6

Partial Gene Duplication and the Formation of Novel Genes

Macarena Toll-Riera[1], Steve Laurie[1], Núria Radó-Trilla[1] and M.Mar Albà[1,2]
[1]*Evolutionary Genomics Group, Biomedical Informatics Programme, Universitat Pompeu Fabra (UPF) - Institut Municipal d'Investigació Mèdica (IMIM)*
[2]*Catalan Institution for Research and Advanced Studies (ICREA), Barcelona Spain*

1. Introduction

The publication of the first fully sequenced genomes represented a landmark in the biological sciences. The comparison of genomes from different organisms provides us with unprecedented opportunities to address many long-standing evolutionary questions in a more comprehensive way.

1.1 Lineage-specific genes

The availability of several genomes from related organisms permits the identification of newly evolved genes in different lineages or species, the study of their mechanisms of formation and the investigation of their role in adapting to new environments or physiological conditions (Domazet-Loso & Tautz, 2003; Guo et al., 2007; Khalturin et al., 2009; Kuo & Kissinger, 2008; Siepel, 2009; Toll-Riera et al., 2009a; Zhou et al., 2008). Recently formed genes give us the opportunity to study the action of natural selection in recent times and to investigate the processes associated with gene creation (Zhou & Wang, 2008).

The number of species-specific genes, or orphan genes, is not insignificant. They represent around 14% of the genes in 60 fully sequenced microbial genomes (Siew & Fischer, 2003) and between 20-30% in *Drosophila* species (Domazet-Loso & Tautz, 2003; Drosophila 12 Genomes Consortium, 2007). Genes restricted to particular lineages include vomeronasal receptors and casein milk proteins in mammals, which are known to be involved in specific physiological adaptations in this lineage (International Chicken Genome Sequencing Consortium, 2004). Additionally, several lineage-specific genes have been found to be involved in defence against pathogens, such as dermcidin in primates (Toll-Riera et al., 2009a) and surface antigens in apicomplexan parasites (Kuo & Kissinger, 2008). Interestingly, it has been noticed that rice orphan genes are more often expressed under environmental pressure (injury and hormone treatment) than non-orphan genes, indicating that novel genes help in adaptation to changing conditions (Guo et al., 2007).

Many newly evolved genes are derived from partial or complete gene duplication of pre-existing genes (Long et al., 2003; Marques et al., 2005; Toll-Riera et al., 2009a; Zhou et al., 2008). Alternative processes of gene formation include exaptation from mobile elements

(Nekrutenko & Li, 2001; Toll-Riera et al., 2009b), gene fusion or fission (Parra et al., 2006) and *de novo* gene formation from non-coding sequences (Cai et al., 2008; Heinen et al., 2009; Knowles & McLysaght, 2009; Levine et al., 2006; Toll-Riera et al., 2009a). A genome-wide study in *Drosophila melanogaster* has reported that gene duplication is the most common mechanism for the formation of novel genes in this species (Zhou et al., 2008).

1.2 Gene duplication

In the early thirties Haldane (Haldane, 1932) and Muller (Muller, 1935) were the first to propose gene duplication as a mechanism for the generation of new genes. Later, in the seventies, Ohno published an influential book about the role of gene duplication in evolution (Ohno, 1970), in which he emphasised the importance of gene duplication in generating protein functional diversity. With the availability of complete genome sequences it has become possible to estimate genomic rates of gene duplication (Lynch & Conery, 2000), analyse the pattern of evolution of the two duplicated copies, and identify lineage-specific gene family expansions. Expanded gene families that have been analysed in detail include olfactory receptors in mouse (Mouse Genome Sequencing Consortium 2002) and human (Gilad et al., 2005), and KRAB-associated zinc-finger in primates (Castresana et al., 2004). Genomic studies have shown that gene duplication is associated with increased coding sequence evolutionary rates (Lynch and Conery 2000; Scannell and Wolfe 2008), higher tissue expression divergence (Gu et al., 2002; Makova & Li, 2003), and higher regulatory sequence divergence (Farre & Alba, 2010).

The molecular mechanisms that have been proposed to be involved in duplication are non-allelic homologous recombination, transposon-mediated transposition and illegitimate recombination. The first two mechanisms imply the presence of sequence homology (Zhou et al., 2008). Yang and colleagues (Yang et al., 2008) found an excess of repetitive sequences at the breakpoints of the duplicated regions of a group of *Drosophila* lineage-specific young duplicates, suggesting the action of non-allelic homologous recombination. Another study in *Drosophila* found that dispersed duplicates have mainly arisen through non-allelic homologous recombination, while tandem duplicates most often arose through illegitimate recombination (Zhou et al., 2008). It has also been hypothesized that segmental duplications may arise from the recombination of Alu repeat sequences (Bailey et al., 2003).

Duplicated genes appear at a very high rate. It has been estimated that, on average, 0.01 duplicates arise per gene per million years (Lynch & Conery, 2000). The most frequent fate following gene duplication is believed to be the silencing of one of the duplicated copies due to the accumulation of degenerative mutations, a process that may take approximately 4 million years to complete (Lynch & Conery, 2000). However, sometimes both copies survive. The duplicated copy can acquire beneficial mutations and consequently gain a novel function with respect to the parental gene (neofunctionalisation), while the parental copy preserves its original function (Ohno, 1970). The duplicated copy may also be retained due to the split of the original function between the two gene copies (subfunctionalisation) (Hughes, 1994). Finally, if an increase in dosage of a particular gene is beneficial, the new copy may become fixed by positive selection maintaining the same gene structure and function as the parental gene (Kondrashov & Koonin, 2004).

Duplicated genes may confer adaptive advantages. For example, trichromatic colour vision in Old World Monkeys is associated with a pigment gene duplication that occurred after the

separation of New World Monkeys, and which gave rise to differentiated red and green pigments (Nathans et al., 1986). Zhang and colleagues (Zhang et al., 1998) reported on another example of the action of positive selection after gene duplication. The eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) genes are present in Old World Monkeys and hominoids, and probably originated by tandem gene duplication after the divergence of New World Monkeys. EDN is an antiviral agent (Domachowske & Rosenberg, 1997) and ECP is a potent toxin for bacteria and parasites (Rosenberg & Dyer, 1995). The authors detected a non-random accumulation of arginine substitutions in ECP, which may contribute to the generation of pores in pathogens' membranes. Another example refers to pancreatic ribonuclease 1B (RNASE1B), which originated through gene duplication of RNASE1, an enzyme used to digest bacteria in the small intestine, in the douc langur (*Pygathrix nemaeus*) around 2-4 million years ago (Zhang et al., 2002). Douc langurs are folivorous monkeys, in which leaves are digested through fermentation by symbiotic bacteria residing in the foregut. The newly duplicated copy, RNASE1B has evolved very rapidly (non-synonymous to synonymous nucleotide substitution rate of 4.03), contrary to the paralogous copy, RNASE1, which has not undergone change. These results indicate a burst of positive selection acting on the duplicated copy. Moreover, most of the substitutions imply the gain of negatively charged residues, lowering the optimal pH for RNASE1B, which could be related to an increase in digestive efficiency, given the lower pH found in the small intestine of douc langurs.

1.3 Partial gene duplication

Not all duplicated proteins are identical to their parental copies at birth. In fact, it has been reported that in *C. elegans* only about 40% of the new duplicates are borne out of complete gene duplications, the remainder representing cases of partial gene duplication (Katju & Lynch, 2003). These partially duplicated genes may recruit sequences from their genomic neighbourhood or from other genes (Katju & Lynch, 2006). In the first case, adjacent non-coding sequences are co-opted for a coding function. Katju and Lynch (Katju & Lynch, 2006) found that about half of the partially duplicated genes did not recruit any surrounding sequences but accumulated mutations, for example in initiation or termination codons, that altered the coding sequence. In *Drosophila melanogaster*, around 30% of the newly formed genes recruited various genomic sequences or formed chimeric gene structures (Zhou et al., 2008). Partially duplicated and chimeric genes are expected to adopt new functions immediately, which may increase their probability of being retained (Patthy, 1999; Zhou et al., 2008). An example of a gene that has arisen by partial duplication is the *Hun* gene in *Drosophila*, located on the X-chromosome. *Hun* arose from a partial duplication of the *Bällchen* gene, which is on chromosome 3R. *Hun* lacks 3' coding sequence with respect to *Bällchen*, but has gained 33 amino acids from a nearby intergenic sequence. Further, while *Bällchen* is expressed ubiquitously, *Hun* shows testes-specific expression (Arguello et al., 2006).

The sequence similarity that exists between completely duplicated gene copies and parental gene copies is often sufficient to detect homologues in a whole range of organisms. However, this is often not the case for partially duplicated genes, especially if the sequence common to both duplicates is short and the rate of divergence of the novel gene duplicate is abnormally high. As a result, many partially duplicated genes are identified as orphan or lineage-specific genes, that is, genes that do not yield any significant hits in database protein

searches of more distant organisms (Chen et al., 2010; Domazet-Loso & Tautz, 2003; Toll-Riera et al., 2009a). In a recent study that showed that newly formed genes in *Drosophila melanogaster* are as likely to perform essential functions as older genes, it was found that 28 out of the 50 new genes that had arisen through gene duplication corresponded to partial duplications (Chen et al., 2010). These young genes were found to evolve very rapidly, showing a median of 47.3% divergence, at the amino-acid level, from their parents. In an analysis of the mechanisms of formation of primate-specific genes, we observed that about 24% of the newly formed genes had originated through gene duplication, frequently involving partial gene duplication and the recruitment of additional sequences (Toll-Riera et al., 2009a). One example is human XAGE-1, a cancer/testis-associated gene that has partial homology to human XAGE-2, a gene that is well conserved in other mammals. The similarity is limited to the C-terminal half of the orphan XAGE-1 protein. We showed that, in the conserved region, the rate of amino acid sequence evolution of XAGE-1 was double that of XAGE-2, suggesting that the recruitment of additional sequences in XAGE-1 resulted in a marked asymmetry in the evolutionary rates of the two copies.

Partial gene duplication is likely to be very important for the formation of novel gene structures and the evolution of new protein functions, but studies focusing on this type of gene duplication are still scarce. To shed new light on this issue, we decided to analyse the evolutionary patterns of several primate-specific genes (orphan genes) formed, at least partially, by gene duplication. The results show that increased evolutionary rates in the partially duplicated copy are the norm, reinforcing the role of partial gene duplication in the formation of novel genes with distinct functions.

2. Results

Here we use a similar approach to that employed in Toll-Riera et al. (Toll-Riera et al., 2009a) to identify a set of primate-specific genes that show significant similarity to human genes (parental genes) that are well conserved in non-primate species. We investigate the differences in the rate of evolution of the novel and parental genes and discuss the role of partial duplication in increasing the protein functional repertoire.

2.1 Identification of primate lineage-specific genes formed by gene duplication

We identified a set of genes present in human and macaque but absent in 13 non-primate genomes (Mus musculus, Rattus norvegicus, Bos Taurus, Canis familiaris, Gallus gallus, Xenopus tropicalis, Danio rerio, Takifugu rubripes, Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Arabidopsis thaliana). The existence of a homologue in a specific genome was determined by the presence of a BLASTP (Altschul et al., 1997) hit with an expectation value (E-value) smaller than 10^{-4}, as previously described (Alba and Castresana 2005). Orphan genes were defined as those for which we could not detect any homologues in any of the species mentioned above. As they were, by definition, present in human and macaque, our collection of orphan genes corresponded to primate-specific genes, presumably formed after the split of the rodent and primate branches and before the speciation of the human and macaque lineages. Once we had this set of orphan genes, we investigated which ones could have arisen through gene duplication by performing BLASTP searches against all human proteins, using a relaxed E-value ($E<0.5$). We kept those cases for which we could identify human

paralogues that were not primate-specific. In such cases, the closest hit in human was considered the putative human parental gene, and the closest non-primate orthologue of the parental gene was taken as the outgroup gene (Figure 1). Protein sequences were aligned with T-Coffee (Notredame et al., 2000), and the alignments between primate-specific genes and parental genes were carefully examined to discard any spurious associations. We also removed any regions that were completely divergent (non-alignable) between the orphan and parental genes.

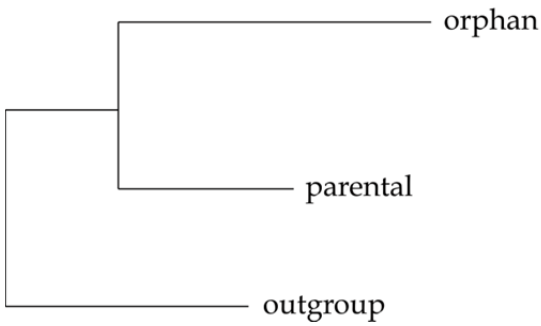


Fig. 1. Tree topology corresponding to gene families containing duplicated orphan genes. The orphan and parental genes are from human, the outgroup gene from a non-primate species.

The final set consisted of 14 orphan genes. Table 1 shows the orphan, parental and outgroup gene names, protein identifiers, and the percent of parental protein that could be reliably aligned with the orphan protein, corresponding to the portion of the protein that had duplicated. Of the 14 orphan genes, 4 represented single copies and the rest belonged to orphan gene families. In only one case, dermcidin, sequence similarity supported a complete gene duplication event.

We used the protein multiple alignments to estimate the number of amino acid substitutions per site (K) in the orphan, parental and outgroup branches. We used PROML, a maximum likelihood based method in the Phylip package for this purpose (Felsenstein, 2005). The results of these computations are discussed below.

2.2 Dermcidin and lacritin

The first example of an orphan gene that arose through gene duplication is dermcidin. This gene encodes a short protein of 110 amino acids in length. The corresponding parental gene is lacritin, which has orthologues in other mammals, and is located on chromosome 12 adjacent to the dermcidin gene. The two genes have a similar exonic structure, and although they are highly divergent, sequence similarity between the two is still detectable (Wang et al., 2006). Dermcidin is secreted in sweat glands, having an antimicrobial activity (Schittek et al., 2001), and may also be involved in neural survival and cancer (Porter et al., 2003), whereas lacritin is expressed in the lacrimal glands (Ma et al., 2008).

Figure 2 shows the alignment of the complete protein sequences of human dermcidin, human lacritin and cat lacritin. The number of amino acid substitutions per site in the orphan branch was 1.026, about double the number of amino acid substitutions per site in the parental and outgroup branches (0.434 and 0.505, respectively).

Orphan Name	Parental name	Parental protein	Outgroup	%
Dermcidin	Lacritin	ENSP00000257867	ENSFCAP00000009317 (cat)	100%
FAM9A	Synaptonemal complex protein 3	ENSP00000266743	ENSMUSP00000020252 (mouse)	87.29%
FAM9B	idem	idem	idem	idem
FAM9C	idem	idem	idem	idem
AL023807.2	AL365202.1	ENSP00000382846	ENSCINP00000011125 (vase tunicate)	64.42%
XAGE-1A	XAGE-2	ENSP00000333775	XP_001249434.1 (cow)	36.03%
XAGE-1B	idem	idem	idem	idem
XAGE-1C	idem	idem	idem	idem
XAGE-1D	idem	idem	idem	idem
XAGE-1E	idem	idem	idem	idem
NPIP-like 1	Acyl-CoA synthetase medium-chain family member 1	ENSP00000428098	ENSMUSP00000036140 (mouse)	18,07%
C2orf27A	Ral guanine nucleotide dissociation stimulator like-4	ENSP00000290691	ENSCAFP00000031102 (dog)	12.05%
C2orf27B	idem	idem	idem	idem
AL133216.1	Arsenite-resistance protein 2	ENSP00000314491	ENSMUSP00000043123 (mouse)	9.36%

Table 1. List of primate-specific genes that have arisen by gene duplication. Protein identifiers are from Ensembl (ENSP) or Genbank (XP). % refers to the percentage of the parental protein that showed homology to the orphan protein.

2.3 Partially duplicated orphan genes

The remaining primate-specific genes that have arisen through gene duplication corresponded to partial duplications of the parental gene (Table 1). They included 3 individual genes (AL023807.2, NPIP-like 1 and AL133216.1) and 3 gene families (FAM9, XAGE-1 and C2orf27). The percentage of protein sequence from the parental protein that could by identified as homologous in the orphan protein ranged from 9.4 to 87.3% (Table 1). With the exception of NPIP-like 1, the orphan gene is located on a different chromosome from the parental gene, although the presence of introns in all orphan genes suggests that they were not retrotransposed copies. We aligned the conserved regions of orphan, parental and outgroup proteins (Figure 3). These alignments were used for the estimation of the

number of amino acid substitutions per site in the orphan, parental and outgroup branches. We also investigated the presence of any known protein domains in the region conserved between parental and orphan proteins, using the Pfam web server (Finn et al., 2010).

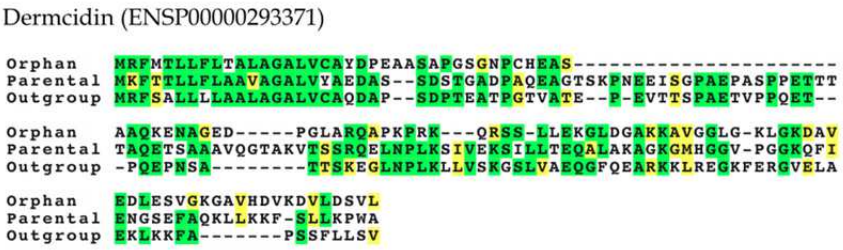


Fig. 2. Alignment of dermcidin (orphan), human lacritin (parental) and cat lacritin (outgroup) proteins. Identical residues are in green, similar residues in yellow.

The FAM9 family (family with sequence similarity 9) is composed of three genes: FAM9A, FAM9B and FAM9C. They are all predicted to have a Cor1/Xlr/Xmr domain in the region of similarity to the parental gene (E-values ranging from 0.049 to $4.4e^{-13}$), related to meiotic prophase chromosomes. The parental gene, synaptonemal complex protein 3 (SYCP3) is involved in the assembly of the synaptonemal complex during meiosis (Martinez-Garay et al., 2002), but the exact physiological functions of the FAM9 proteins remain unknown.

The largest orphan gene family is XAGE-1, which has 5 members with identical amino acid sequences that are contiguous on the X chromosome. The region conserved between the XAGE-1s and XAGE-2 includes the GAGE domain. The function of GAGE (G antigen) and XAGE (X antigen) domains is unknown, but XAGE and GAGE proteins have been implicated in several human cancers (Zendman et al., 2002).

The two genes belonging to the C2orf27 family are contiguous in the genome, though C2orf27A is located on the forward strand of chromosome 2 whereas C2orf27B is located on the reverse strand. Their function is unknown, but they derive from a protein annotated as Ral guanine nucleotide dissociation stimulator-like 4. The parental protein contains the RasGEF domain, which is a guanine nucleotide exchange factor for Ras-like small GTPases. The duplicated region overlaps minimally with this domain (14 amino acids).

NPIP-like 1 belongs to the nuclear pore complex-interacting protein (NPIP) family. The parental protein contains two AMP-binding domains that are at the N-terminal region of the protein, not the area conserved in the orphan protein, which is the C-terminal part. The NPIP family (Nuclear Pore Interacting Protein), also named *morpheus*, is located on a duplicated segment of chromosome 16. It has been suggested to have experienced a burst of positive selection during the emergence of *Homininae* (Johnson et al., 2001).

Finally, AL133216.1 and AL023807.2 are two primate-specific genes of unknown function containing putative coding sequences of length 151 and 121 amino acids respectively. The parental copy of AL133216.1 modulates arsenic sensitivity, is involved in cell cycle progression, and in RNA-mediated gene silencing by microRNA (Gruber et al., 2009). It also contains an arsenite-resistance protein 2 domain (Pfam hit E-value = $3.1e^{-18}$). The orphan

copy does not contain this domain even though it is located in the conserved region, suggesting that this region has lost its ancestral function in the orphan protein.
Table 2 shows the estimated amino acid substitution rates in the orphan, parental and outgroup branches. In the case of identical copies (for example C2orf27A and C2orf27B) only one is taken as representative. In the case of divergent copies (the FAM9 family) the amino acid substitution rates are summed up for all branches from the ancestor to the derived node (see Figure 4). In all cases the duplicated protein is evolving much faster than the parental gene, and in some cases, such as the FAM9 and NPIP-like 1 proteins, more than six times faster. These results indicate that orphan proteins are evolving under much more relaxed constraints, and/or adapting to a new function with respect to their parental copies.

Orphan Name	Orphan protein	Orphan	Parental	Outgroup
Dermcidin	ENSP00000293371	1.02595	0.43458	0.5055
FAM9A	ENSP00000370391	1.28971	0.17014	0.15423
FAM9B	ENSP00000318716	1.13565	0.17014	0.15423
FAM9C	ENSP00000369999	1.15328	0.17014	0.15423
AL023807.2	ENSP00000381423	0.19840	0.12203	0.23096
XAGE-1A	ENSP00000382698	0.52961	0.17820	0.94188
NPIP-like 1	ENSP00000350444	0.40089	0.02020	0.11929
C2orf27B	ENSP00000304065	0.55865	0.21080	0.68342
AL133216.1	ENSP00000382606	1.37580	0.00010	0.00010

Table 2. Estimated number of amino acid substitutions per site (K) for orphan, parental and outgroup branches. Orphan protein identifiers are from Ensembl. See Table 1 for more details.

2.4 Role of low-complexity sequences

Low complexity regions (LCRs) are sequences in which one or a few residues are highly overrepresented. Several studies have shown that duplicated gene copies can gain new functions through the acquisition of LCRs (Fondon & Garner, 2004; Salichs et al., 2009). It has also been shown that young proteins contain more LCRs than old proteins (Alba & Castresana, 2005). Therefore, we inspected the presence of LCRs in our set of orphan proteins using the SEG algorithm with default parameters (Wootton & Federhen, 1996).
We found that the FAM9A protein contained a very conspicuous low-complexity sequence. Figure 4 shows the detailed phylogenetic tree of the FAM9 gene family (including the parental and outgroup SYCP3 genes). The ancestral FAM9 evolved very rapidly and eventually underwent two duplication events, leading to FAM9A, FAM9B and FAM9C. The multiple alignment of the region surrounding the LCR in FAM9A shows how, from a small region containing several acidic residues in SYCP3, a larger acidic region was formed in the common FAM9 ancestor, which finally expanded to a 75 amino acid stretch in FAM9A containing a long glutamic acid repeat, as well as poly-alanine and poly-glycine repeats.
As is the case for the SYCP3 proteins, all three human FAM9 proteins show testis-specific expression. However, the cellular localization is different depending on the protein studied: FAM9B and FAM9C are localized in the nucleus with low protein levels being detectable in the cytoplasm, whereas FAM9A is present at high levels in the nucleolus (Martinez-Garay et al., 2002). The distinct location of FAM9A may be due to the long glutamic acid repeat, as

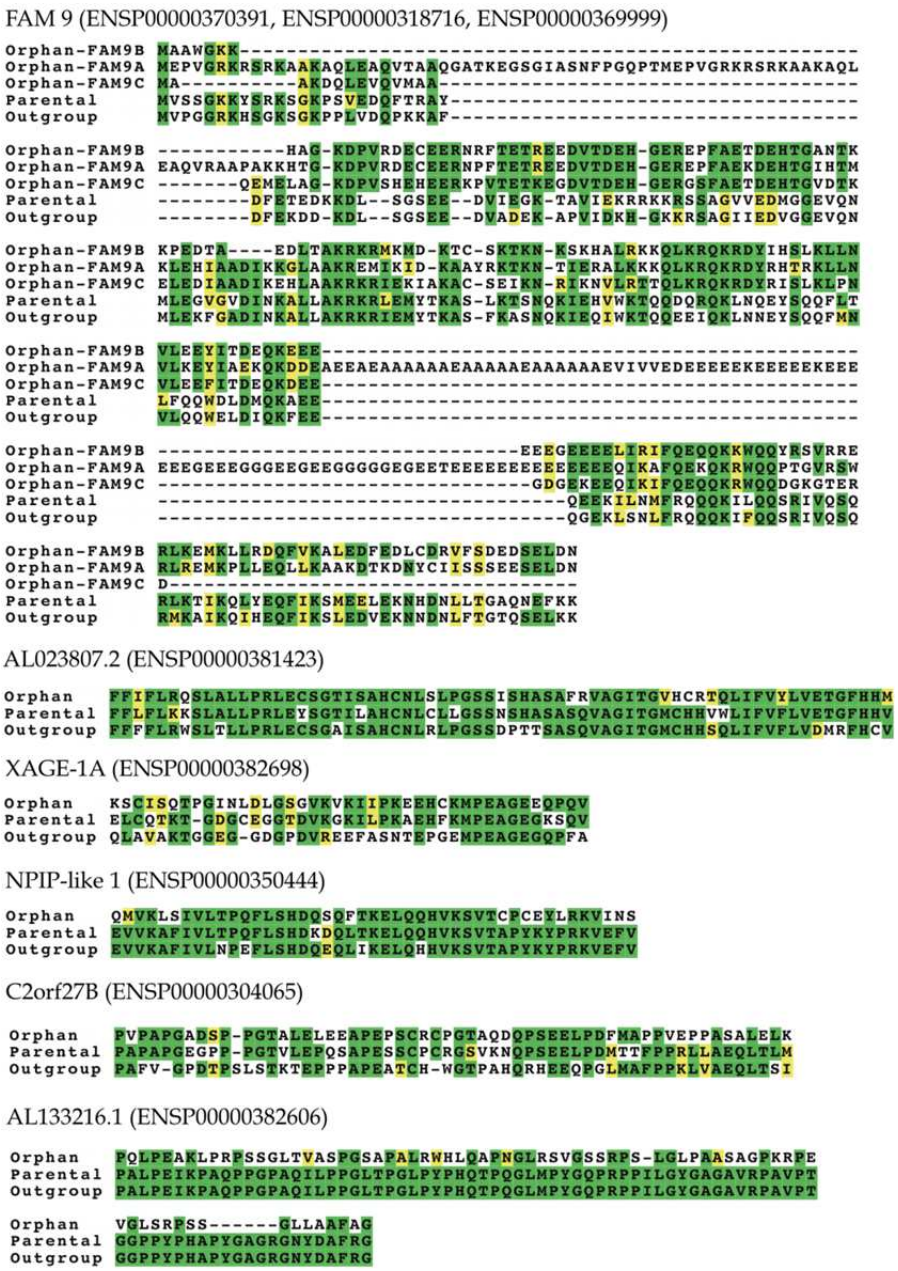


Fig. 3. Multiple alignments of the conserved regions between orphan, parental and outgroup proteins. For the XAGE-1 family only XAGE-1A is shown, as the other orphan sequences were identical at the amino acid level. The same is true for the C2orf27 family. Identical residues are in green, similar residues in yellow. See Table 1 for more details.

Fig. 4. Phylogenetic tree of the FAM9 gene family. Branch lengths correspond to the estimated number of amino acid substitutions per site, using the alignment in Fig 3. The protein alignment shown corresponds to exon 5 in FAM9B and FAM9C and to exon 6 in FAM9A, human SYCP3 and mouse SYCP3. The expanded low-complexity region in FAM9A is depicted above the alignment.

acidic clusters have been shown to mediate protein nucleolar retention (Ochs et al., 1996; Shu-Nu et al., 2000; Ueki et al., 1998).In FAM9A, the low complexity sequence is located within the Cor1/Xlr/Xmr conserved region, perhaps interfering with its function. In fact, FAM9A shows higher sequence divergence from the common ancestor than FAM9B.

3. Discussion

The role of partial gene duplication in the formation of novel genes is still poorly understood, although recent reports in *Drosophila* (Chen et al., 2010; Zhou et al., 2008) and *C.elegans* (Katju & Lynch, 2006; 2003) indicate that partially duplicated gene copies are very frequent. The present study analyses a set of primate-specific genes formed by partial gene duplication. We find that the rate of divergence of the partially duplicated copy is, in all cases, higher that the rate of divergence of the parental copy, generalizing previous observations for XAGE1-A (Toll-Riera et al., 2009a). This, together with the fact that most partially duplicated genes recruit additional sequences, strengthens the notion that partial duplication is a major process for the formation of genes with novel structures and functions. In these genes, any remaining similarity to the homologous proteins is being quickly erased by high sequence turnover. As a consequence, distant homologues are difficult to identify and these proteins end up being classified as orphans. This fits the model of Domazet-Loso and Tautz in explaining the high number of orphan genes in *Drosophila*: orphan genes are created by gene duplication followed by a period of rapid sequence divergence that erases the similarity with its homologues (Domazet-Loso & Tautz, 2003). Although we now have evidence that not all orphan genes are generated in this manner (Toll-Riera et al., 2009a; Toll-Riera et al., 2009b; Zhou et al., 2008), a significant portion is.

A large fraction of the duplicated gene copies that become fixed in a population are subsequently lost, presumably because the new copy is completely redundant and thus dispensable. However, the formation of chimeric gene structures, encoding part of an existing protein together with additional sequences, could in principle favour their retention, as these genes are not going to be functionally equivalent to the ancestral gene (Patthy, 1999; Zhou et al., 2008). In support of this, in *Drosophila* it was found that the proportion of novel genes corresponding to complete gene duplications decreased with gene age, suggesting that complete gene duplications had a shorter lifespan than partial gene duplications (Zhou et al., 2008).

Orphan genes are in general poorly annotated and their function is unknown in most cases (Kuo & Kissinger, 2008). The fact that organisms had lived perfectly well without them until recent times when they made their appearance, has led scientists to think that orphan genes were, for the most part, dispensable. However, a recent study by Chen and colleagues (Chen et al., 2010) has challenged this viewpoint. In their study, the authors identified new young genes in Drosophila melanogaster (around 34 million years old) and designed RNA interference lines to knoch each of them out (KO). Surprisingly, they found that 30% of these young genes KOs were lethal, as Drosophila could not survive without them. These young genes had mainly arisen through duplication and they showed higher evolutionary rates than the parental gene, indicating the action of positive selection, or relaxation of functional constraints. They hypothesized that new genes are quickly integrated into existing pathways, and hence many of them soon become essential for the viability of the organism.

Capra and colleagues (Capra et al., 2010) compared the evolutionary patterns of genes that arose by duplication with those that did not (named novel genes). They argued that the evolutionary pressures should be different in each case as, contrary to novel genes, duplicated genes were functionally and structurally well formed from birth. They showed that although duplicated genes are initially more integrated into cellular networks, both types of new genes gain functions and interactions with time, though novel genes do it more rapidly than duplicated genes. Additionally, novel genes also increase in length through the incorporation of transposable elements or surrounding sequences. This increase in length could be related with the rapid gain of function and interactions experienced by novel genes. They also found that genes tended to interact with genes similar in age and mode of origin. Thus, the mechanism by which a gene originates seems to significantly impact on its subsequent evolution.

Several studies have demonstrated that duplicated genes show increased protein evolutionary rates with respect to non-duplicated genes in the same lineage (Castillo-Davis et al., 2004; Cusack & Wolfe, 2007; Kondrashov et al., 2002; Lynch & Conery, 2000; Nembaware et al., 2002; Scannell & Wolfe, 2008; Van de Peer et al., 2001). Here we identified a very strong asymmetry in the rates of evolution of the newly evolved copy (orphan) and the well-conserved copy (parental), the former evolving much faster than the latter. Surprisingly, the parental protein copy did not evolve consistently faster than the outgroup protein (not duplicated), highlighting the fact that we are dealing with a special type of gene duplication in which the copy containing the partially duplicated segment rapidly departs from the ancestral family, which remains essentially unaffected.

Increased evolutionary rates may reflect either relaxation of purifying selection, positive selection, or the combined effects of both these forces. The orphan genes under study predated the split of the human and macaque lineages, which occurred approximately 25 million years ago so, if relaxed selection was the only factor for their increased rates, the genes should by now have become pseudogenes and not be expressed. However, all genes were expressed at the RNA level in one or several tissues. Therefore we must hypothesize that, at least to some extent, positive selection has influenced the evolution of these genes.

We compared the rates of evolution of the protein regions that were conserved between orphan and parental proteins, but what about the unique sequences contained in the orphan proteins? These sequences lacked any similarity to other protein-coding genes, so they may be ancestral non-coding sequences that have been co-opted for a coding function (Long et al., 2003). Genes generated *de novo* from non-coding sequences are among the fastest evolving genes (Levine et al., 2006), and there is no reason to believe that unique sequences

in orphan proteins will evolve slower than the conserved protein regions, rather the contrary would seem more logical. In a previous study we showed that the non-synonymous to synonymous nucleotide substitution rates of primate-specific genes, measured for human and macaque orthologues, were, on average, twice as high as those of mammalian-specific genes and five times higher than those of deeply conserved eukaryotic proteins (Toll-Riera et al., 2009a). The differences in amino acid substitution rates between orphan and parental genes described here reinforce the idea that the evolution of a new gene is strongly associated with very rapid sequence change.

4. Concluding remarks and future research

We have examined the evolutionary dynamics of a group of novel primate-specific genes (orphan genes) that have arisen by gene duplication. These genes typically form new structures in which only part of the protein sequence is shared with the parental copy, presumably because of partial gene duplication, and the rest of the protein sequence is unique. The orphan proteins accumulate a much larger number of amino acid substitutions per site than the parental proteins, denoting rapid functional diversification. The parental gene copies appear to act as "donors" of sequence but do not experience any obvious sequence evolution alterations, thus they probably preserve their ancestral functions. Future research in this area, using computational as well as experimental studies, should help clarify how frequent is partial gene duplication with respect to complete gene duplication, the differences in gene copy survival in both cases, and how partial and complete gene duplication contribute to the generation of evolutionary novelties.

5. Acknowledgments

We received financial support from Ministerio de Educación, Gobierno de España (FPU to M.T-R, BIO2009-08160), Generalitat de Catalunya (FI to S.L.), Fundació Javier Lamas Miralles (Ajut Predoctoral Javier Lamas Miralles to N.R-T) and Institució Catalana de Recerca i Estudis Avançats (M.M.A).

6. References

Alba, M.M. & Castresana, J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol* 22(3): 598-606.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-3402.

Arguello, J.R., Chen, Y., Yang, S., Wang, W. & Long, M. 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila. *PLoS Genet* 2(5): e77.

Bailey, J.A., Liu, G. & Eichler, E.E. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73(4): 823-834.

Cai, J., Zhao, R., Jiang, H. & Wang, W. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* 179(1): 487-496.

Capra, J.A., Pollard, K.S. & Singh, M. 2010. Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol* 11(12): R127.

Castillo-Davis, C.I., Hartl, D.L. & Achaz, G. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res* 14(8): 1530-1536.

Castresana, J., Guigo, R. & Alba, M.M. 2004. Clustering of genes coding for DNA binding proteins in a region of atypical evolution of the human genome. *J Mol Evol* 59(1): 72-79.

Chen, S., Zhang, Y.E. & Long, M. 2010. New genes in Drosophila quickly become essential. *Science* 330(6011): 1682-1685.

Cusack, B.P. & Wolfe, K.H. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* 24(3): 679-686.

Domachowske, J.B. & Rosenberg, H.F. 1997. Eosinophils inhibit retroviral transduction of human target cells by a ribonuclease-dependent mechanism. *J Leukoc Biol* 62(3): 363-368.

Domazet-Loso, T. & Tautz, D. 2003. An evolutionary analysis of orphan genes in Drosophila. *Genome Res* 13(10): 2213-2219.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450(7167): 203-218.

Farre, D. & Alba, M.M. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Mol Biol Evol* 27(2): 325-335.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. & Bateman, A. 2010. The Pfam protein families database. *Nucleic Acids Res* 38(Database issue): D211-222.

Fondon, J.W., 3rd & Garner, H.R. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A* 101(52): 18058-18063.

Gilad, Y., Man, O. & Glusman, G. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* 15(2): 224-230.

Gruber, J.J., Zatechka, D.S., Sabin, L.R., Yong, J., Lum, J.J., Kong, M., Zong, W.X., Zhang, Z., Lau, C.K., Rawlings, J., Cherry, S., Ihle, J.N., Dreyfuss, G. & Thompson, C.B. 2009. Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell* 138(2): 328-339.

Gu, Z., Nicolae, D., Lu, H.H. & Li, W.H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18(12): 609-613.

Guo, W.J., Li, P., Ling, J. & Ye, S.P. 2007. Significant comparative characteristics between orphan and nonorphan genes in the rice (Oryza sativa L.) genome. *Comp Funct Genomics*: 21676.

Haldane, J.B.S. 1932. The causes of evolution. London: Longmans and Green.

Heinen, T.J., Staubach, F., Haming, D. & Tautz, D. 2009. Emergence of a new gene from an intergenic region. *Curr Biol* 19(18): 1527-1531.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256(1346): 119-124.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018): 695-716.

Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M. & Eichler, E.E. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413(6855): 514-519.

Katju, V. & Lynch, M. 2003. The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. *Genetics* 165(4): 1793-1803.

Katju, V. & Lynch, M. 2006. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. *Mol Biol Evol* 23(5): 1056-1067.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T.C. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* 25(9): 404-413.

Knowles, D.G. & McLysaght, A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* 19(10): 1752-1759.

Kondrashov, F.A. & Koonin, E.V. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* 20(7): 287-290.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biol* 3(2): RESEARCH0008.

Kuo, C.H. & Kissinger, J.C. 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria. *BMC Evol Biol* 8: 108.

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A. & Begun, D.J. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A* 103(26): 9935-9939.

Long, M., Betran, E., Thornton, K. & Wang, W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 4(11): 865-875.

Lynch, M. & Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494): 1151-1155.

Ma, P., Wang, N., McKown, R.L., Raab, R.W. & Laurie, G.W. 2008. Focus on molecules: lacritin. *Exp Eye Res* 86(3): 457-458.

Makova, K.D. & Li, W.H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* 13(7): 1638-1645.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. & Kaessmann, H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* 3(11): e357.

Martinez-Garay, I., Jablonka, S., Sutajova, M., Steuernagel, P., Gal, A. & Kutsche, K. 2002. A new gene family (FAM9) of low-copy repeats in Xp22.3 expressed exclusively in testis: implications for recombinations in this region. *Genomics* 80(3): 259-267.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915): 520-562.

Muller, H.J. 1935. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica* 17: 237-252.

Nathans, J., Thomas, D. & Hogness, D.S. 1986. Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science* 232(4747): 193-202.

Nekrutenko, A. & Li, W.H. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 17(11): 619-621.

Nembaware, V., Crum, K., Kelso, J. & Seoighe, C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res* 12(9): 1370-1376.

Notredame, C., Higgins, D.G. & Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1): 205-217.

Ochs, R.L., Stein, T.W., Jr., Chan, E.K., Ruutu, M. & Tan, E.M. 1996. cDNA cloning and characterization of a novel nucleolar protein. *Mol Biol Cell* 7(7): 1015-1024.

Ohno, S. 1970. Evolution by gene duplication. *New York: Springer-Verlag*.

Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E. & Guigo, R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* 16(1): 37-44.

Patthy, L. 1999. Genome evolution and the evolution of exon-shuffling--a review. *Gene* 238(1): 103-114.

Porter, D., Weremowicz, S., Chin, K., Seth, P., Keshaviah, A., Lahti-Domenici, J., Bae, Y.K., Monitto, C.L., Merlos-Suarez, A., Chan, J., Hulette, C.M., Richardson, A., Morton, C.C., Marks, J., Duyao, M., Hruban, R., Gabrielson, E., Gelman, R. & Polyak, K. 2003. A neural survival factor is a candidate oncogene in breast cancer. *Proc Natl Acad Sci U S A* 100(19): 10931-10936.

Rosenberg, H.F. & Dyer, K.D. 1995. Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J Biol Chem* 270(37): 21539-21544.

Salichs, E., Ledda, A., Mularoni, L., Alba, M.M. & de la Luna, S. 2009. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet* 5(3): e1000397.

Scannell, D.R. & Wolfe, K.H. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res* 18(1): 137-147.

Schittek, B., Hipfel, R., Sauer, B., Bauer, J., Kalbacher, H., Stevanovic, S., Schirle, M., Schroeder, K., Blin, N., Meier, F., Rassner, G. & Garbe, C. 2001. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. *Nat Immunol* 2(12): 1133-1137.

Shu-Nu, C., Lin, C.H. & Lin, A. 2000. An acidic amino acid cluster regulates the nucleolar localization and ribosome assembly of human ribosomal protein L22. *FEBS Lett* 484(1): 22-28.

Siepel, A. 2009. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res* 19(10): 1693-1695.

Siew, N. & Fischer, D. 2003. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* 53(2): 241-251.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. & Alba, M.M. 2009a. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* 26(3): 603-612.

Toll-Riera, M., Castelo, R., Bellora, N. & Alba, M.M. 2009b. Evolution of primate orphan proteins. *Biochem Soc Trans* 37(Pt 4): 778-782.

Ueki, N., Kondo, M., Seki, N., Yano, K., Oda, T., Masuho, Y. & Muramatsu, M. 1998. NOLP: identification of a novel human nucleolar protein and determination of sequence requirements for its nucleolar localization. *Biochem Biophys Res Commun* 252(1): 97-102.

Van de Peer, Y., Taylor, J.S., Braasch, I. & Meyer, A. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol* 53(4-5): 436-446.

Wang, J., Wang, N., Xie, J., Walton, S.C., McKown, R.L., Raab, R.W., Ma, P., Beck, S.L., Coffman, G.L., Hussaini, I.M. & Laurie, G.W. 2006. Restricted epithelial proliferation by lacritin via PKCalpha-dependent NFAT and mTOR pathways. *J Cell Biol* 174(5): 689-700.

Wootton, J.C. & Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 266: 554-571.

Yang, S., Arguello, J.R., Li, X., Ding, Y., Zhou, Q., Chen, Y., Zhang, Y., Zhao, R., Brunet, F., Peng, L., Long, M. & Wang, W. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. *PLoS Genet* 4(1): e3.

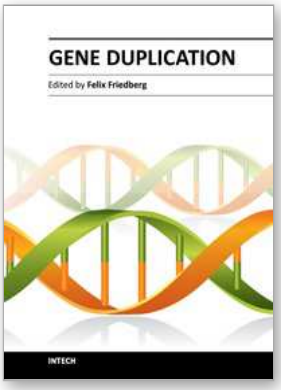
Zendman, A.J., Van Kraats, A.A., Weidle, U.H., Ruiter, D.J. & Van Muijen, G.N. 2002. The XAGE family of cancer/testis-associated genes: alignment and expression profile in normal tissues, melanoma lesions and Ewing's sarcoma. *Int J Cancer* 99(3): 361-369.

Zhang, J., Rosenberg, H.F. & Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* 95(7): 3708-3713.

Zhang, J., Zhang, Y.P. & Rosenberg, H.F. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet* 30(4): 411-415.

Zhou, Q. & Wang, W. 2008. On the origin and evolution of new genes--a genomic and experimental perspective. *J Genet Genomics* 35(11): 639-648.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. & Wang, W. 2008. On the origin of new genes in Drosophila. *Genome Res* 18(9): 1446-1455.



Gene Duplication

Edited by Prof. Felix Friedberg

ISBN 978-953-307-387-3
Hard cover, 400 pages
Publisher InTech
Published online 17, October, 2011
Published in print edition October, 2011

The book Gene Duplication consists of 21 chapters divided in 3 parts: General Aspects, A Look at Some Gene Families and Examining Bundles of Genes. The importance of the study of Gene Duplication stems from the realization that the dynamic process of duplication is the "sine qua non" underlying the evolution of all living matter. Genes may be altered before or after the duplication process thereby undergoing neofunctionalization, thus creating in time new organisms which populate the Earth.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Macarena Toll-Riera, Steve Laurie, Núria Radó-Trilla and M.Mar Alba (2011). Partial Gene Duplication and the Formation of Novel Genes, Gene Duplication, Prof. Felix Friedberg (Ed.), ISBN: 978-953-307-387-3, InTech, Available from: http://www.intechopen.com/books/gene-duplication/partial-gene-duplication-and-the-formation-of-novel-genes

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821