Origin of Primate Orphan Genes: A Comparative Genomics Approach

Macarena Toll-Riera,† Nina Bosch,‡ Nicolás Bellora,* Robert Castelo,† Lluis Armengol,‡ Xavier Estivill,†‡ and M. Mar Albà*†§*

*Evolutionary Genomics Group, Biomedical Informatics Research Programme, Fundació Institut Municipal d'Investigació Mèdica, Barcelona, Spain; †Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain; ‡Genes and Disease Program, Centre for Genomic Regulation (CRG-UPF) and CIBERESP, Barcelona, Catalonia, Spain; and §Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

Genomes contain a large number of genes that do not have recognizable homologues in other species and that are likely to be involved in important species-specific adaptive processes. The origin of many such "orphan" genes remains unknown. Here we present the first systematic study of the characteristics and mechanisms of formation of primate-specific orphan genes. We determine that codon usage values for most orphan genes fall within the bulk of the codon usage distribution of bona fide human proteins, supporting their current protein-coding annotation. We also show that primate orphan genes display distinctive features in relation to genes of wider phylogenetic distribution: higher tissue specificity, more rapid evolution, and shorter peptide size. We estimate that around 24% are highly divergent members of mammalian protein families. Interestingly, around 53% of the orphan genes contain sequences derived from transposable elements (TEs) and are mostly located in primate-specific genomic regions. This indicates frequent recruitment of TEs as part of novel genes. Finally, we also obtain evidence that a small fraction of primate orphan genes, around 5.5%, might have originated de novo from mammalian noncoding genomic regions.

Introduction

The gene content of genomes varies in different lineages, indicating the existence of general and widespread mechanisms of gene birth and loss (Long et al. 2003). The study of lineage-specific genes has generated much interest in recent years as these genes are particularly important in helping understand recent adaptive processes. It is well established that gene duplication is a major mechanism for the formation of novel genes (Ohno 1970; Zhou et al. 2008), including segmental DNA duplications and retrotransposition events (Fortna et al. 2004; Marques et al. 2005). Genes formed by gene duplication can normally be grouped into gene families that include members from distant species. For example, Kruppel-associated box-Zinc finger proteins (KRAB-ZNF) have undergone numerous lineage-specific gene duplications in mammals (Eichler et al. 1998; Castresana et al. 2004), but the copies retain significant sequence similarity to vertebrate KRAB proteins.

A special class of lineage-specific genes are "orphan" genes, which are genes that do not show homology to sequences in other species (Fischer and Eisenberg 1999). Each newly sequenced genome contains a significant number of such genes. For example, among 60 fully sequenced microbial genomes, 14% of genes are species-specific orphans (Siew and Fischer 2003), and about 18% of genes in *Drosophila* are restricted to the *Drosophila* group (Zhang et al. 2007). They typically encode short proteins and show high nonsynonymous substitution rates, but their functions are largely unknown (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004).

Due to their lack of phylogenetic conservation, the origin of orphan genes has remained elusive. One proposed scenario is that they derive from gene duplication events in which one copy has accumulated so many sequence changes that the ancestral similarity is no longer detectable (Domazet-Loso and Tautz 2003). This process should involve abnormally high sequence divergence rates, given that for short timescales (e.g., within mammals), sequence similarity methods should suffice to detect homology for most genes, including those that are rapidly evolving (Alba and Castresana 2007). Non-deleterious frameshift mutations after gene duplication could also potentially generate novel protein-coding genes (Ohno 1984), as recently shown in mouse (Okamura et al. 2006). A second scenario, which does not involve gene duplication, is direct birth of new protein-coding genes from noncoding genomic regions, such as introns, gene untranslated regions, or intergenic regions. This mechanism has recently been reported in *Drosophila* (Levine et al. 2006; Zhou et al. 2008) and yeast (Cai et al. 2008) but has not been observed in mammals. Sequences derived from transposable elements (TEs), such as Alu repeats in primates, can be incorporated into preexisting human genes, often forming new exons (Makalowski et al. 1994). Interestingly, TE insertions have been suggested to be implicated in the creation of two new mouse genes (Nekrutenko and Li 2001), but the global impact of this process in the formation of completely new genes in mammalian genomes remains unknown.

The aim of the present study is to estimate the relative importance of the different mechanisms of gene formation in primate genomes. For this, we examine the patterns of conservation of mammalian syntenic genomic regions, perform extensive gene homology searches, and map the positions of TEs onto human genes. We conclude that about one-fourth of the orphan genes are likely to have been formed by gene duplication processes, that TEs are likely to have played a very important role in the formation of the remaining genes, and that some new genes may have arisen de novo from noncoding sequences by TE-independent mechanisms.

Materials and Methods

Sequence Data Sets

Human–macaque orthologous protein pairs, their corresponding gene coding and protein sequences, expression

Key words: primate-specific gene, orphan gene, novel gene formation.

E-mail: malba@imim.es.

Mol. Biol. Evol. 26(3):603–612. 2009
doi:10.1093/molbev/msn281
Advance Access publication December 8, 2008

data, and chromosome number were obtained using Biomart at Ensembl, version 48 (Flicek et al. 2008). We used build NCBI36 of the human genome and Mmul_1 of the macaque genome. When more than one protein sequence per gene was available, we chose the longest one. In order to build groups of primate genes of different phylogenetic distribution, we obtained protein sequences from 14 additional eukaryotic complete genomes. Sequences from *Saccharomyces pombe* and *Arabidopsis thaliana* were downloaded from the Cogent Database release 153 (Goldovsky et al. 2005). Sequences from *Pan troglodytes* (CHIMP1), *Mus musculus* (NCBIM36), *Rattus norvegicus* (RGSC3.4), *Bos Taurus* (Btau_2.0), *Canis familiaris* (BROADD1), *Gallus gallus* (WASHUC1), *Xenopus tropicalis* (JGI4.1), *Danio rerio* (ZFISH6), *Saccharomyces cerevisiae* (SGD1.01), *Caenorhabditis elegans* (WS180), *Drosophila melanogaster* (BDGP54), and *Takifugu rubripes* (FUGU4) were downloaded from Ensembl.

Classification by Phylogenetic Distribution

We classified the human proteins into four gene age groups—Primates, Mammals, Vertebrates, and Eukarya—using BlastP sequence similarity searches (Altschul et al. 1997). The data set comprised 20,764 human proteins (Ensembl 48). We considered that there existed a homologue in another genome if there was at least one BlastP hit with an expectation value (*E* value) smaller than 10^{-4}, as previously described (Alba and Castresana 2005). To avoid false positives caused by low-complexity sequences, we filtered this type of region from the human sequences using the SEG program (Wootton and Federhen 1996). If the human protein had any homologue in *P. troglodytes* and *Macaca mulatta* but not in the other genomes, it was classified as Primates (primate orphan genes). If the protein had homologues in the other 4 mammalian species but not in the rest of the eukaryotes, it was classified as Mammals. If it had homologues in all the species mentioned above, and also in all other vertebrates "tested" but not in the rest of the eukaryotes, it was classified as Vertebrates. Finally, if it had homologues in all 15 eukaryotes tested, it was classified as Eukarya. This classification was quite strict, and some genes with more complex conservation patterns could not be classified, but it provided robust classes of well-defined phylogenetic age. Human pseudogenes were downloaded from the psedogene.org database (Karro et al. 2007), and an R script was used to determine if there were pseudogenes overlapping the coding region of primate orphan genes. Single exon genes that overlapped TEs were eliminated to discard possible contamination of TEs incorrectly annotated as genes. The final data set contained 270 Primate genes, 364 Mammal genes, 1,958 Vertebrate genes, and 6,153 Eukarya genes.

Calculation of Nucleotide Substitution Rates

For each human and macaque orthologous protein pair, we obtained the corresponding coding sequences from Ensembl. The coding sequence alignments were based on protein alignments, which were obtained by ClustalW (Thompson et al. 1994). Nonsynonymous substitutions per nonsynonymous site (*K*a) and synonymous substitutions per synonymous site (*K*s) were estimated using the maximum likelihood method implemented in the codeml program of the PAML software package (Yang 2007). Pairs with high substitution rates ($Ka > 0.5$ and/or $Ks > 0.5$ substitutions/site) were not used in the analysis of evolutionary rates to avoid the inclusion of non bona fide orthologues. We obtained 7,203 gene pairs classified in different age groups: 5,814 Eukarya, 1,664 Vertebrates, 276 Mammals, and 120 Primates.

Codon Usage

We built a codon usage table using 35,882 nonredundant human coding sequences from Ensembl, after filtering out those sequences in which start or stop codons were missing or which had stop codons in frame. For each codon, its relative frequency of use was estimated as the number of occurrences of that codon throughout the sequence data set divided by the total number of codons. The ATG codon (Methionine) as translation initiation codon was not taken into account as this position is used to predict the start of open reading frames (ORFs). Using this codon usage table, we estimated codon usage scores for each different codon dividing its relative frequency of use by the uniform probability of occurrence (1/64), in log scale. Using these scores, for any given DNA sequence divided in codons, we estimated its capacity of coding for a protein (often referred to as coding potential or coding bias) by adding up the scores of every codon along the sequence, as previously described (Guigo 1999). To compare sequences of different length, we divided the previous score by the number of codons in the sequence obtaining a measure of average codon usage score per codon. Positive values indicate a codon usage similar to that observed in human proteins. Negative values indicate a codon usage typical of noncoding sequences.

We examined the distribution of codon usage scores in six data sets, one that contained all the human coding sequences from Ensembl (35,882), a subset of these comprising those sequences $\leq$100 amino acids (1,668), a subset containing coding sequences with evidence at protein level (20,689), a subset of these coding sequences with evidence at protein level and $\leq$100 amino acids (1,659), a library of noncoding RNA genes curated from the literature from the database RNAdb (939 genes) (Pang et al. 2007), and our data set of primate orphan coding sequences (270 genes). As a control, we also examined noncoding portions of protein-coding genes and alternative noncoding frames.

Expression and Functional Data

Tissue gene expression data were obtained from Ensembl, using the eGenetics/SANBI annotated expressed sequence tag (EST) collection. The number of annotated tissues was 70, and for each gene, we recorded the number of tissues in which there was EST-based expression evidence. We also retrieved Gene Ontology (GO) functional annotations from Ensembl for primate orphan genes and their human paralogues. We used the database Uniprot

(Bairoch et al. 2005) to retrieve further functional information and literature as well as to obtain a list of human genes with evidence at the protein level using an in-house Perl program.

Analysis of Syntenic Genomic Regions

Syntenic regions to the list of 270 primate orphan human genes were extracted for *M. musculus*, *R. norvegicus*, *C. familiaris*, and *B. taurus*, using the Galaxy server (Giardine et al. 2005). The alignments were based on build hg18 for the human genome, build mm8 for the mouse genome, build rn4 for the rat, build canFam2 for the dog genome, and build bosTau2 for the cow genome.

For each gene, we calculated the syntenic sequence coverage and percentage identity. The coverage was the fraction of the alignment that did not contain gaps. We used this measure to classify the genes in primate orphan genes showing conserved synteny if coverage was between 0.70 and 1 in at least two mammalian species and in primate orphan genes with no conserved synteny if the coverage was between 0 and 0.2 in all four mammalian species.

For genes with conserved synteny, we calculated the percentage identity of the genomic alignment corresponding to the human coding sequence and the human noncoding sequence separately. We also translated the nonprimate genomic sequences in the three possible reading frames and obtained artificial translations that were as similar as possible to the human protein sequences. In some cases, we needed to introduce frameshifts to maximize the similarity to the human protein. We used in-house Perl programs for these tasks. Protein multiple alignments were obtained by T-Coffee (Notredame et al. 2000), which was followed by manual editing if necessary. Vertebrate genomic syntenic conservation was also visually inspected using the University of California–Santa Cruz (UCSC) Genome Browser (Karolchik et al. 2008).

Sequence Similarity Searches

We extracted the genomic positions of TE sequences (comprising short interspersed element (SINE), long interspersed element (LINE), long terminal repeat [LTR], and DNA) identified by RepeatMasker and available from the UCSC Genome Browser and used these coordinates to identify all orphan genes that contained TE sequences in their coding regions.

To identify paralogues of primate orphan genes, we performed BlastP searches against all human proteins. If there were no significant hits, we additionally performed TBlastN searches against all human chromosomes to identify possibly unannotated proteins. BlastP cutoff *E* value was 0.5, which allowed detection of the paralogy between dermcidin and lacritin. We identified 66 genes with paralogues that showed significant similarity to mammalian proteins. In 12 of these genes (18%), the similarity included TE regions but was also significant when these regions were not considered. We used T-Coffee to build sequence alignments (Notredame et al. 2000) and the program "neighbor" in the PHYLIP package for distance-based tree reconstruction (Felsenstein 2005).

The best BlastP hit among nonprimate-specific paralogues was recovered (here named "parental" gene) and the positions of introns compared with those of the orphan genes. When at least one splice site was located in the same position in the alignment of the parental and orphan genes, we assumed that gene duplication had occurred by unequal crossing over (or segmental duplication, S); when none of the splice sites in the parental gene could be matched to the orphan gene, we assumed that it had occurred by retrotransposition (R). In some cases, there were no splice sites in the aligned part of the parental gene; these cases were left undetermined (S/R).

We performed TBlastN searches with primate orphan gene human proteins against GenBank EST human and mouse databases, using default parameters.

Regulatory Motif Analysis

We identified significant motifs in proximal promoter sequences (-600 to $+100$) from 14,678 human genes from Ensembl, using the program PEAKS (Bellora et al. 2007a). This algorithm identifies motifs based on their positional bias with respect to the transcription start site. We used libraries of known vertebrate transcription factor–binding sites from TRANSFAC (Matys et al. 2006) and JASPAR (Vlieghe et al. 2006). Redundant motif matches were clustered as described previously (Bellora et al. 2007b). The parameters used were window size 31 and P value $< 10^{-5}$. We obtained a list of significant motifs in the complete human gene data set, the average number of motifs in different subsets of genes, and P values using random sequences of the same composition as the test sequence (Bellora et al. 2007b).

Statistical Tests and Graphics

As evolutionary rates greatly depart from a Normal distribution, we used the nonparametric Kolmogorov–Smirnov test to detect any statistical differences between *K*a, *K*s, and *K*a/*K*s in different gene groups. In the case of codon usage scores, both the Shapiro–Wilk normality test and the one-sample Kolmogorov–Smirnov test indicated normality for all sequence data sets (P value $< 10^{-3}$), and therefore, we used a *t*-test to compare data set pairs. In order to determine if there were differences in the expression patterns between all human genes and primate orphan genes, we used a Fisher test. We used the R statistical software package (R DCT 2007) for all calculations.

Results

Features of Primate Orphan Genes

We identified all human proteins from Ensembl (Flicek et al. 2008) that showed significant sequence similarity to chimpanzee (*P. troglodytes*) and macaque (*M. mulatta*) gene products but lacked homologues in 13 other complete eukaryotic genomes, including 3 nonprimate mammalian species. We only considered human genes with putative homologues in other primates in order to increase

Table 1
Codon Usage Scores for Coding and Noncoding Human Sequence Data Sets

Data Set	Mean Codon Usage Score
Protein-coding	
Ensembl genes	0.140 ± 0.069 (35,882)
Ensembl genes with evidence at protein level	0.153 ± 0.065 (20,689)
Protein-coding <100 amino acids	
Ensembl genes	0.079 ± 0.090 (1,668)
Ensembl genes with evidence at protein level	0.081 ± 0.09 (1,659)
Protein-coding primate-specific	
Coding frame primate-specific genes	0.063 ± 0.071 (270)
Noncoding	
Noncoding frames from primate-specific genes	−0.064 ± 0.101 (540)
Noncoding regions from primate-specific genes	−0.098 ± 0.072 (921)
Noncoding frames from Ensembl genes	−0.115 ± 0.087 (71,764)
Noncoding frames Ensembl genes <100 amino acids	−0.097 ± 0.124 (3,336)
Noncoding RNA genes	−0.082 ± 0.114 (939)

NOTE.—Mean and standard deviation are indicated. Number of sequences is in brackets.

the confidence in the genes in our data set. This comprised 270 primate orphan genes (Materials and Methods, supplementary file 1, Supplementary Material online), with an estimated pseudogene inclusion rate below 5% (supplementary file 2, S1, Supplementary Material online).

All genes were annotated as protein-coding genes in Ensembl, but direct protein evidence was missing except in 6 cases, and the encoded peptides were in general short. For this reason, we analyzed their coding potential (Materials and Methods) and compared it with several coding and noncoding sequence data sets (table 1). In all coding sequence groups, including our data set of primate orphan genes, the average codon usage score was positive. In contrast, all noncoding sequence groups showed negative scores. The difference between primate orphan genes and noncoding RNA genes, or noncoding gene frames, was highly significant ($P < 10^{-5}$). Therefore, the data support the current annotation of these genes as protein-coding genes.

Orphan genes in *Drosophila* have been reported to evolve rapidly (Domazet-Loso and Tautz 2003; Zhang et al. 2007). We found that the nonsynonymous to synonymous (*K*a/*K*s) substitution rate ratio of primate orphan genes was markedly high, with a median *K*a/*K*s of 0.91 for human and macaque orthologous sequence comparisons. Similarly, high *K*a/*K*s values were observed in the subset of genes encoding experimentally verified proteins, such as the primate-specific antibacterial peptide dermcidin (*K*a/*K*s 0.89). We compared these results with groups of genes of an increasingly wider phylogenetic distribution: mammalian-specific genes (Mammals), vertebrate-specific genes (Vertebrates), and widely distributed eukaryotic genes (Eukarya) (see Materials and Methods). Consistent with previous reports (Alba and Castresana 2005, 2007; Cai et al. 2006; Luz et al. 2006; Zhang et al. 2007), we found an accelerated evolutionary rate in younger genes with respect to older ones (table 2). The *K*a and *K*a/*K*s values of human and macaque orthologous pairs were significantly different in all group-to-group comparisons ($P < 0.01$). Besides, protein size was shortest in Primate, intermediate in Vertebrates and Mammals, and longest in Eukarya.

Table 2
Features of Human Genes with Different Phylogenetic Distribution

	N	*K*a	*K*a/*K*s	Protein Length
Primates				
Mean	120	0.14	1.15	100
Median		0.11	0.91	90
SD		0.09	0.98	53
Mammals				
Mean	276	0.12	0.64	305
Median		0.05	0.53	202
SD		0.21	0.47	300
Vertebrates				
Mean	1,664	0.06	0.38	329
Median		0.03	0.32	259
SD		0.13	0.33	282
Eukarya				
Mean	5,814	0.03	0.23	624
Median		0.01	0.17	506
SD		0.06	0.24	511

NOTE.—N, number of genes; *K*a, nonsynonymous substitutions per nonsynonymous site; *K*a/*K*s, ratio nonsynonymous to synonymous substitutions; and SD, standard deviation.

Functionality of Primate Orphan Genes

The functions of orphan genes are generally poorly characterized (Daubin and Ochman 2004; Domazet-Loso et al. 2007), and most primate orphan genes in our data set were of unknown function. However, there were some exceptions. A well-characterized gene was dermcidin, encoding a peptide secreted in sweat glands with antimicrobial activity, which has also been reported to be involved in neural survival and cancer (Schittek et al. 2001; Porter et al. 2003). Another primate orphan protein with a role in immune response was minor histocompatibility protein HB-1, which is able to stimulate T-cell responses (Dolstra et al. 1999). A third example of a gene with a described function was the SPHAR gene (S-phase response), involved in the regulation of DNA synthesis (Digweed et al. 1995). Two more genes, FAM9B and FAM9C, exclusively expressed in testis, have been suggested to play roles in mediating recombination during meiosis (Martinez-Garay et al. 2002). Finally, one protein from the primate-specific morpheus gene family has been shown to locate in the nuclear pore complex (Johnson et al. 2001).

To gain further insight into the functions of primate orphan proteins, we inspected the available gene expression data in Ensembl (eGenetics/SANBI). We found that primate orphan genes were expressed in significantly less tissues than human genes in general (fig. 1, *P* value $< 10^{-5}$). In particular, the fraction of primate orphan genes expressed in only one tissue was 19% in comparison to 3.8% for all human genes. We did not find any significant tissue expression bias in tissue-specific orphan genes when compared with the complete gene data set.

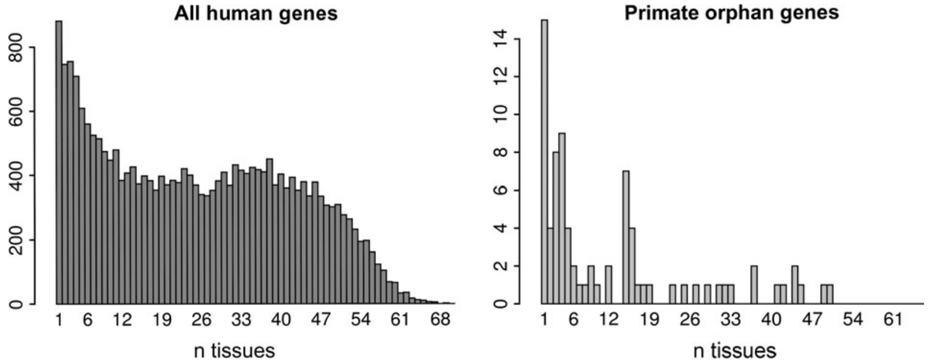


FIG. 1.—Number of tissues where a gene is expressed in the data set of all human genes and in the data set of primate orphan genes.

Gene expression is controlled to an important extent by transcription factors that bind to cis-regulatory motifs in the promoter sequence. Common motifs in human promoters include Sp1 sites, the CAAT box and the GA-binding protein GABP box, among others. In general, tissue-specific genes contain a smaller number of such motifs (Bellora et al. 2007b), so one should expect less motifs in primate orphan genes. Consistently, we found that the number of regulatory motifs in these genes was inferior to nonorphan genes, although significantly higher than in a random sequence control data set (supplementary file 2, S2, Supplementary Material online).

Mechanisms of Formation of Primate Orphan Genes

To obtain clues as to the processes underlying the formation of primate orphan genes, we performed several gene and genome comparative sequence analysis. In this section, we first describe the nature of the analysis and then provide a classification of the genes in relation to their possible mechanism of formation.

First, we investigated the similarity of the orphan genes to other human genes that were conserved in other mammals (hereafter "phylogenetically conserved genes"). This allowed us to test the hypothesis that many orphan genes could be the result of complete or partial gene duplication events (Domazet-Loso and Tautz 2003). One known example was the orphan gene dermcidin, which sits next to the lacritin gene on chromosome 12. Both genes show similar exonic structure and, although they appear to have diverged to an important extent, sequence similarity between the two is still detectable (Ma et al. 2008). Unlike dermcidin, the lacritin gene has homologues in other mammals, pointing to an earlier origin of the gene family. Using BlastP searches, we obtained a list of genes that showed complete or partial significant similarity to human phylogenetically conserved genes.

Second, we analyzed the similarity of the primate orphan genes to TE sequences (including SINE, LINE, LTR, and DNA transposons), using the genomic mapping of such elements available at the UCSC Genome Database (Karolchik et al. 2008). These searches were motivated by previous observations that many new exonic sequences in human genes are derived from TEs (Makalowski et al. 1994). We obtained a list of orphan gene coding sequences that showed some degree of overlap with TE sequences.

Third, we determined the degree of conservation of the orphan gene genomic regions in four other mammalian species (mouse, rat, dog, and cow), using prebuilt genomic alignments from the UCSC (Karolchik et al. 2008). Lack of synteny was more frequent in the human–rodent comparisons than in human–dog or human–cow comparisons (supplementary file 2, S3, Supplementary Material online), which may be due to the previously described high rate of sequence deletion in rodents, resulting in the mouse genome being about 14% smaller than the human genome (Waterston et al. 2002). The majority of genes could be classified in two well-defined groups: with no conserved synteny (105 genes), when the genes were located in genomic regions that did not align to any of the nonprimate mammalian genomes, and with conserved synteny, when they were located in regions that matched genomic sequences in at least two other mammals (56 genes). In the latter case, a similar percentage sequence identity was observed for genomic alignments corresponding to human coding and noncoding gene regions, indicating that these regions essentially lacked functional genes in the other mammalian species (supplementary file 2, S4, Supplementary Material online).

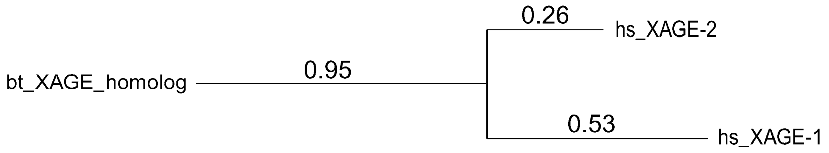
With these data, we could identify three main mechanisms associated with the formation of orphan genes in primates (table 3, supplementary file 1, Supplementary Material online).

Gene Duplication

About one-fourth of the primate orphan genes showed similarity to phylogenetically conserved human genes (66 genes), indicating that they had been formed by gene duplication. In these cases, rapid sequence divergence, often accompanied by duplication of only a part of the ancestral gene, had initially hindered the identification of homologues

Table 3
Categories of Primate Orphan Genes

Mechanism of Formation of Primate Orphan Gene	*N* Ensembl
Gene duplication	66 (24%)
Exaptation from TEs	142 (53%)
De novo formation from noncoding regions	15 (5.5%)
Unknown	47 (17%)



```
hs_XAGE-1              KSCISQTPGINLDLGSGVKVKIIPKEEHCKMPEAGEEQPQV
hs_XAGE-2              ELCQTKT-GDGCEGGTDVKGKILPKAEHFKMPEAGEGKSQV
bt_XAGE_homolog        QLAVAKT-GGEGGDGPDVREEFASNIEPVEMPEAGEGQPFA
                        : . ::*  *      *..*: :: .:  *  :******  :. .
```

FIG. 2.—Multiple alignment and evolutionary tree of XAGE sequences. Sequences were human XAGE-1 (ENSP00000364766, orphan protein), human XAGE-2 (ENSP00000286049), and cow (*Bos Taurus*, bt) homologous protein (XP_001787281). The alignment corresponds to the conserved C-terminal half of the orphan protein (GAGE domain). Estimated amino acid substitution rates are indicated in the branches; the tree was obtained by Neighbor-Joining (Jones, Taylor, and Thornton matrix).

in other mammals. For example, XAGE-1, a cancer/testis-associated gene, hit the human protein XAGE-2, another member of this family (Zendman et al. 2002). However, only XAGE-2 showed detectable similarity to XAGE homologues in other mammalian species, whereas the XAGE-1 sequence was too highly divergent (fig. 2).

Gene duplication may occur by unequal crossing over (segmental duplication) or retrotransposition. In the latter case, the new gene will initially have no introns, but there exists the possibility that, at a later stage, new introns may be inserted or new exons recruited. To estimate the relative frequency of these two mechanisms, we compared the position of splice sites in the orphan protein and its closest human relative. If at least one splice site was located in approximately the same position, we took it as evidence of segmental duplication. We estimated that 59% of the genes would have been formed by segmental duplication, 14% by retrotransposition, and the rest (27%) would be compatible with both mechanisms given the lack of splice sites in the aligned region (supplementary file 1, Supplementary Material online).

The identification of homologues with known functions may provide some hints as to the possible functions of orphan genes. Analysis of GO terms revealed that about one-third of the GO annotated paralogous genes had DNA binding–related functions. However, given the high divergence of orphan genes, it is unclear how many of these functions will be conserved.

Exaptation from TEs

Interestingly, about 70% of the genes with no similarity to phylogenetically conserved human genes matched TEs (142 genes). SINE elements (mostly Alus) comprised 93% of cases, either alone or accompanied by other TEs. These genes were mostly found in primate-specific genomic regions. TE-derived sequences corresponded to one or both exon boundaries, indicating TE exonization, or were embedded into an exon as a TE cassette (fig. 3). The genes often included non–TE-derived exons, indicating alternative mechanisms of de novo exon formation. This suggests that coding sequence exaptation from TE may trigger or contribute in a very significant manner to the formation of novel genes in primate genomes.

De Novo Formation from Noncoding Genomic Regions

The remaining genes showed no similarity to TEs or to phylogenetically conserved human genes (62 genes). Interestingly, these genes were mostly found in genomic regions with conserved synteny in other mammals. In searching for

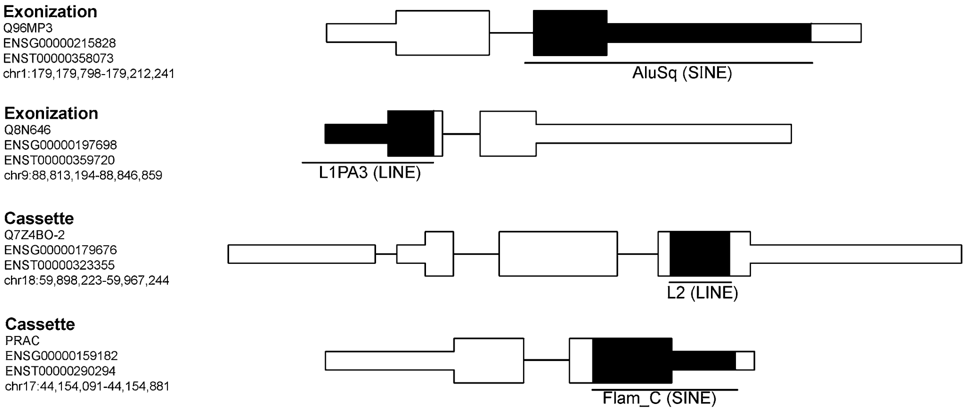
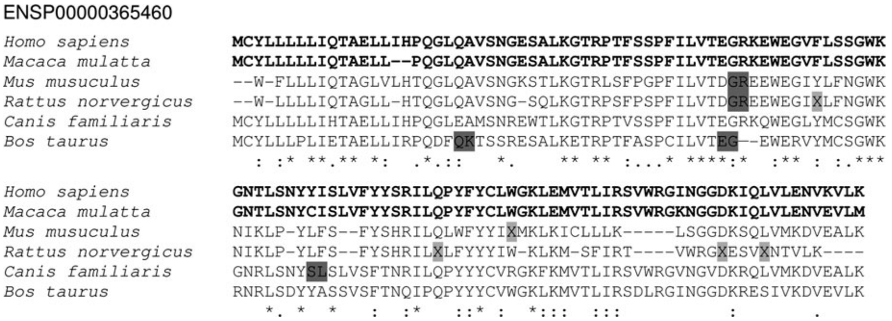


FIG. 3.—Examples of primate orphan genes containing sequences derived from TEs. The type of TE and amino acid positions at exon boundaries are indicated. Exonic sequences are indicated as boxes, which are wider for protein-coding parts; TE-derived sequences as black boxes; and introns are indicated as horizontal lines and are not at scale.



```
ENSP00000365460
Homo sapiens          MCYLLLLLIQTAELLIHPQGLQAVSNGESALKGTRPTFSSPFILVTEGRKEWEGVFLSSGWK
Macaca mulatta        MCYLLLLLIQTAELL--PQGLQAVSNGESALKGTRPTFSSPFILVTEGRKEWEGVFLSSGWK
Mus musuculus         --W-FLLLIQTAGLVLHTQGLQAVSNGKSTLKGTRLSFPGPFILVTDGREEWEGIYLFNGWK
Rattus norvergicus    --W-LLLLIQTAGLL-HTQGLQAVSNG-SQLKGTRPSFPSPFILVTDGREEWEGIXLFNGWK
Canis familiaris      MCYLLLLLIHTAELLIHPQGLEAMSNREWTLKGTRPTVSSPFILVTEGRKQWEGLYMCSGWK
Bos taurus            MCYLLLPLIETAELLIRPQDFQKTSSRESALKETRPTFASPCILVTEG--EWERVYMCSGWK
                        : :* **.** *:  .*.::   *.    ** ** :...* ****:*  :** : :.***

Homo sapiens          GNTLSNYYISLVFYYSRILQPYFYCLWGKLEMVTLIRSVWRGINGGDKIQLVLENVKVLK
Macaca mulatta        GNTLSNYCISLVFYYSRILQPYFYCLWGKLEMVTLIRSVWRGKNGGDKIQLVLENVEVLM
Mus musuculus         NIKLP-YLFS--FYSHRILQLWFYYIXMKLKICLLLK-----LSGGDKSQLVMKDVEALK
Rattus norvergicus    NIKLP-YLFS--FYSHRILXLFYYYIW-KLKM-SFIRT----VWRGXESVXNTVLK----
Canis familiaris      GNRLSNYSLSLVSFTNRILQPYYYCVRGKFKMVTLIRSVWRGVNGVDKRQLVMKDVEALK
Bos taurus            RNRLSDYYASSVSFTNQIPQPYYYCVWGKLKMVTLIRSDLRGINGGDKRESIVKDVEVLK
                         *. *  *   :  :*   ::* :  *:::  :::          :       .
```

FIG. 4.—Multiple alignment corresponding to a primate orphan protein. The alignment corresponds to Ensembl human protein sequence (ENSP00000326030, in bold), macaque protein (ENSMMUP00000001646, in bold), and artificial translations from other mammalian syntenic regions. Stop codons are X. Amino acids at frameshift boundaries are shown with a dark gray background.

clues as to their origin, we examined whether the corresponding nonprimate genomic regions had the capability to encode a similar, uninterrupted protein sequence. We identified 15 genes in which there was a strong evidence that the nonprimate sequence could not encode a similar protein due to the presence of multiple stop codons, frameshifts, and/or long gaps in several of the mammalian species (fig. 4, supplementary file 2, S5, Supplementary Material online). In addition, none of these genes showed similarity to mouse EST in the GenBank database. We also confirmed a lack of mammalian conservation of human exonic regions in the UCSC Genome Browser mammalian conservation track (supplementary file 2, S6, Supplementary Material online). These genes may thus have originated de novo from noncoding mammalian genomic regions. Interestingly, a number of them could be involved in cancer processes: PART-1 (ENSG00000152931), a gene that shows increased expression when exposed to androgens and that has been suggested to be involved in the etiology of prostate carcinogenesis (Lin et al. 2000); ENSG00000174613, an imprinted gene located in the critical region of Wilm's tumor 2 that shows reduced expression in this kind of tumor (Xin et al. 2000); and ENSG00000173046, overexpressed in colon carcinoma (Pibouin et al. 2002). The characteristics of a selected list of genes are shown in table 4.

Discussion

The formation of new genes is an important source of functional innovation, which contributes to the adaptation of the organism to new conditions of life. For example, some novel genes formed in the primate lineage play roles in the defense against pathogens (e.g., dermcidin) or may be involved in spermatogenesis (Kouprina et al. 2004). However, many of these genes lack homologues in other species and are poorly characterized. Here we have attempted to shed new light on the nature and origin of primate orphan genes.

We have determined that primate orphan genes evolve about four times faster than the average gene. Such rapid evolutionary rates may be related to relaxed functional constraints on proteins with newly acquired functions and/or to positive selection linked to adaptive evolution (Johnson et al. 2001; Levine et al. 2006). Besides, the genes generally encode short proteins, with a median length of 101 amino acids in the complete data set. Short ORFs are to be expected in

Table 4
Primate Orphan Genes Mapping to Noncoding Regions in Other Mammals

Ensembl Gene and Protein Identifiers	UniProt	Chromosomes	Protein Length	*N* Exons[a]	ESTs GenBank	eVOC
ENSG00000204537 ENSP00000365460	Q6ZTQ9	17	122	1 (1)	BX090949.1 (breast)	NA
ENSG00000204323 ENSP00000364363	Q71RC9	17	101	2 (1)	CA413024.1 (chondrosarcoma), BM930242.1 (eye), BI712423.1 (pancreas), BQ015596.1 (placenta), and DB305396.1 (brain)	NA
ENSG00000180547 ENSP00000326404	Q9P1G1	12	83	1 (1)	AI133322.1 (fetal liver)	Liver
ENSG00000174613 ENSP00000310973	KCQ1D	11	68	2 (2)	AA828167.1 (kidney tumor) and AI732937.1 (kidney tumor)	Kidney, lung, and testis
ENSG00000152931 ENSP00000354582	PART1	5	59	1 (1)	DA869513.1 (prostata), DA857587.1 (trachea), and DB206471.1 (placenta)	Broad expression
ENSG00000180441 ENSP00000317176	Q9UHU7	1	58	1 (1)	DA045082.1 (bladder), CN362190.1 (embryonic stem cells), BX451446.1 (fetal brain), BM457313.1 (testis), and BG057354.1 (lymphocyte)	Brain, blood, and testis
ENSG00000125899 ENSP00000367499	Q9UGB4	20	50	2 (2)	BX281604.1 (pooled), AW269959.1 (pooled), and AA912145.1 (pooled)	NA

NOTE.—ESTs GenBank, a maximum of five different ESTs are shown. NA: not available.

[a] Coding exons in parentheses.

genes that have arisen de novo, although this can also be a result of partial gene duplications of preexisting older genes. Interestingly, a recent estimation of the proportion of short proteins (<100 amino acids) in the mouse genome has elevated the number of such proteins from 3% to 10% of the proteome, including some peptides localized to the secretory pathway (Frith et al. 2006). So a fraction of the primate orphan proteins could be located in the membrane or in the extracellular space, which is consistent with the observation that some of them contain putative signal peptides (~18%).

It has recently been argued that most of the annotated human orphan proteins are likely to be spurious ORFs that are not functional (Clamp et al. 2007). Here we only considered human gene products that showed significant similarity to putative macaque and chimpanzee proteins and, with this data set, we reached quite different conclusions regarding the possible functionality of orphan genes. First, codon usage values were largely consistent with the current annotation of these genes as protein-coding genes. Second, the expression of many of these genes was well supported in EST and/or cDNA libraries. Third, about one-fourth of them showed sequence similarity to protein families that included nonprimate homologues. Fourth, although the number of experimentally verified proteins in the data set was small, these proteins showed similar characteristics to the rest, including high substitution rates between human and macaque, indicating that high divergence is not necessarily linked to lack of protein functionality.

In *Drosophila*, lineage-specific genes are often expressed in testis (Levine et al. 2006; Zhou et al. 2008), and the same has been observed for recent primate retrogenes (Marques et al. 2005). In contrast, primate orphan genes showed high tissue specificity but not testis-specific enrichment when compared with all human tissue–specific genes. One interesting question is how newly formed genes acquire functional promoter sequences, particularly in the case of retrogene copies or genes originated de novo. Previous studies suggest that widely expressed genes require a larger number of regulatory motifs for basic transcription factors (Sp1, GABP/ETS, CAAT-BP, etc.) than tissue-specific genes (Bellora et al. 2007b). We found that, although the number of motifs in orphan genes was significantly higher than the number expected in random sequences of similar compositional and CpG content, it was, in general, smaller than in genes of deeper phylogenetic conservation. So, the reduced number of functional regulatory motifs in new promoter sequences may explain why so many orphan genes are expressed in a tissue-specific manner.

An important fraction of the primate orphan genes showed significant sequence similarity to human genes that in turn had homologues in other mammalian species. This is consistent with the previously proposed mechanism of gene duplication followed by rapid sequence divergence as an explanation for the high number of orphan genes detected in *Drosophila* (Domazet-Loso and Tautz 2003). Gene duplication has a prominent role in the creation of novel functional lineage-specific genes (Ohno 1970; Long et al. 2003), either by unequal crossing over or by retrotransposition. The latter has been shown to be a very active process for the formation of *Drosophila* and human lineage–specific genes (Marques et al. 2005; Vinckenbosch et al. 2006; Bai et al. 2007). None of the genes in our list matched the list of recently originated functional human gene retrocopies provided by Marques et al. (2005) (supplementary table S1, Supplementary Material online), probably because our study only comprised orphan genes, but, in this context, our results also indicated that retrotransposition is a relevant process for the formation of rapidly evolving, lineage-specific gene copies.

One striking result was that a very large fraction of the primate orphan genes showing no homology to human conserved genes contained TE-like sequences. Overall, genes containing TE represented 53% of the orphan primate genes, far greater than the estimated 4% of human genes containing these elements (Nekrutenko and Li 2001). This result is not so surprising if we consider that TEs, specially Alus, are a major substrate for exon formation in primates (Krull et al. 2005; Corvelo and Eyras 2008). This has been attributed to the fact that they contain motifs that can become functional splice sites via specific mutations, allowing the exonization of part of the element (Gal-Mark et al. 2008). Our results strongly suggest that TEs may also promote the formation of completely new genes, which is reinforced by the observation that TE-containing primate orphan genes were essentially located in primate-specific genomic regions. Two mouse genes, lungerkine and mNSC1, have previously been suggested to originate from TEs as they lack orthologues in human and rat and a large part of their coding sequence is composed of rodent TEs (Nekrutenko and Li 2001). The insertion of TE-derived exons in coding sequences can generate protein functional variants (Gerber et al. 1997). TE-containing orphan genes were very poorly annotated at the functional level, although many had putative macaque orthologues. However, only about 30% of them corresponded to entries in Uniprot, whereas the general figure for primate orphan genes was 52%. Future studies will be required to better assess how many of these genes encode functional proteins and to fully understand their relevance in the generation of evolutionary gene novelty in the primate lineage.

We obtained evidence that about 5.5% of primate orphan genes could have originated de novo from noncoding genomic regions. The syntenic regions in other mammalian species lacked the potential to code for similar genes, which indicates that these genes may only have become functional in the primate lineage, although independent pseudogeneization in different mammalian lineages cannot be completely ruled out. To our knowledge, this is the first time that this mechanism is proposed to have made a significant contribution to the formation of novel genes in mammals, although it has been previously proposed to explain the formation of several novel *Drosophila* (Begun et al. 2006; Levine et al. 2006; Zhou et al. 2008) and *S. cerevisiae* genes (Cai et al. 2008).

Independently of the mechanism by which the gene has originated, many of the new proteins are markedly short; in addition, there is a direct relationship between the length of the gene and its age (table 1) (Alba and Castresana 2005; Choi and Kim 2006). It has previously been proposed that proteins tend to become longer and to evolve toward complex α/β structures (Choi and Kim 2006), as they become older. The acquisition of new domains, for example, by exon shuffling (Long et al. 2003), exonization (Sorek 2007), or expansion of short repetitive elements (Mularoni et al. 2007), may result in protein size increase in a time-dependent fashion. Concomitantly, functional constraints may also become stronger with time, which would be reflected in the lowest evolutionary rates being observed in the most ancient proteins (Alba and Castresana 2005).

Important advances have been made in deciphering the human transcriptome, through the development of new high-throughput technologies such as tiling microarrays or large-scale determination of transcript ends (Birney et al. 2007; Kapranov et al. 2007). These studies have revealed that a much larger fraction of the genome than previously thought is found in primary transcripts, potentially increasing the opportunities for new gene functions to arise. Short ORFs present in such transcripts could occasionally be translated into new peptides, which would then be tested by natural selection. If advantageous, the new function would be retained and continue to evolve. We hope our results will encourage further studies on the evolutionary and functional implications of newly formed genes.

Supplementary Material

Supplementary files 1 and 2 and table S1 are available at *Molecular Biology Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Mireya Plass, André Corvelo, Francesc Castro, and Steve Laurie for technical advice and useful comments on the manuscript. We are also grateful to Jose Castresana for fruitful discussions during this work. We received financial support from Ministerio de Educación e Innovación (FPU to M.T.-R., Plan Nacional BFU2006-07120, Ramón y Cajal Programme to R.C.), Ministerio de Sanidad y Consumo (REEM to N.B.), Generalitat de Catalunya ("Grups de Qualitat" to X.E.), and Fundació ICREA (M.M.A.).

Literature Cited

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol. 22:598–606.

Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 7:53.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bai Y, Casola C, Feschotte C, Betran E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol. 8:R11.

Bairoch A, Apweiler R, Wu CH, et al. (15 co-authors). 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res. 33:D154–D159.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. Genetics. 172:1675–1681.

Bellora N, Farre D, Alba MM. 2007a. PEAKS: identification of regulatory motifs by their position in DNA sequences. Bioinformatics. 23:243–244.

Bellora N, Farre D, Alba MM. 2007b. Positional bias of general and tissue-specific regulatory motifs in mouse promoters. BMC Genomics. 8:459.

Birney E, Stamatoyannopoulos JA, Dutta A, et al. (318 co-authors). 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447:799–816.

Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY. 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. J Mol Evol. 63:1–11.

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

Castresana J, Guigo R, Alba MM. 2004. Clustering of genes coding for DNA binding proteins in a region of atypical evolution of the human genome. J Mol Evol. 59:72–79.

Choi IG, Kim SH. 2006. Evolution of protein structural classes and protein sequence families. Proc Natl Acad Sci USA. 103:14056–14061.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA. 104:19428–19433.

Corvelo A, Eyras E. 2008. Exon creation and establishment in human genes. Genome Biol. 9:R141.

Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res. 14: 1036–1042.

Digweed M, Gunthert U, Schneider R, Seyschab H, Friedl R, Sperling K. 1995. Irreversible repression of DNA synthesis in Fanconi anemia cells is alleviated by the product of a novel cyclin-related gene. Mol Cell Biol. 15:305–314.

Dolstra H, Fredrix H, Maas F, Coulie PG, Brasseur F, Mensink E, Adema GJ, de Witte TM, Figdor CG, van de Wiel-van Kemenade E. 1999. A human minor histocompatibility antigen specific for B cell acute lymphoblastic leukemia. J Exp Med. 189:301–308.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 13:2213–2219.

Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.

Eichler EE, Hoffman SM, Adamson AA, Gordon LA, McCready P, Lamerdin JE, Mohrenweiser HW. 1998. Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. Genome Res. 8: 791–808.

Felsenstein J. 2005. PHYLIP (phylogeny inference package). In: 3.6 edn.

Fischer D, Eisenberg D. 1999. Finding families for genomic ORFans. Bioinformatics. 15:759–762.

Flicek P, Aken BL, Beal K, et al. (59 co-authors). 2008. Ensembl 2008. Nucleic Acids Res. 36:D707–D714.

Fortna A, Kim Y, MacLaren E, et al. (16 co-authors). 2004. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. 2:E207.

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM. 2006. The abundance of short proteins in the mammalian proteome. PLoS Genet. 2:e52.

Gal-Mark N, Schwartz S, Ast G. 2008. Alternative splicing of Alu exons—two arms are better than one. Nucleic Acids Res. 36:2012–2023.

Gerber A, O'Connell MA, Keller W. 1997. Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. RNA. 3:453–463.

Giardine B, Riemer C, Hardison RC, et al. (13 co-authors). 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15:1451–1455.

Goldovsky L, Janssen P, Ahren D, et al. (13 co-authors). 2005. CoGenT++: an extensive and extensible data environment for computational genomics. Bioinformatics. 21:3806–3810.

Guigo R. 1999. DNA composition, codon usage and exon prediction. In: Bishop M, editor. Genetics databases. Oxford (UK): Academic Press.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature. 413:514–519.

Kapranov P, Cheng J, Dike S, et al. (22 co-authors). 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 316:1484–1488.

Karolchik D, Kuhn RM, Baertsch R, et al. (25 co-authors). 2008. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res. 36:D773–D779.

Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 35:D55–D60.

Kouprina N, Mullokandov M, Rogozin IB, Collins NK, Solomon G, Otstot J, Risinger JI, Koonin EV, Barrett JC, Larionov V. 2004. The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. Proc Natl Acad Sci USA. 101:3077–3082.

Krull M, Brosius J, Schmitz J. 2005. Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol. 22:1702–1711.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci USA. 103:9935–9939.

Lin B, White JT, Ferguson C, et al. (10 co-authors). 2000. PART-1: A noval human prostate-specific, androgen-regulated gene that maps to chromosome 5q12. Cancer Res. 60:858–863.

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

Luz H, Staub E, Vingron M. 2006. About the interrelation of evolutionary rate and protein age. Genome Inform. 17:240–250.

Ma P, Wang N, McKown RL, Raab RW, Laurie GW. 2008. Focus on molecules: lacritin. Exp Eye Res. 86:457–458.

Makalowski W, Mitchell GA, Labuda D. 1994. Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10:188–193.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3.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. 80:259–267.

Matys V, Kel-Margoulis OV, Fricke E, et al. (16 co-authors). 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34:D108–D110.

Mularoni L, Veitia RA, Alba MM. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. Genomics. 89:316–325.

Nekrutenko A, Li WH. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 17:619–621.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

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

Ohno S. 1984. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci USA. 81:2421–2425.

Okamura K, Feuk L, Marques-Bonet T, Navarro A, Scherer SW. 2006. Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics. 88:690–697.

Pang KC, Stephen S, Dinger ME, Engstrom PG, Lenhard B, Mattick JS. 2007. RNAdb 2.0—an expanded database of mammalian non-coding RNAs. Nucleic Acids Res. 35:D178–D182.

Pibouin L, Villaudy J, Ferbus D, Muleris M, Prosperi MT, Remvikos Y, Goubin G. 2002. Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. Cancer Genet Cytogenet. 133:55–60.

Porter D, Weremowicz S, Chin K, et al. (19 co-authors). 2003. A neural survival factor is a candidate oncogene in breast cancer. Proc Natl Acad Sci USA. 100:10931–10936.

R DCT. 2007. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Schittek B, Hipfel R, Sauer B, et al. (12 co-authors). 2001. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. Nat Immunol. 2:1133–1137.

Siew N, Fischer D. 2003. Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins. 53:241–251.

Sorek R. 2007. The birth of new exons: mechanisms and evolutionary consequences. RNA. 13:1603–1608.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A. 103:3220–3225.

Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res. 34:D95–D97.

Waterston RH, Lindblad-Toh K, Birney E, et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Wootton JC, Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266:554–571.

Xin Z, Soejima H, Higashimoto K, et al. (12 co-authors). 2000. A novel imprinted gene, KCNQ1DN, within the WT2 critical region of human chromosome 11p15.5 and its reduced expression in Wilms' tumors. J Biochem. 128:847–853.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Zendman AJ, Van Kraats AA, Weidle UH, Ruiter DJ, Van Muijen GN. 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. 99:361–369.

Zhang G, Wang H, Shi J, Wang X, Zheng H, Wong GK, Clark T, Wang W, Wang J, Kang L. 2007. Identification and characterization of insect-specific proteins by genome data analysis. BMC Genomics. 8:93.

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. 18:1446–1455.

Diethard Tautz, Associate Editor

Accepted December 3, 2008