Pseudogenes and Their Evolution

Ondrej Podlaha, *University of California, Davis, California, USA*
Jianzhi Zhang, *University of Michigan, Ann Arbor, Michigan, USA*

Advanced article

Article Contents

Online posting date: 15th November 2010

Pseudogenes are relics of former genes that no longer possess biological functions. They are abundant in the genomes of complex organisms such as vertebrates and flowering plants and provide a useful resource for studying mutation rates and neutral evolutionary patterns. Processed pseudogenes can be regarded as fossilised footprints of past gene expression, permitting a peek into ancient transcriptomes. Occasionally adaptive, pseudogenisation is usually a neutral process. Yet, it can have a substantial impact on future evolution by limiting or opening certain evolutionary possibilities. Studies of pseudogenisation may help date key phenotypic changes in evolution and understand the genetic basis of phenotypic evolution. Recent studies found some pseudogenes to possess apparent functions in gene regulation, creating a difficulty in defining pseudogenes.

Introduction

The word 'pseudogene' was first used by Jacq and colleagues to describe a deoxyribonucleic acid (DNA) sequence that, while resembling a gene coding for the frog 5S ribosomal ribonucleic acid (RNA), contained several degenerative features that rendered its RNA product nonfunctional (Jacq *et al*., 1977). Since then, the term 'pseudogene' has been widely used to encompass all DNA sequences that display two distinct characteristics: (i) the sequence is highly similar to that of a functional gene, yet (ii) the sequence contains degenerative features such that its (RNA or protein) product is nonfunctional. In other words, pseudogenes are severely crippled gene copies that cannot generate a functional product. However, it was not until the last decade when many prokaryotic and eukaryotic genomes were completely sequenced had scientists begun to appreciate the abundance of pseudogenes in living organisms and to understand their origins and their roles in evolution. Despite their seemingly useless existence, pseudogenes taught us much about the process of DNA evolution. Very recently, their role in gene regulation was discovered and their potential utility in the study of transcriptome evolution was explored. **See also**: Human Lineage-specific Gene Inactivation; Pseudogene Evolution in the Human Genome

ELS subject area: Evolution and Diversity of Life

How to cite:
Podlaha, Ondrej; and Zhang, Jianzhi (November 2010) Pseudogenes and Their Evolution. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005118.pub2

Origins of Pseudogenes

Most pseudogenes came from duplicate genes that were generated by either DNA or RNA mediated duplication (**Figure 1**). In DNA-mediated duplication, an extra copy of a gene, often with the full complement of its coding sequence and regulatory noncoding sequence, is generated. This can happen through the mechanism of unequal crossing-over during meiosis, resulting in tandem duplicates linked in a chromosome. It can also happen by nondisjunction in meiosis or polyploidisation, resulting in whole chromosome or even whole genome duplication (WGD). These processes usually generate paralogs that are capable of transcription and translation into functional proteins. Because the parent gene and the newly arisen paralog are functionally redundant, one of them gradually accumulates degenerative mutations that lead to the loss of gene function, that is, one copy becomes a pseudogene (Zhang, 2003). Pseudogenisation, however, does not have to be the end result of gene duplication. In rare instances when mutations confer novel functions or divide ancestral functions, both gene copies can be stably retained, albeit with different functions (Ohno, 1970; Force *et al*., 1999).

In RNA-mediated duplication, the messenger RNA (mRNA) of a gene is reversed transcribed into complimentary DNA, which is then inserted back into the genome at random, a process that is also known as retroposition

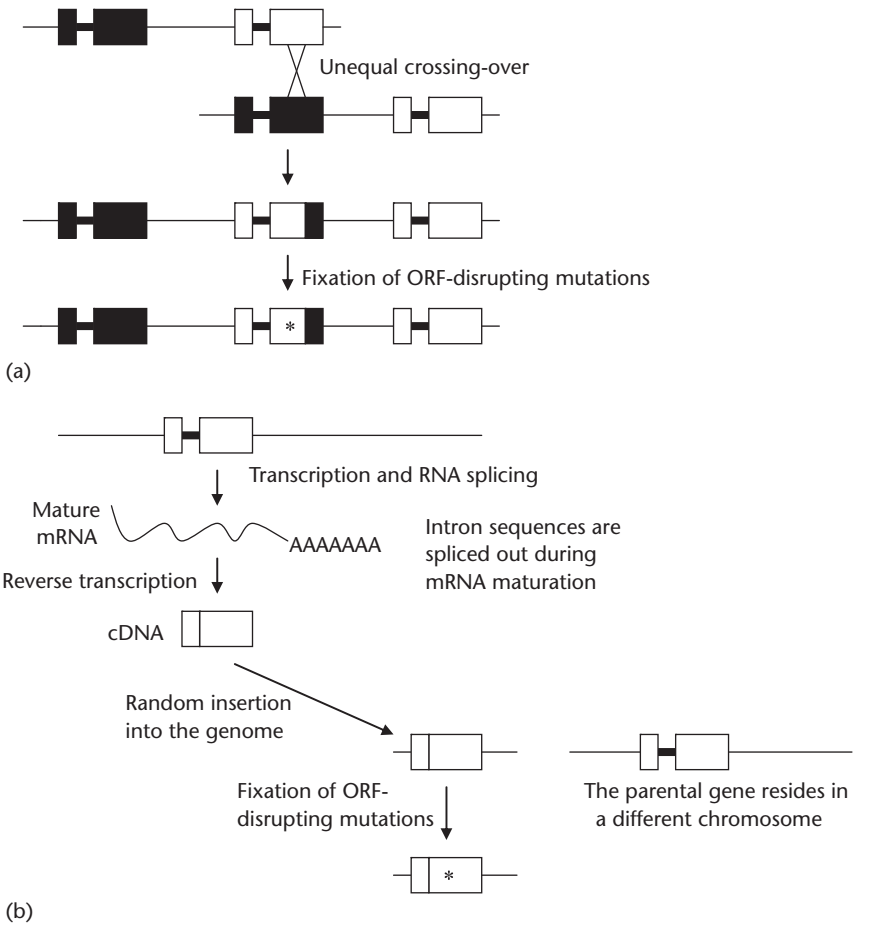

Figure 1 Pseudogene formation after DNA- and RNA-mediated duplication. (a) Unequal crossing-over leads to the generation of an extra gene with a full complement of coding and noncoding sequences. (b) During retroposition, mature mRNA is reverse-transcribed and randomly inserted into the genome, giving rise to a processed pseudogene. Retroposition copies only the coding sequence, leaving out any regulatory sequence. Rectangles and bold lines represent exons and introns, respectively. Asterisks indicate open reading frame (ORF)-disrupting mutations. Reproduced with modification from Zhang (2003), with permission from Elsevier.

(**Figure 1**). Unlike the duplicates from DNA-mediated duplication, those from retroposition do not contain introns or a promoter sequence. Unless the retroposition happens to occur at a locus that already contains a regulatory sequence, the retroposed copy is 'dead on arrival', because it cannot be transcribed (Zhang, 2003). Consequently, the coding sequence gradually degenerates by accumulating random mutations. These nonfunctional copies are called processed pseudogenes for their absence of introns and presence of polyadenylated 3′ end that are both typical of processed mRNAs.

Whereas pseudogene formation is most often preceded by DNA- or RNA-mediated duplication, functional genes can also turn into pseudogenes without duplication, simply when their selective constraints are drastically reduced or removed altogether due to either changes of the genetic background or more often the environment. **See also**: Processed Pseudogenes and Their Functional Resurrection in the Human and Mouse Genomes

Prevalence of Pseudogenes

The number of pseudogenes in a genome depends on the balance between the birth and the loss of pseudogenes. The birth rate of pseudogenes is determined primarily by the rates of DNA- and RNA-mediated duplications. In addition, mutation rate will also affect the pseudogene birth rate, particularly when the ecology of a species changes suddenly such that a large number of formerly useful genes become useless (e.g. when a free-living organism becomes an endosymbiont). The pseudogene loss rate depends on factors such as the neutral substitution and deletion rates. The effective population size can affect both the birth and the death rates, because it affects the likelihood that a duplicate is retained in the genome and the probability that the cost of carrying a piece of useless DNA in the genome (due to the energetic waste of replication or expression) is detected by natural selection.

Table 1 Variation of the number of pseudogenes among species

Species	Genome size (Mb)[a]	Number of annotated genes[a]	Number of pseudogenes[b]
Human	3272	22 286	16 946
Chimpanzee	2929	17 741	16 785
Mouse	3420	22 931	19 044
Rat	2507	21 673	13 962
Dog	2384	15 900	12 852
Chicken	1050	15 908	5539
Zebrafish	1564	23 569	16 357
Pufferfish	342	15 315	3250
Fruitfly	168	14 076	2208
Mosquito	278	12 604	4019
Caenorhabditis elegans	103	20 158	2445
Arabidopsis thaliana	135	24 405	2698
Rice	487	44 659	5608
Escherichia coli (O157 H7)	4.7	4263	224

[a]Data from www.ensembl.org.
[b]Data from www.pseudogene.org.

Pseudogene birth and loss rates vary substantially across species, and therefore, the prevalence of pseudogenes is vastly different in different genomes (**Table 1**). For example, both the human and the mouse genomes contain approximately 22 000 annotated protein coding genes and similar numbers of pseudogenes, approximately 17 000 and approximately 19 000, respectively. The *Drosophila melanogaster* genome has approximately 14 000 protein-coding genes, but only approximately 2200 pseudogenes. This contrast between mammals and *Drosophila* in pseudogene number is likely due to *Drosophila*'s high genomic deletion rate (Petrov *et al*., 1996). Mouse pseudogenes show greater levels of decay than human pseudogenes, which is likely to be a consequence of a higher neutral mutation rate in rodents than in primates (Graur *et al*., 1989; Waterston *et al*., 2002).

Differential mutation rates can complicate pseudogene identification and skew overall abundance estimates. Pseudogene identification is not a trivial task and usually involves the comparison of a DNA sequence with all protein-coding sequences. Frequently, very stringent criteria that require major disruptive mutations are imposed on pseudogene candidates to minimise the false-positive rate. By default, this approach will underestimate the number of very young pseudogenes, because it typically takes several million years for an open reading frame (ORF)-disrupting mutation to occur and become fixed in a gene that lacks any selective constraint (Zhang and Webb, 2003). Likewise, very old pseudogenes will also be underdetected because of the difficulty of establishing homology to any protein-coding gene.

Although the total number of detected pseudogenes varies among vertebrates, the number of processed pseudogenes is usually approximately three to four times that of nonprocessed pseudogenes (www.pseudogene.org). This suggests that the retroposition rate is likely the key determinant in pseudogene formation, at least in mammals.

Recent studies revealed that plant genomes also contain many pseudogenes. *Arabidopsis thaliana* and *Oryza sativa* (rice) contain approximately 2700 and approximately 5600 well-defined pseudogenes, respectively (Benovoy and Drouin, 2006; Zou *et al*., 2009). This disparity is also reflected in the genome size difference between the two plants, as *Arabidopsis* genome is 135 Mb in size and contains approximately 24 000 protein-coding genes (*Arabidopsis* Genome Initiative, 2000), whereas the rice genome is 487 Mb in size and contains approximately 45 000 genes (International Rice Genome Sequencing Project, 2005). The contrast in pseudogene number between these two species may also be related to their different histories of WGD. The relatively recent WGD in *Arabidopsis* (approximately 20–40 Mya), compared to that in the rice (Blanc *et al*., 2003; Yu *et al*., 2005), implies that some selectively unconstrained duplicate genes may not have had sufficient time to acquire premature stop codons and frame-shifting mutations, which leads to an underestimation of the pseudogene number.

In both animals and plants, the distribution of pseudogenes along chromosomes seems to be generally proportional to the chromosome length. On a more microscopic level, however, analysis of processed pseudogenes in human and mouse shows that their density is uneven given the G + C content (Zhang *et al*., 2004). Despite the fact that both organisms share similar LINE (long interspersed element) machinery, the preferential retroposition location in mouse is in regions with low G + C frequencies, whereas in humans, this tendency is not so extreme and most processed pseudogenes are found in regions with intermediate G + C frequencies (Zhang *et al*., 2004). One known vertebrate exception to the relative uniformity of pseudogene distribution along the

chromosome is the pufferfish *Tetraodon nigroviridis*. The genome of *Tetraodon* is among the smallest in vertebrates and shows high levels of compartmentalisation. Pseudogenes, along with transposable elements, are exclusively located in the heterochromatic regions of the short arms of subtelomeric chromosomes in *Tetraodon* (Dasilva *et al*., 2002).

Given the high number of pseudogenes in the mammalian genome, one would be tempted to think that a large fraction of protein-coding genes have their pseudogenes by either DNA- or RNA-mediated duplication present in the genome. Surprisingly, certain genes are prone to generating large numbers of pseudogenes while others produce none. Two major factors influence the likelihood that a functional gene produces a processed pseudogene: the tissue of expression and the level of expression. For any newly arisen pseudogene to be heritable, it has to originate in the germline, or embryonic stem cells that will give rise to the germline. Thus, genes that are expressed only in somatic cells do not have processed pseudogenes. Additionally, the level of expression affects the probability of retroposition (Podlaha and Zhang, 2009). Highly expressed genes, such as house-keeping genes, have a higher probability of retroposition, simply because of the high numbers of mRNA molecules that can be reverse-transcribed and then inserted into the genome. For example, the human ribosomal proteins, which are encoded by roughly 80 genes, have approximately 2000 pseudogenes (Balasubramanian *et al*., 2009). In general, the classes of genes that produce the highest numbers of pseudogenes belong to only a few functional categories: ribosomal proteins, DNA- and RNA-binding proteins, certain structural proteins and metabolic enzymes.

One can study the age distribution of pseudogenes by comparing the sequences of pseudogenes with those of their parental functional genes. It has been observed that this age distribution is skewed such that there appears to have more young pseudogenes than old pseudogenes within each time unit (Zhang *et al*., 2004; Marques *et al*., 2005). This pattern reflects the combined effects of a decay of nonfunctional sequences in the genome (i.e. very old pseuodgenes tend to be lost or hard to detect) and the changes in the activity of the reverse transcriptase in the past. **See also**: Evolution of Nuclear Receptor Pseudogenes; Processed Pseudogenes and Their Functional Resurrection in the Human and Mouse Genomes; Pseudogenes: Age; Pseudogene Evolution in the Human Genome

Utilities of Pseudogenes

In the 1980s and the 1990s, pseudogenes were extremely useful for studying the rates of neutral substitutions, because whole genome sequences were not yet available at the time and pseudogenes are closest to DNA sequences lacking any biological function and selective constraints. From these studies, we learned that C–phosphate–G (CpG) dinucleotides are mutational hotspots and that the relative frequencies of deletions and insertions are not equal (Ophir and Graur, 1997). Many initial comparative studies were performed between the human and the mouse, revealing a considerably higher mutation rate in rodents than in primates (Graur *et al*., 1989). Early studies of pseudogenes also elucidated the mutational mechanisms underlying some human-inherited diseases such as the polycystic kidney disease (Watnick *et al*., 1998) and the Gaucher syndrome (Eyal *et al*., 1990). Interestingly, gene conversion between a pseudogene and its parental gene served as the mutational mechanism causing these human diseases.

The mechanism of RNA-mediated duplication, through which processed pseudogenes are produced, lends these pseudogenes a useful characteristic. As retroposed copies of spliced mRNAs, processed pseudogenes represent an archive of splice variants. Alternative splicing plays a key role in generating the diversity of proteins from a limited number of genes, and distinct splice forms of a single gene's transcript may be translated into proteins with diverse functions (Keren *et al*., 2010). It is often difficult to discern how many splice variants are produced by a single gene, let alone how splicing of a particular transcript evolved over time. In a study by Shemesh *et al*. (2006), the authors compared the sequences of thousands of processed pseudogenes to the exon architecture of their parent genes and discovered that many genes produce previously unknown transcript variants. A large sample of these splice variants were confirmed experimentally (Shemesh *et al*., 2006). Furthermore, this comparison enabled them to identify ancient splice variants that are no longer being produced today, giving us a unique glimpse into the transcriptome's past.

The observation that genes expressed at high levels in the germline generate the majority of processed pseudogenes prompted the hypothesis that the level of retroposition correlates positively with the amount of expression in the germline. Only recently has this correlation been verified (Podlaha and Zhang, 2009) and the implication of this finding is interesting. After establishing the correlation between the expression of the parent gene and the number of processed pseudogene offspring in mouse, Podlaha and Zhang (2009) used the genomic age distribution of processed pseudogenes to infer the historical expression levels for the parent genes. They observed that approximately 3% (at the 1% false discovery rate) of mouse and human genes expressed in the embryonic tissue or germline significantly changed expression levels through evolutionary time. Thus, processed pseudogenes serve as a fossilised record of past gene expression.

Pseudogenisation and Evolution

Pseudogenisation refers to the process by which a functional gene becomes a pseudogene during evolution. Although the majority of pseudogenisation events are

neutral and occur through the random accumulation of mutations in genes on which the functional constraints are relaxed, pseudogenisation can sometimes be adaptive. This occurs when the inactivation of a previously functional gene increases the organismal fitness upon a change in the environment or genetic background. One example of such adaptive pseudogenisation is the human cysteine-aspartic protease 12 (*CASPASE12*) gene (Wang *et al*., 2006). *CASPASE12* encodes a cysteinyl aspartate proteinase involved in the suppression of immune response to endotoxins. Human populations outside of Africa carry a null *CASPASE12* allele due to a point mutation that creates a premature stop codon, whereas 10% of the Africans still harbour the functional allele. Epidemiological studies showed that the null allele is associated with a significantly reduced incidence of severe sepsis (Saleh *et al*., 2004), suggesting that the spread of the null allele may have been promoted by positive selection (Wang *et al*., 2006). Indeed, population genetic analysis revealed strong signatures of a selective sweep of the null allele that started shortly before the out-of-Africa migration of modern humans (Wang *et al*., 2006).

Even when pseudogenisation is not immediately beneficial, it may open up future evolutionary possibilities or channel future evolutionary paths. For example, the loss of the sarcomeric myosin heavy chain 16 gene *MYH16* during human origins was probably not advantageous at the time of pseudogenisation, but this event may have allowed subsequent changes in the cranial morphology that would not have been previously possible (Stedman *et al*., 2004). MYH16 is associated with the masticatory muscle fibre size in primates, and the loss of this gene in the human lineage was argued to have had profound effects on the relative size of jaw muscles and consequently cranial size. The powerful jaw muscles in primates run through massive zygomatic arch opening and attach at the top of the cranium. Modern humans and closest ancestors, however, have much weaker muscles that only extend up the side of the skull and attach in the temple area. Stedman *et al*. (2004) argued that the loss of *MYH16* in hominin evolution resulted in the reduction of the masticatory muscle, allowed for later cranio-facial modifications and possibly even the brain size expansion.

Studying pseudogenisation can also help dating important evolutionary changes of certain phenotypic traits when the gene–trait relationship is clear. For example, humans and related primates lack a functional vomeronasal organ, an olfactory organ used for sensing pheromones and environmental odorants. TRPC2 (transient receptor potential cation channel, subfamily C, member 2) is a channel protein absolutely required for vomeronasal signal transduction; mice lacking TRPC2 show altered sexual and social behaviours (reviewed in Grus and Zhang, 2006). Interestingly, the human *TRPC2* gene contains multiple ORF-disrupting mutations. Zhang and Webb sequenced *TRPC2* from all major lineages of higher primates and found that the gene was inactivated in the common ancestor of catarrhines (humans, apes and Old World monkeys; Zhang and Webb, 2003). Intriguingly, the full trichromatic colour vision also evolved in the catarrhine ancestor, through the duplication of the X-linked red/green opsin gene. It is possible that the vomeronasal organ-mediated chemical communication is replaced by the visual communication if the latter is superior to the former (Zhang and Webb, 2003). **See also**: Human Lineage-specific Gene Inactivation

'Functional Pseudogenes'

In rare cases, pseudogenes that have clear signs of sequence deterioration, including frame-shifting mutations and premature stop codons, show unusual patterns of evolution that are normally associated with functional genes. The duplicated *Adh* (alcohol dehydrogenase) pseudogene in the *repleta* group of *Drosophila* suffered several mutations that render the sequence incapable of coding for a functional protein. However, other molecular characteristics of this *Adh* locus are very atypical of a pseudogene. For example, the rate of nucleotide substitutions is lower in exons than in introns; the codon bias in this pseudogene is still retained and the silent substitution rate is significantly higher than the replacement substitution rate. These features prompted the speculation that the *Adh* pseudogene in the *repleta* group is perhaps a chimeric functional gene (Begun, 1997). Even more astonishing behaviours of atypical pseudogenes can be found in plants. Tiling-array experiments looking for transcribed regions in the genome showed that approximately 20% of annotated pseudogenes in *Arabidopsis* and rice are actively transcribed (Yamada *et al*., 2003; Johnson *et al*., 2005; Zou *et al*., 2009). The 5′ noncoding regions of most pseudogenes in *Arabidopsis* and rice show a higher sequence similarity to those of their functional paralogs than the 3′ regions. In fact, the average nonsynonymous to synonymous substitution rate ratio for nearly half of *Arabidopsis*' pseudogenes is approximately 0.4, which is significantly lower than the expected ratio of approximately 1 for nonfunctional sequences (Zou *et al*., 2009). All these peculiar properties are reminiscent of functional constraints; a closer examination is needed to test whether these apparent pseudogenes actually perform any biological function.

Clear evidence for a 'functional pseudogene' is difficult to come by. Probably the first and most frequently cited example of biological function of an operationally defined pseudogene is the nitric oxide synthase (NOS) pseudogene in the snail *Lymnaea stagnalis*. Owing to a small inversion within the NOS pseudogene, the nonfunctional transcript forms a stable RNA–RNA duplex with the functional NOS mRNA, causing a significant reduction in NOS protein expression (Korneev *et al*., 1999). The second striking example of a functioning pseudogene could be the *makorin1-p1* in mice. The conclusion that *makorin1-p1* has a biological function has been seriously contested, but Hirotsune *et al*.'s (2003) original experiments advocated that the stability of the functional *makorin1* mRNA is

enhanced in the presence of *makorin1-p1* transcript. In the absence of the *makorin1-p1* transcript, the expression of *makorin1* was drastically decreased and 80% of newborn mice suffered mortality (Hirotsune *et al*., 2003). Findings by Gray *et al*. (2006), however, contradict the verdict of *makorin1-p1* functionality. First, Gray *et al*. (2006) showed that the *makorin1-p1* locus is completely methylated, which is symptomatic of transcriptionally silent loci; second, the putative *makorin1-p1* transcript reported by Hirotsune *et al*. (2003) was found to be very similar to the truncated *makorin1* mRNA and finally, the disruption of the functional *makorin1* in mice produced a different phenotype than the one previously attributed by Hirotsune *et al*. (2003) to the partial reduction of *makorin1* through the disruption of *makorin1-p1*. The third example, discovered very recently, involves the human tumour suppressor gene *PTEN* (phosphatase and tensin homolog) and its pseudogenised paralog *PTENP1* (Poliseno *et al*., 2010). It was reported that the mRNA of *PTENP1* regulates cellular levels of PTEN and exerts a growth-suppressive role through competing for binding with microRNAs that suppresses PTEN production. Furthermore, *PTENP1* is selectively lost in human cancer.

Although the precise regulation mechanism is still unknown for NOS and *makorin1*, a clearer picture is emerging from the studies of gene expression regulation through RNA interference (RNAi). RNAi, in very simple terms, involves small RNA molecules (less than 30 nucleotides in length) that are capable of targeting specific mRNAs for degradation. The source of these small interfering RNAs (siRNAs) is a subject of extensive investigation, but several recent studies point to pseudogenes (Tam *et al*., 2008; Watanabe *et al*., 2008). In particular, pseudogenes can generate siRNAs through the formation of a specific double-stranded RNA (dsRNA) complex. This dsRNA complex can form in two ways, either through the interaction of antisense pseudogene transcript and the parent gene mRNA or through the formation of a hairpin structure in the presence of inverted repeats. After the formation of dsRNA complex, the structure is cut into 21-nucleotide endo-siRNAs, which are then guided to interact and degrade the parent's mRNA. Reports by Watanabe *et al*. (2008) and Tam *et al*. (2008) also indicate that the genes regulated by pseudogene-derived siRNAs are disproportionately involved in cytoskeletal dynamics, indicating that this type of regulation is probably not caused by pseudogenes simply by chance, but by selection.

Difficulty with the Pseudogene Definition

The term 'pseudogene' was originally coined to describe a degenerated RNA- or protein-coding sequence that is incapable of being transcribed or translated into functional RNA or protein products. The key in this definition is that pseudogenes are biologically nonfunctional. However, in practice, it is virtually impossible to experimentally establish nonfunctionality; the lack of any observable phenotypic effect upon the deletion of a putative pseudogene does not necessarily mean that the deletion has no phenotypic effect, because the effect may be too subtle to observe. When more and more research groups are coming across cases where a so-called pseudogene is potentially involved in a meaningful biological interaction, primarily in gene regulation (Tam *et al*., 2008; Watanabe *et al*., 2008), it becomes increasingly difficult to define pseudogenes. Is it appropriate to call such noncoding yet functional sequences pseudogenes? This is where the boundaries of the pseudogene definition become hazy. As mentioned, experimental demonstration of nonfunctionality of a putative pseudogene would be impractical and inconclusive. Can transcriptional activity of a particular locus serve as a sufficient criterion of functionality? Hardly. Recent transcriptome analyses from various model organisms showed that a large portion of noncoding sequences, including annotated pseudogenes, produce transcripts (Yamada *et al*., 2003; Johnson *et al*., 2005; Zou *et al*., 2009). Furthermore, although most examples of 'functional pseudogenes' involve some interactions between the pseudogene transcript and the mRNA of the functional paralog, pseudogenes have also been found to function in ways that do not require transcription, such as serving as the material source of gene conversion events to enhance the antibody diversity in birds, humans and other vertebrates (Ota and Nei, 1995; Balakirev and Ayala, 2003).

To eliminate controversies involved in the pseudogene nomenclature, Zheng and Gerstein (2007) proposed a novel term 'exapted pseudogene'. This name indicates the recruitment of a pseudogene for a new function. Although this terminology does not clearly differentiate among the myriad of mechanisms through which pseudogenes can biologically function, it may be a more appropriate name than the 'functional pseudogene' oxymoron.

References

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Balakirev ES and Ayala FJ (2003) Pseudogenes: are they "junk" or functional DNA? *Annual Review of Genetics* **37**: 123–151.

Balasubramanian S, Zheng D, Liu YJ *et al*. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biology* **10**: R2.

Begun DJ (1997) Origin and evolution of a new gene descended from alcohol dehydrogenase in *Drosophila*. *Genetics* **145**: 375–382.

Benovoy D and Drouin G (2006) Processed pseudogenes, processed genes, and spontaneous mutations in the *Arabidopsis* genome. *Journal of Molecular Evolution* **62**: 511–522.

Blanc G, Hokamp K and Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Research* **13**: 137–144.

Dasilva C, Hadji H, Ozouf-Costaz C *et al*. (2002) Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. *Proceedings of the National Academy of Sciences of the USA* **99**: 13636–13641.

Eyal N, Wilder S and Horowitz M (1990) Prevalent and rare mutations among Gaucher patients. *Gene* **96**: 277–283.

Force A, Lynch M, Pickett FB *et al*. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Graur D, Shuali Y and Li WH (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. *Journal of Molecular Evolution* **28**: 279–285.

Gray TA, Wilson A, Fortin PJ and Nicholls RD (2006) The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. *Proceedings of the National Academy of Sciences of the USA* **103**: 12039–12044.

Grus WE and Zhang J (2006) Origin and evolution of the vertebrate vomeronasal system viewed through system-specific genes. *BioEssays* **28**: 709–718.

Hirotsune S, Yoshida N, Chen A *et al*. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**: 91–96.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**: 793–800.

Jacq C, Miller J and Brownlee G (1977) Pseudogene structure in 5S-DNA of Xenopus laevis. *Cell* **12**: 109–120.

Johnson JM, Edwards S, Shoemaker D and Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics* **21**: 93–102.

Keren H, Lev-Maor G and Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews. Genetics* **11**: 345–355.

Korneev SA, Park JH and O'Shea M (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *Journal of Neuroscience* **19**: 7711–7720.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A and Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biology* **3**: e357.

Ohno S (1970) *Evolution by Gene Duplication*. Berlin: Springer.

Ophir R and Graur D (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205**: 191–202.

Ota T and Nei M (1995) Evolution of immunoglobulin VH pseudogenes in chickens. *Molecular Biology and Evolution* **12**: 94–102.

Petrov DA, Lozovskaya ER and Hartl DL (1996) High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**: 346–349.

Podlaha O and Zhang J (2009) Processed pseudogenes: the 'fossilized footprints' of past gene expression. *Trends in Genetics* **25**: 429–434.

Poliseno L, Salmena L, Zhang J *et al*. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**: 1033–1038.

Saleh M, Vaillancourt JP, Graham RK *et al*. (2004) Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**: 75–79.

Shemesh R, Novik A, Edelheit S and Sorek R (2006) Genomic fossils as a snapshot of the human transcriptome. *Proceedings of the National Academy of Sciences of the USA* **103**: 1364–1369.

Stedman HH, Kozyak BW, Nelson A *et al*. (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* **428**: 415–418.

Tam O, Aravin A, Stein P *et al*. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**: 534–538.

Wang X, Grus WE and Zhang J (2006) Gene losses during human origins. *PLoS Biology* **4**: e52.

Watanabe T, Totoki Y, Toyoda A *et al*. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**: 539.

Waterston RH, Lindblad-Toh K, Birney E *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Watnick TJ, Gandolph MA, Weber H, Neumann HPH and Germino GG (1998) Gene conversion is a likely cause of mutation in PKD1. *Human Molecular Genetics* **7**: 1239–1243.

Yamada K, Lim J, Dale JM *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**: 842–846.

Yu J, Wang J, Lin W *et al*. (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biology* **3**: e38.

Zhang J (2003) Evolution by gene duplication – an update. *Trends in Ecology & Evolution* **18**: 292–298.

Zhang J and Webb DM (2003) Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proceedings of the National Academy of Sciences of the USA* **100**: 8337–8341.

Zhang Z, Carriero N and Gerstein M (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in Genetics* **20**: 62–67.

Zheng D and Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends in Genetics* **23**: 219–224.

Zou C, Lehti-Shiu MD, Thibaud-Nissen F *et al*. (2009) Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiology* **151**: 3–15.

Furthur Reading

Chiefari E, Iiritano S, Paonessa F *et al*. (2010) Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. *Nature Communications* **1**: 40.

Karro JE, Yan Y, Zheng D *et al*. (2007) Pseudogene. org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Research* **35**: D55–D60.

Lam HY, Khurana E, Fang G *et al*. (2009) Pseudofam: the pseudogene families database. *Nucleic Acids Research* **37**: D738–D743.

Li WH (1997) *Molecular Evolution*. Sunderland, MA: Sinauer.

Li WH, Gojobori T and Nei M (1981) Pseudogenes as a paradigm of neutral evolution. *Nature* **292**: 237–239.

Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics* **64**: 18–23.

Svensson O, Arvestad L and Lagergren J (2006) Genome-wide survey for biologically functional pseudogenes. *PLoS Computational Biology* **2**: e46.

Zhang J and Zhang YP (2003) Pseudogenization of the tumor-growth promoter angiogenin in a leaf-eating monkey. *Gene* **308**: 95–101.

Zhao H, Yang JR, Xu H and Zhang J (2010) Pseudogenization of the umami taste receptor gene Tas1r1 in the giant panda coincided with its dietary switch to bamboo. *Molecular Biology and Evolution* [Epub ahead of print].