Genome Analysis

Functional divergence of proteins through frameshift mutations

Jeroen Raes* **and Yves Van de Peer**

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Frameshift mutations are generally considered to be deleterious and of little importance for the evolution of novel gene functions. However, by screening an exhaustive set of vertebrate gene families, we found that, when a second transcript encoding the original gene product compensates for this mutation, frameshift mutations can be retained for millions of years and enable new gene functions to be acquired.

Introduction

Most proteins are encoded by genes that are part of gene families, many of them consisting of tens or hundreds of genes. The evolutionary and functional diversification of such genes and their proteins is of increasing interest, and many mechanisms to explain the evolution of novel gene functions (e.g. point mutations [1,2] and exon shuffling [3]) have been described. More radical mutational events, such as insertions and deletions that change the reading frame – frameshift mutations – are generally considered to be detrimental (e.g. by causing nonfunctional transcripts and/or proteins [4], through premature stop codons) and of little evolutionary importance, because they seriously alter the sequence and structure of the protein. However, when a protein is temporarily freed from selective pressure, frameshift mutations do not necessarily have to be deleterious. When selective pressure is relieved, for example, through the presence of a second copy of a gene, this duplicate can compensate for the possible loss of function caused by the frameshift mutation in the first gene and enable such mutations to lead to functional divergence. Alternatively, if the frameshift originates from an alternative splicing event, the original splice variant still produces the original gene product. Not surprisingly, gene duplication and alternative splicing are generally thought to be responsible for the majority of protein diversity and the evolution of organismal complexity [5–8]. Figure 1 illustrates a model in which frameshift mutations might survive as a result of the existence of a 'compensating' transcript that performs the function of the original gene.

If such functionally important frameshift mutations have been retained during evolution, we should be able to detect them by searching for a conserved sequence in one subgroup of a gene family that is – if its frame is (artificially) shifted – homologous to a conserved sequence in another subgroup of the same family. By screening an exhaustive set of vertebrate gene families (Figure 2), we found 16 examples (in 15 families) in which a frameshift mutation was retained for up to hundreds of millions of years (Table 1; supplementary material online). The majority of detected frameshifts originated through the emergence of an alternative splice form, in which a new splice site or the use of a different exon leads to a shift in frame (Figure 1). A frameshift mutation in one of the gene copies after duplication occurred in only one out of 16 cases. In addition, one example of conserved frameshift after speciation was found (Table 1 and supplementary data online). Apparently, alternative splicing seems to constitute the principal mechanism leading to conserved frameshifts. Alternatively, the reason for this bias might be that these examples are easier to detect: the two proteins (normal and frameshifted) are – in the case of alternative splicing – encoded by the same locus, and the frameshifted sequence will be more restrained in its evolution because mutations in one frame are also under selection pressure in the other frame. Therefore, these examples might be more easily recognizable than those that occurred after duplication and were preferentially identified by our approach. However, the survival of frameshift mutations after duplication occurs. For examples, in plants, where alternative splicing is believed to be less prevalent [9], three examples (albeit in different clades of the same large MADS-box gene family) of functional divergence through frameshift after duplication have been observed [10].

To gain further insight into the functional relevance of these frameshift mutations, the function of the detected genes was investigated in greater detail. In several examples, the frameshift seems to have been responsible for functional divergence. The natural killer (NK) complex genes constitute one example: NKG2 proteins are transmembrane receptors of the surface of NK cells that recognize the HLA-E antigen of the major histocompatibility complex when dimerized with CD94. They are involved in the specific recognition of self versus non-self antigens by the NK cells [11]. In a rodent-specific branch of the NKG2 family, a frameshift occurred after gene duplication, leading to the production of two gene subfamilies: (i) NKG2C and NKG2E; and (ii) NKG2A and NKG2B. Both gene families have different

Corresponding author: Van de Peer, Y. (Yves.VandePeer@psb.ugent.be).
* *Current address:* European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
Available online 13 June 2005

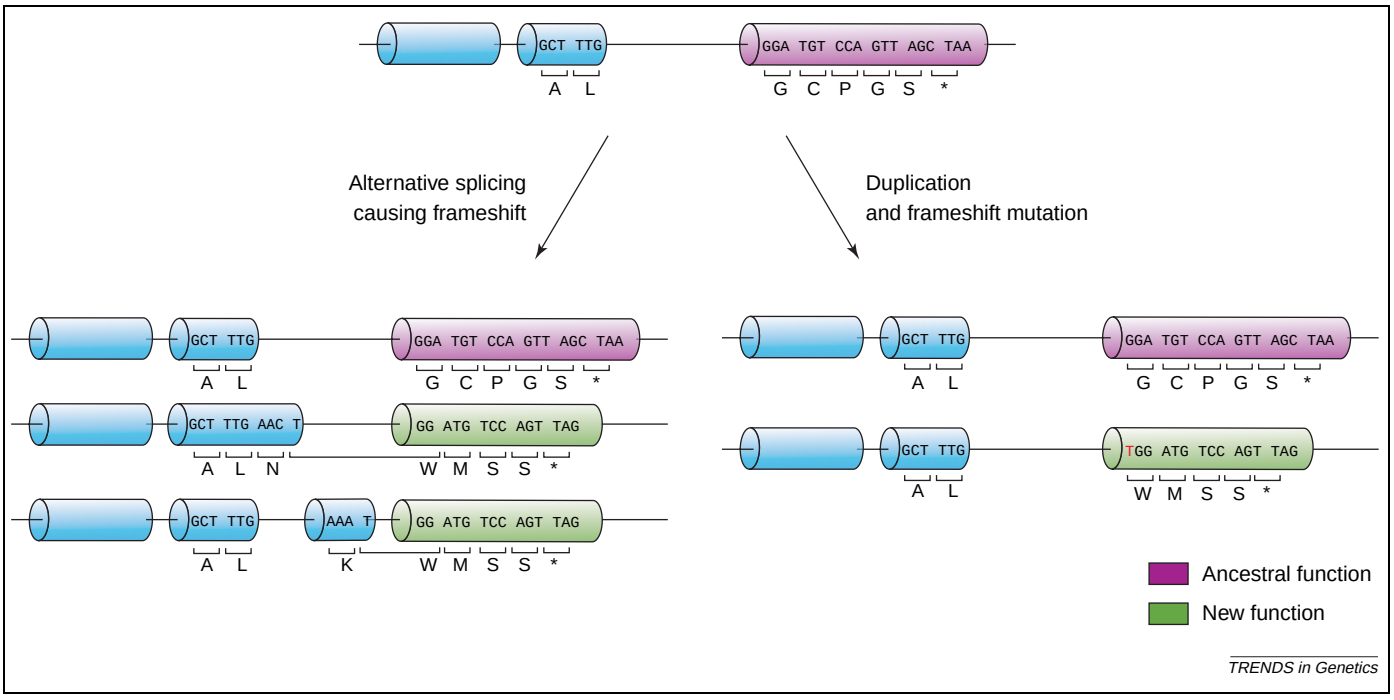

Figure 1. Evolving novel protein functions by frameshift mutations. After gene duplication, a frameshift mutation can escape 'the relentless pressure of natural selection' through the existence of a second copy [4]. In alternative splicing, an additional transcript containing a frameshift can arise, for example, through the use of a new splice site changing the phase of the intron, which alters the reading frame for the remainder of the transcript. However, alternative splicing can lead to the insertion of an additional, frameshifting exon.

functions: NKG2C-containing complexes activate NK cells on binding of the antigen, whereas NKG2-A inhibits them. The functional difference can be clearly linked to the frameshift mutation because the cytoplasmic N-terminal tail of the NKG2A receptor contains an immunoreceptor tyrosine-based inhibitory motif (ITIM) that is crucial for the transmission of the inhibitory signal. This motif is altered through the frameshift in the NKG2C receptor, producing the activating functionality of this protein [12].

Similarly, the C-terminal frameshift in the paired box gene 8 (PAX8) family of transcription factors leads to isoforms with reduced *in vitro* transactivation efficiency, which, in thyroid tissues, results from the loss of interaction with the synergistically acting thyroid transcription factor 1 (TTF-1) [13,14].

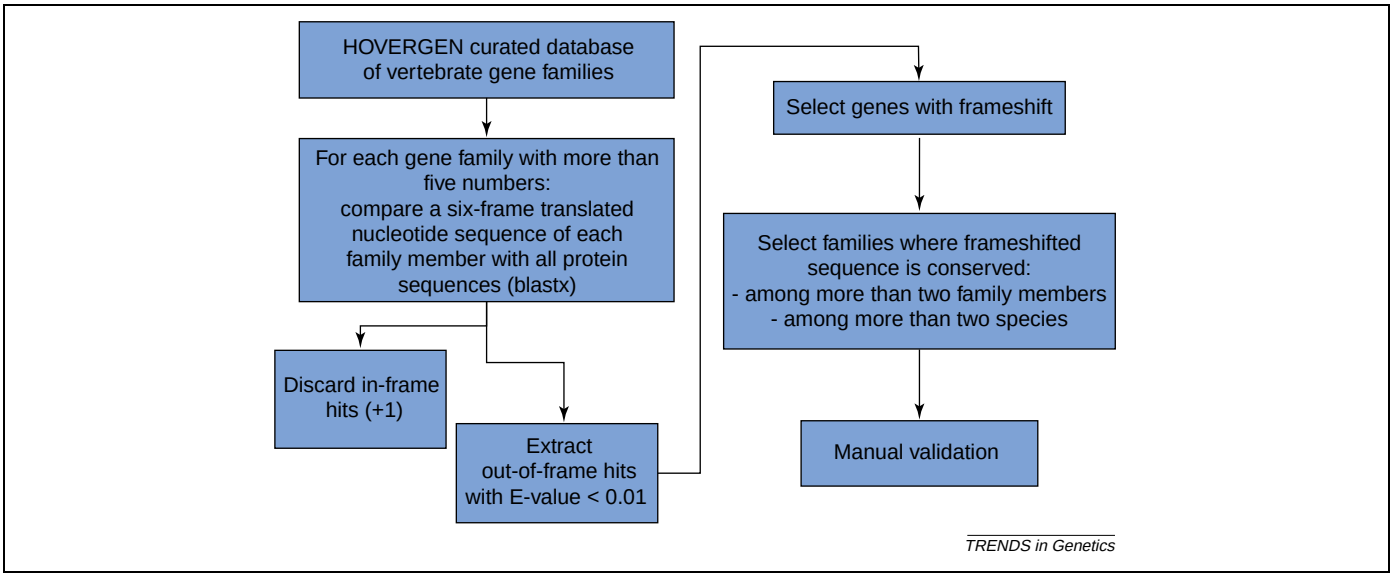

Figure 2. A flowchart of the approach used to screen vertebrate gene families. Gene families with more than five members were extracted from the HOVERGEN (release 43) curated gene family database (http://pbil.univ-lyon1.fr/databases/hovergen.html) [26]. For each gene family, the coding (nucleotide) sequence of each family member was translated in all six reading frames and compared with the protein sequence of all other members using BLASTX at default-parameter settings [27]. From all significant within-family, out-of-frame hits, the genes that contained the frameshift mutations were inferred. For this set of genes, the frameshifted regions were extracted and grouped based on their sequence similarity. Clusters of similar sequences were retained if they contained at least two family members from different species, to exclude the possibility of sequencing errors. Gene families yielding retained clusters (i.e. gene families containing within-family frameshifted sequences conserved in multiple species) were finally subjected to manual verification. Furthermore, expression of the transcripts was confirmed by the presence of a full-length cDNA in the databases and/or from literature (Table 1). Further details of the methodology can be found in the supplementary data online.

Table 1. Gene families containing conserved frameshift mutations[a]

Hovergen family[b]	Function	Mechanism	Position	Downstream effect	Predicted NMD candidate?	Expressed?	Divergence
Transcription factors							
HBG000122	Pbx homeodomain protein	Alternative splicing	C-terminal	Premature stop	No	Yes	Fish–human (450 Mya)
HBG000419	T-cell transcription factor	Alternative splicing	C-terminal	Premature stop	No	Yes	Frog–human (350 Mya)
HBG006561	Nuclear factor 1	Alternative splicing	C-terminal	Premature stop	No	Yes	Frog–human (350 Mya)
HBG009115	Paired box (Pax) transcription factor	Alternative splicing	C-terminal	Premature stop	No	Yes	Dog–human (95 Mya)
HBG017385	Odd-skipped related transcription factor	Alternative splicing	C-terminal	Longer protein	N.A.	Unknown	Rodents–human (80 Mya)
HBG020791	RBCK	Alternative splicing	C-terminal	Premature stop	Yes	Yes	Rodents–human (80 Mya)
Transmembrane proteins							
HBG001463	Equilbrative nucleoside transporter	Alternative splicing	C-terminal	Premature stop	No	Yes	Rodents–human (80 Mya)
HBG004374	Rhesus blood group protein	Alternative splicing	C-terminal	Premature stop	No	Yes	Within primates (10 Mya)
HBG007562	CIRL or latrophilin G-protein-coupled receptor	Alternative splicing	C-terminal	Premature stop	No	Yes	Cow–human (95 Mya)
HBG012748	NKG2 natural killer-cell receptor	Duplication	N-terminal	Corrected	N.A.	Yes	Within rodents (40 Mya)
		Speciation	C-terminal	Premature stop	No	Yes	Within primates (10 Mya)
HBG016641	Granuphilin	Alternative splicing	C-terminal	Premature stop	No	Yes	Within rodents (40 Mya)
Other							
HBG004532	G-protein-coupled receptor kinase	Alternative splicing	C-terminal	Premature stop	No	Yes	Rodents–human (80 Mya)
HBG014652	p58 protein kinase	Alternative splicing	N-terminal	Corrected	N.A.	Yes	Within rodents (40 Mya)
HBG014779	Testis-specific protein Y	Alternative splicing	C-terminal	Premature stop	No	Yes	Within monkeys (20 Mya)
HBG015298	Epididymis-specific EP2 protein	Alternative splicing	C-terminal	Premature stop	No	Yes	Within primates (10 Mya)

[a]Abbreviations: CIRL, Ca^{2+}-independent receptor for α-latrotoxin; Mya, million years ago; N.A., not applicable.
[b]For more details, see http://pbil.univ-lyon1.fr/databases/hovergen.html [26].

By compiling the functions of all detected mutations, we noticed that approximately two-thirds of the gene families with conserved frameshifts consisted of either transcription factors or transmembrane proteins (Table 1). One possible explanation for the apparent preferential retention of frameshift mutations in these classes lies in the structural organization of functional domains in these proteins. In 14 of the 16 cases, frameshift mutations are positioned in the C-terminus. This is not unexpected, because frameshift mutations are generally less harmful if they are near the end of the sequence. However, for both transcription factors and transmembrane proteins, mutations in the C-terminus especially influence those regions that are sensitive to functional diversification. In transcription factors, C-termini are regularly involved in transactivation, repression or in protein–protein interactions [15], whereas in transmembrane proteins, they usually constitute the cytoplasmic or extracellular regions of the protein and have, for example, ligand-binding- or signal-propagation functions.

In addition to frameshifts that are compensated by a second transcript (Figure 1), it should be noted that we have also found one example of a frameshift mutation that occurs after speciation. A recent frameshift within the NK-cell receptors that is primate-specific could be important for recent adaptations of the primate (and thus human) immune system. Therefore, in these cases, the frameshift has been retained without the existence of a compensating transcript, although it remains possible that the original function is exerted by a more distant member of the family or through an alternative route in the pathway [16].

One might argue that the frameshift mutations described here do not cause the acquisition of a new function, but simply infer a partial loss of function in one of the transcripts by 'erasing' a downstream protein domain [17]. For example, if a C-terminally truncated transcription factor contained a functioning DNA-binding domain, but lost its transactivation domain, it could act as a repressor solely by using the former domain and competing with the latter, functional copy [18]. The frameshifted sequence would then not be of crucial functional importance, contrary to our hypothesis. However, the alternative frame is still conserved during evolution, indicating that it is under considerable purifying selection. The same locus is also used in the original frame, and any synonymous (e.g. third codon) mutation in the original frame will, in most cases, result in a nonsynonymous mutation in the shifted frame. Therefore,

given that synonymous mutations in the original frame are under low selectional pressure, the conservation of the frameshifted sequence, despite the 'easy' accumulation of mutations in the original frame, shows that this sequence is selected for and thus of functional importance. We would argue that the conserved frameshifted sequences described here might prove useful in understanding the function of these genes.

Because 13 of the 16 examples reported here result in premature stop codons, one might hypothesize that they are probable targets of nonsense-mediated decay (NMD), a process that limits the synthesis of abnormal proteins by degrading truncated transcripts [19]. However, based on the generally accepted '50 nucleotide rule' (i.e. transcripts with premature stop codons >50 nucleotides upstream of the final exon are degraded) [20,21], only one case – within the RBCK family – is predicted to be a target of NMD (Table 1). RBCK2 is expressed [22], so it seems that even this probable NMD-target escapes degradation. Indeed, NMD was shown to display differences in mRNA-degradation efficiency among tissues and even among individuals; it does not completely breakdown all premature stop-codon-containing transcripts, leading to truncated, but functionally important, proteins [19,23]. Therefore, it seems that NMD is an additional factor in the evolutionary conservation of frameshift mutations: if the resulting transcript is not subject to NMD or can somehow escape it, then the frameshifted sequence can persist and change its function.

Concluding remarks

We propose that frameshift mutations can cause functional divergence of proteins and that, at least in vertebrates, this is usually linked to the presence of NMD avoiding alternatively spliced transcripts. However, because conserved alternatively spliced exons are rarely frame-disturbing [24,25], the conservation of the downstream frameshifted sequence and its subsequent use for neofunctionalization is even less obvious. As Susumu Ohno put it: 'Starting with a frame-shift mutation, a duplicate might acquire a new function, which is totally different from that assigned to the original gene. Admittedly, this is a one in a million chance, but in evolution, events with the odds of one in one million occurred time and time again.' [4] The examples described in this article, mostly transcription factors and transmembrane proteins, have beaten these odds.

Acknowledgements

We thank Cedric Simillion for his help in algorithmic issues, Stephane Rombauts, Evi Michels and Michiel Vandenbussche for stimulating discussions, and Martine De Cock for help in preparing the article. Furthermore, we thank four anonymous reviewers for their constructive remarks. J.R. is a postdoctoral fellow of the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.05.013

References

1 Winkler, K. *et al*. (2000) Changing the antigen binding specificity by single point mutations of an anti-p24 (HIV-1) antibody. *J. Immunol.* 165, 4505–4514
2 Zhang, J. *et al*. (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* 30, 411–415
3 Long, M. *et al*. (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875
4 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag
5 Van de Peer, Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* 5, 752–763
6 International Human Genome Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
7 Taylor, J. and Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* 38, 615–643
8 Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418, 236–243
9 Brett, D. *et al*. (2002) Alternative splicing and genome complexity. *Nat. Genet.* 30, 29–30
10 Vandenbussche, M. *et al*. (2003) Structural diversification and neofunctionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Res.* 31, 4401–4409
11 Vance, R.E. *et al*. (1998) Mouse CD94/NKG2A is a natural killer cell receptor for the nonclassical major histocompatibility complex (MHC) class I molecule Qa-1(b). *J. Exp. Med.* 188, 1841–1848
12 Lohwasser, S. *et al*. (1999) Cloning of murine NKG2A, B and C: second family of C-type lectin receptors on murine NK cells. *Eur. J. Immunol.* 29, 755–761
13 Di Palma, T. *et al*. (2003) The paired domain-containing factor Pax8 and the homeodomain-containing factor TTF-1 directly interact and synergistically activate transcription. *J. Biol. Chem.* 278, 3395–3402
14 Poleev, A. *et al*. (1995) Distinct functional properties of three human paired-box-protein, PAX8, isoforms generated by alternative splicing in thyroid, kidney and Wilms' tumors. *Eur. J. Biochem.* 228, 899–911
15 Latchman, D.S. (1998) *Eukaryotic transcription factors*, Academic Press
16 Wagner, A. (2000) Robustness against mutations in genetic networks of yeast. *Nat. Genet.* 24, 355–361
17 Liu, S. and Altman, R.B. (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res.* 31, 4828–4835
18 Foulkes, N.S. and Sassone-Corsi, P. (1992) More is better: activators and repressors from the same gene. *Cell* 68, 411–414
19 Holbrook, J.A. *et al*. (2004) Nonsense-mediated decay approaches the clinic. *Nat. Genet.* 36, 801–808
20 Hillman, R.T. *et al*. (2004) An unappreciated role for RNA surveillance. *Genome Biol.* 5, R8
21 Nagy, E. and Maquat, L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* 23, 198–199
22 Tokunaga, C. *et al*. (1998) Molecular cloning and characterization of RBCK2, a splicing variant of a RBCC family protein, RBCK1. *FEBS Lett.* 435, 11–15
23 Neu-Yilik, G. *et al*. (2004) Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. *Genome Biol.* 5, 218
24 Sorek, R. *et al*. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.* 20, 68–71
25 Resch, A. *et al*. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* 32, 1261–1269
26 Duret, L. *et al*. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22, 2360–2365
27 Altschul, S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

doi:10.1016/j.tig.2005.05.013