isoform lacking part of exon 4 has not yet been described. The authors’ studies on mtTFA processed pseudogenes have revealed that these sequences are almost equally distributed in all the analysed mammals (rat, mouse, human and other primates), with their percentages of similarity to mtTFA CDS always higher than 70 per cent and that the number of these sequences in human

and rat

is quite similar, with the processed pseudogenes deriving either from the full-length mRNA or from the alternatively spliced mRNA.

References

Pseudogenes in metazoa: Origin and features


166 D’Errico, Gadaleta and Saccone


