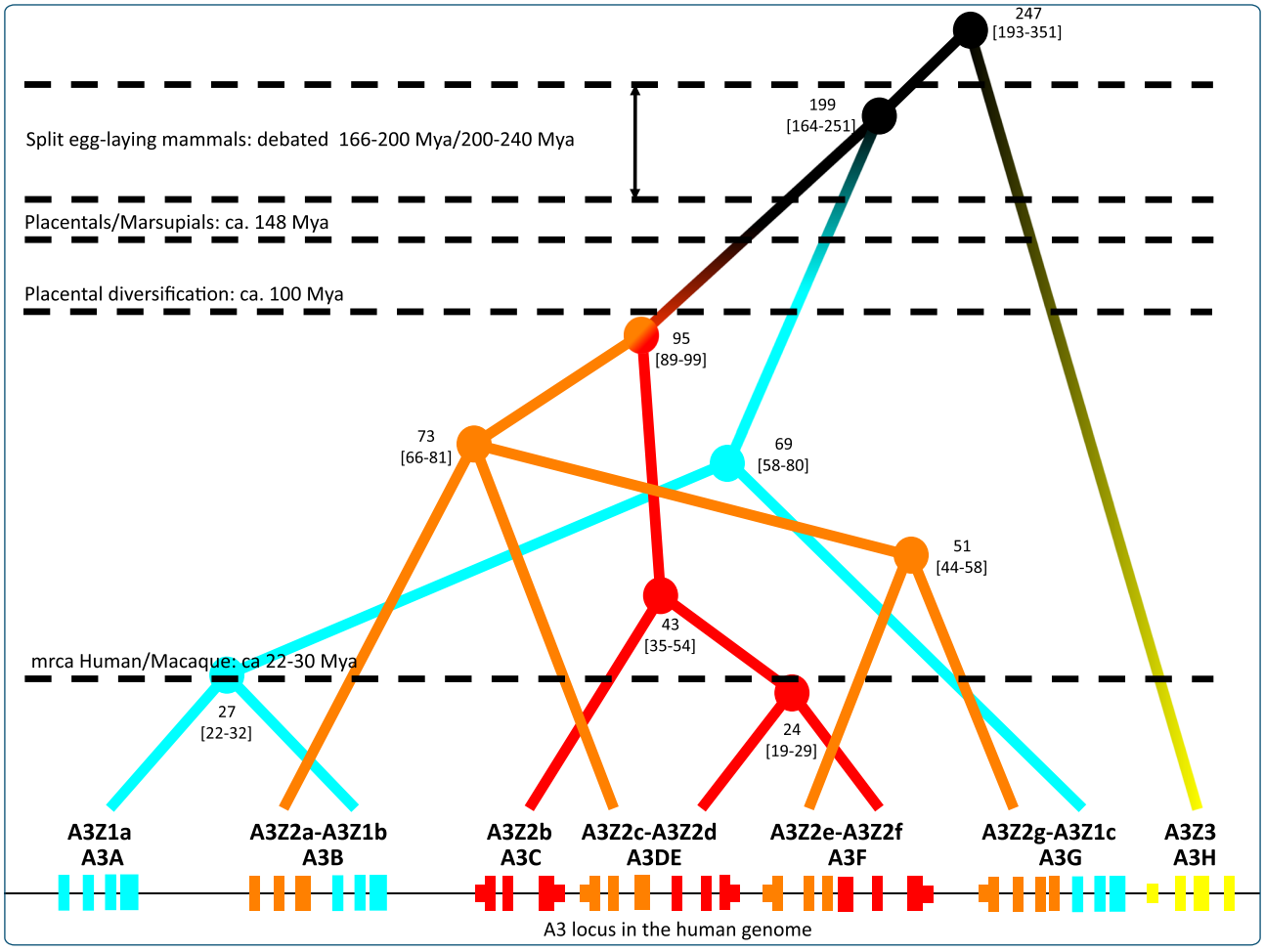
BMC
Evolutionary Biology



An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals

Münk *et al.*

BioMed Central

Münk *et al. BMC Evolutionary Biology* 2012, **12**:71
http://www.biomedcentral.com/1471-2148/12/71

Münk et al. BMC Evolutionary Biology 2012, 12:71
http://www.biomedcentral.com/1471-2148/12/71

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals

Carsten Münk[1], Anouk Willemsen[2] and Ignacio G Bravo[2,3,4*]

Abstract

Background: The APOBEC3 (A3) genes play a key role in innate antiviral defense in mammals by introducing directed mutations in the DNA. The human genome encodes for seven A3 genes, with multiple splice alternatives. Different A3 proteins display different substrate specificity, but the very basic question on how discerning self from non-self still remains unresolved. Further, the expression of A3 activity/ies shapes the way both viral and host genomes evolve.

Results: We present here a detailed temporal analysis of the origin and expansion of the A3 repertoire in mammals. Our data support an evolutionary scenario where the genome of the mammalian ancestor encoded for at least one ancestral A3 gene, and where the genome of the ancestor of placental mammals (and possibly of the ancestor of all mammals) already encoded for an A3Z1-A3Z2-A3Z3 arrangement. Duplication events of the A3 genes have occurred independently in different lineages: humans, cats and horses. In all of them, gene duplication has resulted in changes in enzyme activity and/or substrate specificity, in a paradigmatic example of convergent adaptive evolution at the genomic level. Finally, our results show that evolutionary rates for the three A3Z1, A3Z2 and A3Z3 motifs have significantly decreased in the last 100 Mya. The analysis constitutes a textbook example of the evolution of a gene locus by duplication and sub/neofunctionalization in the context of virus-host arms race.

Conclusions: Our results provide a time framework for identifying ancestral and derived genomic arrangements in the APOBEC loci, and to date the expansion of this gene family for different lineages through time, as a response to changes in viral/retroviral/retrotransposon pressure.

Keywords: APOBEC, Cytydine deaminase, Gene duplication, Subfunctionalisation, Virus/host arms race

Background

Cytidine deaminases of the APOBEC3 (A3) gene family have a broad antiviral activity against retroviruses, can inhibit LTR- and non-LTR-retrotransposons, parvoviruses, hepadnaviruses, flaviviruses and paramyxoviruses, and might repress also TT-viruses, papillomaviruses and herpesviruses [1-9]. Under the already tested additional viruses, adenoviruses, poxviruses and influenza viruses replicate well irrespective of A3s [3,10,11]. The hunt is on for the identification of additional viral targets [12]. The APOBEC3 gene family encodes a characteristic zinc (Z)-coordinating catalytic motif (His-X-Glu-X_{23-28}-Pro-Cys-X_{2-4}-Cys) [13] and the A3 proteins can be classified according to the presence of an A3Z1, A3Z2 or A3Z3 motif [14-16]. A3s act by deaminating cytidine into uridine (EC 3.5.4.5) using single-stranded DNA as a substrate. This DNA editing results in the introduction of mutations that eventually render the target genome inactive. However, uncontrolled chemical editing of DNA sequences puts at risk the genomic information in the cell [17,18] and many mechanistic questions regarding A3 activity and substrate specificity remain open.

During retroviral budding, A3 molecules incorporated into progeny retroviral particles are carried over within the virion and counteract the new infection upon release in the cytoplasm of the newly infected cell. Retroviruses

* Correspondence: igbravo@iconcologia.net
[2]Genomics and Health, Centre for Public Health Research (CSISP), Valencia, Spain
[3]Infections and Cancer, Catalan Institute of Oncology (ICO) | Bellvitge Institute of Biomedical Research (IDIBELL), Barcelona, Spain
Full list of author information is available at the end of the article

BioMed Central
© 2012 Münk et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

have evolved mechanisms to prevent encapsidation of A3s into viral particles. The Vif protein in lentiviruses, the Bet protein in foamyviruses and the nucleocapsid protein in *Human T-cell lymphotrophic virus* accomplish this anti-antiviral activity [19-24]. The expression of A3s is not restricted to the immune system [25], and different cell types may express different A3 repertoires, and even different variants may present different subcellular location [26]. Thus, some retroviruses without *vif* or *bet* genes might escape A3-mediated antiviral inhibition by a restriction of their cellular tropisms, as it is discussed for the *Equine infectious anemia virus* [27]. Several aspects on how retroviruses cope with the deaminase activity of the A3s encoded by their host species are still a matter of debate [28,29]. Thus, and despite several studies, it remains unknown which strategies the *Moloney murine leukemia virus* or the *Mouse mammary tumor virus* have evolved to prevent inhibition by the murine A3 [30-36].

In mammals the A3 locus appears always flanked by the CBX6 and CBX7 genes but there is ample variation across species, regarding the number and arrangement of A3 individual genes, presence of fused genes, expression pattern, splice alternatives and read-through mechanisms and substrate specificity, even in related species. Thus, although Primates and Rodents are relatives and belong together within Euarchontoglires, the human genome contains seven A3 genes, four of them resulting from the fusion of two A3 domains, while the mouse genome encodes for a single A3Z2-A3Z3 fused gene [37]. Additionally the dog genome presents two A3 genes, while there are four A3 genes in cats, two to three in pigs, sheep and cow and six in horses [15,16,27], and all these species belong together within Laurasiatheria. The finding of A3Z1, A3Z2, and A3Z3 genes in these two mammalian lineages strongly indicates that their ancestor (Boreoeutheria) was probably equipped with a single copy of each gene.

One important mechanism of genome evolution and for the appearance of novel gene functions is gene duplication [38]. Genes within a genome that are descendants of gene duplications are paralogs, while two genes in different species that derive from a single gene in the last common ancestor of both species are orthologs. Paralogs derive either from an ancestral duplication ("outparalogs") or they derive from a lineage-specific duplication ("inparalogs"), giving rise to co-orthologous relationships [39]. Several models for the emergence, maintenance and evolution of gene copies have been proposed (for a review see [40]). The duplication of genes may in some cases have no immediate consequence for the host, but in other cases can be deleterious or linked to disease [41], or confer an selective advantage [42]. For the antiviral A3 genes evolutionary solutions reflect the trade off between a potential self-toxicity against cellular DNA -in cases of an A3 exacerbated response- and the emergence of viral pathogens if low A3 activity and/or diversity allow for restriction escape variants. Fixation of duplicated A3 genes and the subsequent preservation in certain population is likely driven by a strong selective advantage for the individuals carrying additional copies of the gene/s. Any subsequent acquisition of genetic differences between the gene copies can alter the chances of both copies being preserved and might change the function of the encoded proteins and result either in gene loss, 'neo-functionalization' [38] or 'sub-functionalization' [43].

We present here a time scale of the evolution of the A3 loci, arising from their common origin with other cytidine deaminases. According to our results, the track of duplications in the A3 locus started with the ancestral gene itself. The appearance of the three clades within the A3 family dates back to emergence of placental mammals. The evolutionary history of A3s has since then been landmarked by duplication events especially in Primates, but also in Perissodactyla and Carnivora, as well as by deletions events, such as in Rodentia. Three main trends can be observed: first, a sustained decrease in evolutionary rate for the A3 subfamilies in the last 100 Mya; second, duplication events have occurred in the A3Z1 and A3Z2 subfamilies, but not in the A3Z3; third, duplication events accumulate in the last 50 Mya. The molecular mechanisms that generate the tandem duplication of A3, and the evolutionary pressures that drive the sub/neofunctionalisation and eventually selection of the duplicated genes still need to be identified.

Results

There are three main clades of A3 genes

We have found sequences corresponding to A3 genes in extant members of Laurasiatheria (seven Carnivora, seven Cetartiodactyla, one Perissodactyla and one Chiroptera species), Euarchontoglires (three Rodentia and seventeen Primates species) and Afrotheria (one Hyracoidea and one Proboscidea species). However, BLAST and PSI-BLAST searches of genomic, EST and other sequence databases failed to find A3 in any other taxon within other placental mammals (e.g. Xenarthra) marsupials, monotremes, or in non-mammals. A selection of 34 AICDA sequences (23 Eutheria species, three Metatheria, one Prototheria, two Aves, two Amphibia and three Actinopterygii species) was identified as sister taxa to all A3s, while a selection of 24 A1 sequences (21 Eutheria species, two Metatheria and one Aves species) was identified as outgroup. An exhaustive list of species and accession numbers is given in Additional file 1: Table S1. Using this dataset, we performed first phylogenetic reconstruction using both maximum likelihood and Bayesian inference, identified local topologies suitable for molecular dating, and then performed time inference introducing this information as temporal constrains using a relaxed clock approach. The

global results are shown in Figure 1. More detailed results are shown in Additional file 2: Figure S1 which shows the bootstrap values of the ML analysis and Additional file 3: Figure S2 which contains the 95% highest posterior density (HPD) of the node ages.

The most recent common ancestor (MRCA) of the A1s, A3s and AICDA sequences can be dated 441 Mya (95% HPD, 431–456 Mya). The MRCA of the A3s and AICDA sequences can be dated 424 Mya (95% HPD, 417–438 Mya), while the MRCA of extant A3 sequences can be dated 247 Mya (95% HPD 193–351 Mya) (Figure 1). The three members of the A3 family radiated later at comparable time points: 112 Mya (95% HPD 109–113 Mya) for A3Z1; 111 Mya (95% HPD 106–113 Mya) for A3Z2; and 109 Mya (95% HPD 102–113 Mya) for A3Z3. The relationships among the three A3 subfamilies could not be solved with certainty, but both ML (44% bootstrap support) and Bayesian inference (0.52 Bayesian posterior probability) suggested a sisterhood relationship between A3Z1 and A3Z2, with a MRCA dated 199 Mya (95% HPD 164–251 Mya). Only more quality data from Afrotheria, or the discovery of A3 genes in Monotremes, Marsupials and/or Xenarthra may help us define and date these basal nodes with confidence.

Evolution of the A3 loci in Euarchontoglires

Rodents and Primates are the two main orders in Euarchontoglires. We could identify different A3 genes in the genomes of platyrrhines (New World monkeys) and catarrhines (Old World monkeys and apes). In modern humans eleven A3 open reading frames exist, forming seven genes encoding a single Z domain or a Z-Z domain, either fused A3Z2-A3Z2 or A3Z2-A3Z1 (Figure 2). This is by far the most complex known organisation of the A3 loci. We

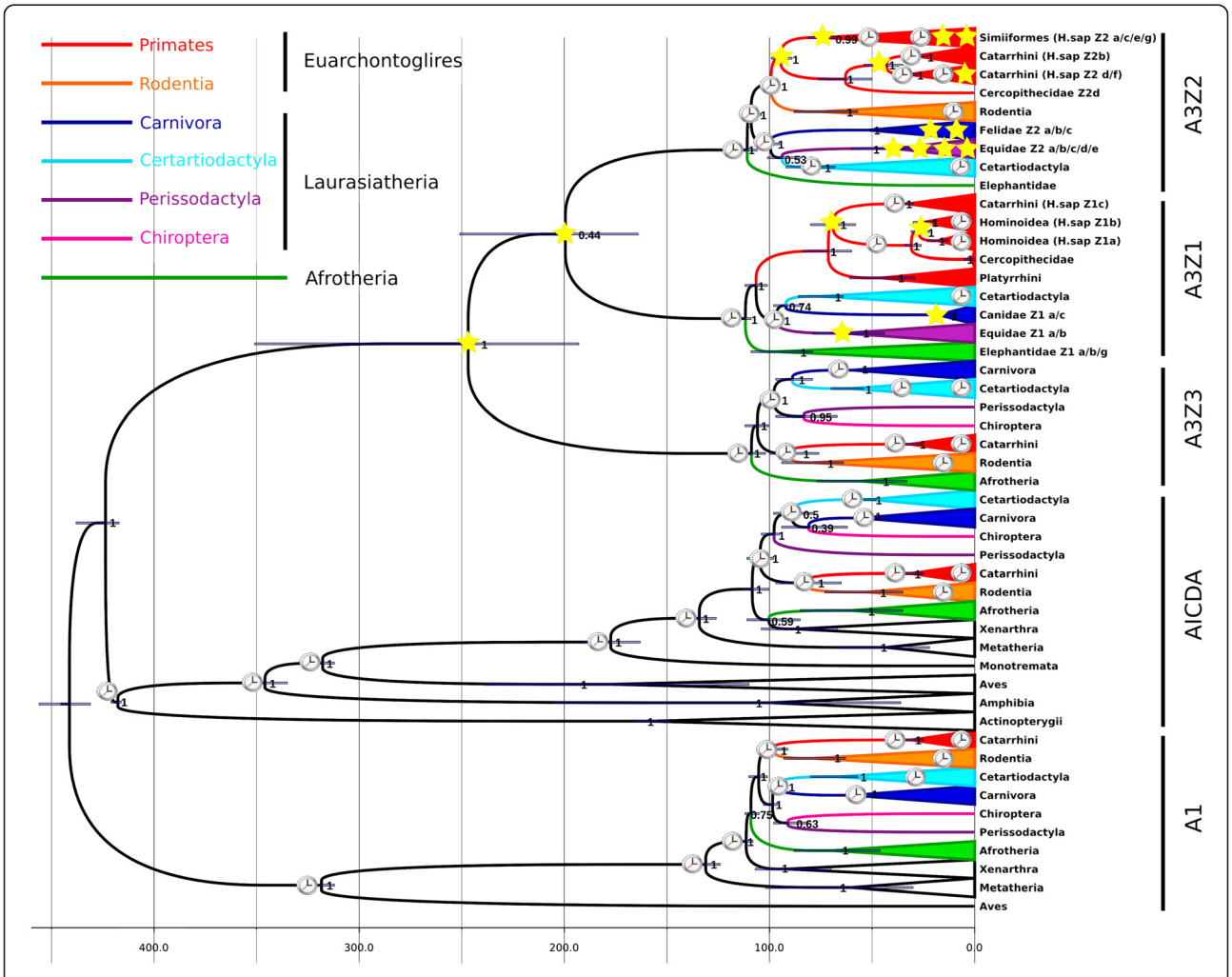


Figure 1 Dated Bayesian phylogenetic tree for the A3 family using A1 sequences as outgroup. Scale bar given in Mya. Values at the nodes correspond to ML bootstrap support. Error bars encompass 95% HPD for the age of the nodes. Star symbols indicate duplication events in the corresponding lineage and moment in time. Clock cartoons labels the nodes used for calibration. Mammalian clades are labelled using colours (code in the inset). No sequences from Xenarthra could be found in our searches.

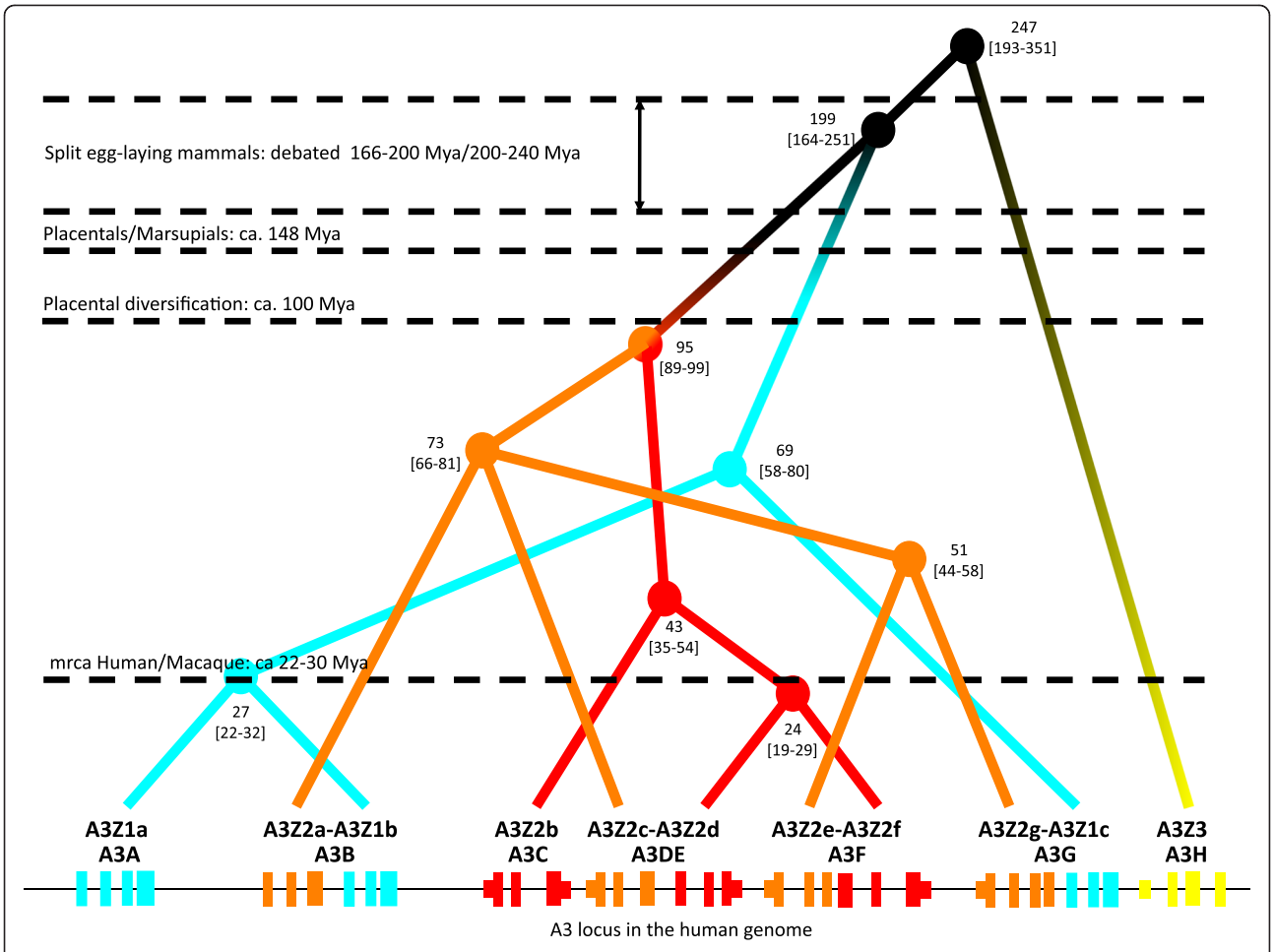


Figure 2 Projection of the A3 phylogenetic relationships onto the human genomic locus. Exon composition and gene arrangement of the seven A3 genes are displayed. Labels indicate median age [95% HPD] of the corresponding nodes. Several milestones in the evolution of mammals are indicated.

must however keep in mind that our knowledge on the structure of the A3 loci in other primate genomes, especially regarding transcriptomic data, is still fragmentary.

A3Z1 genes in primates

The A3Z1 gene appears in humans in three copies, named Z1a, Z1b and Z1c, and we will name the corresponding orthologs according to the nomenclature in humans (Figure 2). The MRCA of A3Z1 genes can be dated 71 Mya (95% HPD 60–84), which grossly corresponds to the basal divergence time of primates, some 73–87 Mya. Orthologs of the A3Z1 genes can be found in the genomes of several Haplorrhini species, including Catarrhini (Old World monkeys and apes) and Platyrrhini (New World monkeys). We could only detect one A3Z1 gene in the platyrrhines genomes we have browsed, and these sequences are the outgroup of the Z1a, Z1b and Z1c found in catarrhines. Phylogenetic relationships and inferred divergence times among the Z1c orthologs grossly correspond to those among the corresponding species. Regarding the Z1a and Z1b genes, we could identify the orthologs of both human genes in the genomes of the common chimpanzee, western gorilla, northern white-cheeked gibbon and Rhesus macaque. The ortholog of the Z1b human gene could be identified in the Sumatran orangutang. The duplication event that generated the Z1a and Z1b genes can be dated some 27 Mya (95% HPD 22–32). This timing could be compatible with the Hylobatidae/Hominidae split some 18–24 Mya, although it matches well the Cercopithecidae/Hominoidea split some 26–34 Mya [44].

A plausible scenario for the evolution of the A3Z1 loci in primates would include a basal split, generating the ancestors of the Z1a/Z1b (hereafter Z1ab) and the Z1c genes. Whether this first duplication event predates the early split Haplorrhini (tarsiers, monkeys and apes)/Strepsirrhini (non-tarsier prosimians) or is rather exclusive to the Haplorrhini lineage is not clear and must still be solved, since there are no sequence data available for Strepsirrhini. In Catarrhini, the ancestral Z1c gene may

have evolved without further duplication event. The ancestral Z1ab gene underwent duplication and generated the ancestral Z1a and Z1b genes. Orthologs of both Z1a and Z1b are present in all hominids [45], as the MRCA to both genes predates well the split Ponginae/Homininae. Future identification of Z1ab genes in hylobatids and in cercopithecoids will clear whether the Z1a/Z1b duplication event is basal to catarrhini or to Hominoidea.

A3Z2 genes in primates

Our fragmentary information on the A3 gene content makes it impossible to reconstruct with confidence the evolutionary relationships among the A3Z2 genes in primates. Additionally, a number of A3Z2 sequences arise from genomic material while others belong to cDNA, and the complex alternative splicing of the A3 genes may disturb further the algorithms for phylogenetic inference. Finally, in this locus gene conversion may have played a key role, given the large number of gene duplication events in this A3 subfamilly -up to six in humans- and the nature of the locus, where the copies generated remain in a tandem arrangement. A proper phylogenetic reconstruction will thence need to wait until genomic sequences with better coverage are available for a larger number of primate species, especially if encompassing Strepsirrhini and Tarsiidae. A basal split event some 94 Mya (95% HPD 89–99) generated the two main lineages Z2bdf and Z2aceg. This timing could match the basal strepsirrhini/haplorrhini diversification time in primates (some 87 Mya), although it would be more compatible with one duplication exclusive to haplorrhini (MRCA some 85 Mya) or even to simiiformes. In either case our interpretation implies that the ancestral simiiformes already carried two copies of the A3Z2 genes in their genomes. The Z2bdf group comprises exclusively sequences from catarrhines, with an estimated MRCA around 43 Mya (95% HPD 35–54). Sequences here cluster into two groups: Z2b and Z2df. Both clusters contain sequences from cercopithecoids and from hominoids. It can thus be inferred that the duplication event that generated both groups took place before the divergence between cercopithecoids and hominoids some 29 Mya. This timing is compatible with the estimated MRCAs around 28 Mya (95% HPD 23–34) for Z2b and 32 Mya (95% HPD 28–34) for Z2df. Within the Z2aceg group, phylogenetic inference cannot solve the fine relationships. Only a major, monophyletic group containing the Z2g genes can be clearly defined, containing sequences of both catarrhines and platyrrhines. Phylogenetic relationships among Z2g genes grossly correspond to those of the corresponding species: platyrrhines are basal to the cluster, with calculated split time 39 Mya (95% HPD 34–45), and cercopithecoids and hominids cluster separately, with a MRCA around 24 Mya (95% HPD 23–28).

A plausible scenario for the evolution of the A3Z2 genes in primates would include a first basal split Z2aceg/Z2bdf, previous to the split between platyrrhini and catarrhini. The ancestral Z2aceg underwent a second duplication Z2ace/Z2g also before this split event. Modern Z2a, Z2c and Z2e human genes appeared after subsequent duplication events. On the other hand, the ancestral Z2bdf underwent a first duplication Z2b/Z2df at least before the divergence between cercopithecoids and hominoids. The basal position of human Z2d with respect to Z2f sequences suggests also that the duplication event Z2d/Z2f predates diversification within catarrhines, but a larger repertoire of sequences is needed here to reconstruct phylogenetic relationships with confidence.

A3Z3 genes in primates

Orthologs of the A3Z3 gene can be found in the genomes of the catarrhines chimpanzee, bonobo, human, gorilla, Bornean and Sumatran orangutan, gibbon and macaque (Figure 1). The MRCA of these A3Z3 genes dates back to 31 Mya (95% HPD 26–34). In all these species the A3Z3 gene appears as a single copy, with no evidences of gene duplication. Additionally, the phylogenetic relationships among these genes and the inferred timing for the nodes perfectly match those of the corresponding species. The most parsimonious explanation for the absence of A3Z3 in platyrrhines and in other primates is therefore our lack of information about the locus, and it can be anticipated that the missing A3Z3 genes should be identified in the future, as genome coverage increases.

A3 genes in rodents

In the rodent *Cavia porcellus*, as in rat and mouse, the Z2Z3 genes are fused and the A3Z1 gene is missing. Considering the presence of A3Z1 in Primates, the last common ancestor of Muridae and Caviidae probably possessed already 50–85 Mya the genome organization found in the extant rat and mouse genomes. Precise answers on the timing for the A3Z1 deletion event and for the Z2Z3 fusion event will need to wait until new genomic information is available. Interesting sequences could come from the genomes of squirrels, which are basal to Muridae/Caviidae, or from rabbits, which constitute together with rodents the Glires taxon, sister to Primates.

Evolution of the A3 loci in Laurasiatheria

A3 genes in cetartiodactyla

In the group of Cetartiodactyla (even-toed ungulates) the evolution of the individual A3 genes matches the evolution of the host genomes (Figure 1), although there is controversy about the phylogenetic relationships among mammalian orders here [46]. The genomes of cow and sheep present single copies of the A3Z1, A3Z2

and A3Z3 genes, with no evidences for gene duplications. In the A3 locus in the pig, the A3Z1 gene however seems to have been lost during evolution [15,47]. The presence of a A3Z1 gene in the genomes of cow and sheep suggests that the loss of the A3Z1 gene in pig genome occurred after the split that generated the Suidae lineage.

A3 genes in horses

Modern horses have six A3 genes (Z1a, Z1b, Z2a, Z2b, Z2c, Z2d, Z2e, A3Z3) [27,48], which arose from the ancestral A3Z1, A3Z2 and A3Z3 genes after relatively recent duplication events (Figure 3a). The A3Z1 locus in the horse genome experienced a duplication event ca. 64 Mya (95% HPD 44–79), whereas the A3Z2 locus underwent three rounds of expansion between 39 and 18 Mya. The phylogenetic relationships as well as the accord in timing for the last two duplications suggest that the tandem of ancestral Z2ac and Z2bd genes underwent duplication in a single step, generating the present day arrangement Z2a Z2b Z2c Z2d Z2e [27]. The Equus genus dates back to ca. 4 Mya, posterior to the duplication events that shaped the A3 loci in horses. Thus, it is highly likely that other Equidae, such as zebra and donkey also present a horse-like A3 gene locus.

A3 genes in carnivora

Evidences of A3 genes can be found in the genome of canids and felids. The A3Z3 gene is present in all carnivores genomes analysed, but the A3Z1 gene is missing in Felidae and the A3Z2 gene is missing in Canidae (Figure 1, 3b). Both gene loss events must have occurred independently after the split between cat and dog lineages, some 43–64 Mya. In the genome of the domestic cat (*F. catus*) we identified four A3 genes (Z2a, Z2b, Z2c, and Z3), while only one expressed A3Z2 and one A3Z3 gene could be found in diverse species in the Panthera genus [16]. The duplication events that generated the three A3Z2 genes in the genome of the cat date 30 Mya (95% HPD 13–48) and 14 Mya (95% HPD 4.7-30) (Figure 3b). Although fine details in the evolution of felids are a matter of debate, the MRCA of Pantherinae and Felinae could have lived some 12 Mya [49]. The fact that we have found the A3Z2 gene in lion, tiger and leopard –all of them Pantherinae- basal to the A3Z2 genes in cat, puma and lynx –all of them Felinae- suggests that the split between Pantherinae and Felinae predated the duplication events that generated the A3Z2 diversity in cats. Further, three different mRNA sequences were retrieved from lion and two from tiger [16]. If these mRNA sequences originated from different genes this would imply that independent gene duplications could have occurred in the Panthera lineage. However, considering again the potential for multiple alternative splicing in the A3 genes, the analysis of genomic sequences is required to determine whether panthers and cats have found similar, convergent evolutionary answers to a selective pressure that we still need to identify.

Evidences for selection in the evolution of A3 loci

The A3 genes of primates, rodents, felids, horses and pigs have been described to be under a positive selection [15,16,50-53]. We have searched for hints of positive selection in the sequences in our dataset using Bayesian inference. The AICDA gene, the sister taxon of the A3s, showed to be under strict purifying selection while all A3Z1, A3Z2 and A3Z3 genes contained residues under positive selection (Table 1). Considering all sequences, the distribution of Ka/Ks values was clearly multimodal (Figure 4) for all three A3 genes, with around 25% of positions under strict purifying selection, with *Ka/Ks* values below 0.25, and above 10% of positions under positive selection, with *Ka/Ks* values above 1.1. To exclude that these results were driven by the sequences gathered from Primates the calculations were repeated after excluding them from the dataset, with a similar outcome (Table 1). A slide-window analysis showed that the *Ka/Ks* profiles were similar along the sequences of the three genes (Figure 5), reflecting that the repertoire of sites that are allowed to mutate and to explore sequence space are not evenly distributed and that to a certain extent localise to similar positions in the three A3Z1, A3Z2 and A3Z3 genes. Positions around codons 55 and 145 exhibit very low *Ka/Ks* values in all three A3 genes. There is a better coincidence of the *Ka/Ks* profiles in the C-terminus of the sequences analysed, while there is ample variation in the first 100 codons of the alignment, with A3Z1 and A3Z2 presenting islands of increased the *Ka/Ks* values that do not always overlap. The global distribution of *Ka/Ks* values in all sequences is depicted in Figure 4, where the complexity of the distribution is evident, showing the existence of different site levels of purifying selection, a large proportion of sites evolving close to neutrality and a 10-20% of the area below the curve corresponding to sites under diversifying selection (Table 1).

Evolutionary rate has decreased in the three A3 subfamilies

The introduction of the variable time in the phylogenetic analyses allows to identify variations of the evolutionary rates in different taxa. The global mutation rate for our sequences set was $1.62 \cdot 10^{-3}$ substitutions per site per My (given as Huber M-estimator, Huber-M; median absolute deviation, MAD, $1.07 \cdot 10^{-3}$). The evolutionary rates were not homogeneous for the five gene families analysed (one-way ANOVA, F ratio = 7.414, $p < 0.0001$). Regarding the A1 genes the evolutionary rate is around 1.11 10^{-3}

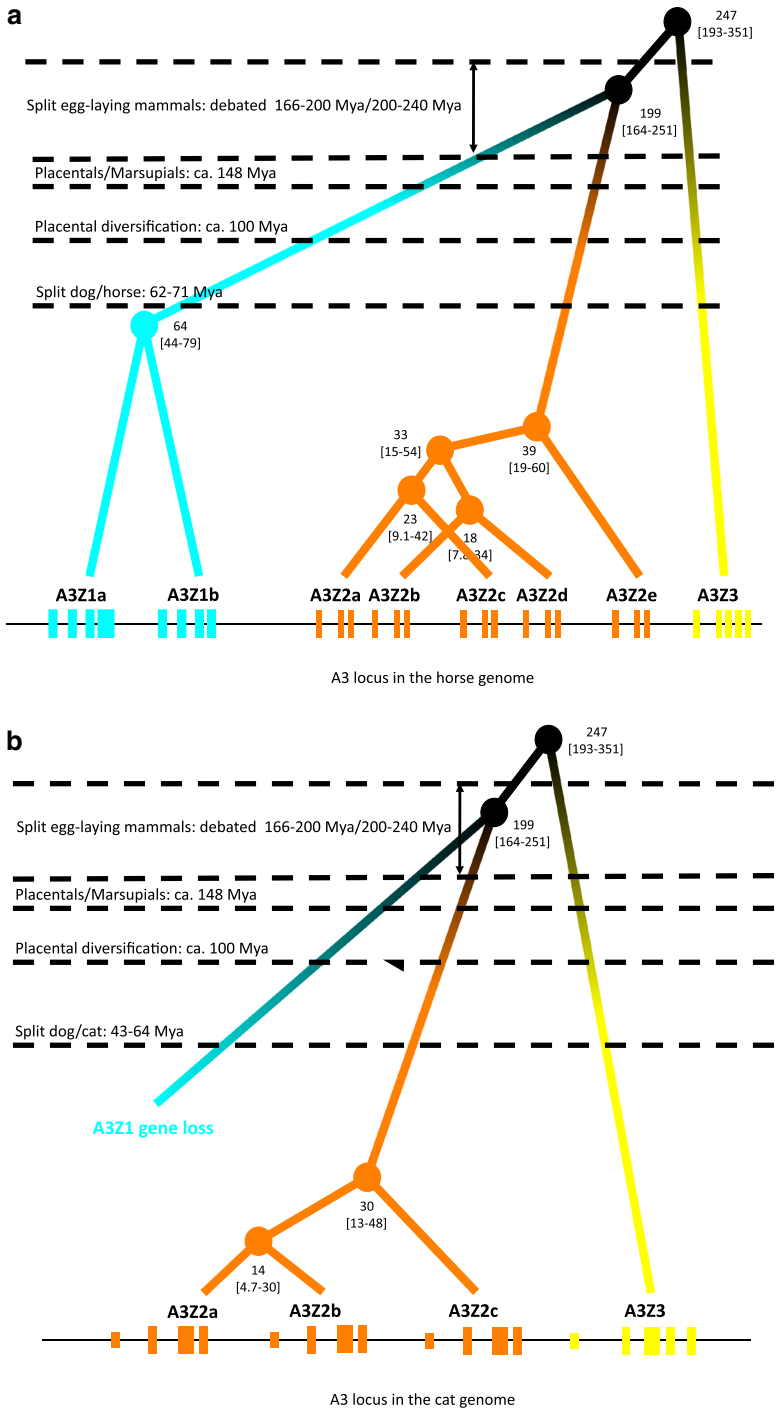


Figure 3 Projection of the A3 phylogenetic relationships onto the horse (a) and cat (b) genomic loci. Exon composition and gene arrangement of the A3 genes are displayed. Labels indicate median age [95% HPD] of the corresponding nodes. Several milestones in the evolution of mammals are indicated.

substitutions per site per My, (Huber-M, Table 2). For this gene family there has been a significant decrease in evolutionary rate with time ($P = 0.00069$, Figure 6a), but it is mainly lead by the values sampled for the oldest node at the root of the family. When this node is not included in the analyses, the evolutionary rate has not significantly varied with time in the last 150 My ($P = 0.089$, Figure 6a). The AICDA genes, the sister group to A3s, displayed the lowest evolutionary rate among the sequences studied (Huber-M $7.41 \cdot 10^{-4}$

Table 1 Statistics for the *Ka/Ks* values obtained either with the Mechanistic Empirical model (MEC) or with Random Effects Likelihood (REL) estimates

		All sequences					Excluding primates sequences		
		A1	AICDA	Z1	Z2	Z3	Z1	QZ2	Z3
MEC ka/ks	quantil 25%	0.10	0.0067	0.29	0.26	0.19	0.44	0.26	0.15
	median	0.40	0.036	0.90	0.88	0.65	0.92	0.8	0.61
	quantil 90%	1.0	0.19	1.7	1.4	1.7	2.2	1.2	1.7
	maximum value	1.9	0.34	3.6	3.8	2.4	3.5	2.6	2.4
REL dN-dS	percentage of sites under positive|puryfying selection	5.4|34	0|100	1.0|15	7.5|14	5.0|12	2.0|8.2	16|13	6.8|12

substitutions per site per My) and this value has not experienced changes with time in the last 400 My (Figure 6a). All A3s however displayed increased evolutionary rate values, around 2.5 times higher than the AICDA sistergroup (Table 2). Very interestingly, the evolutionary rate for each of the three A3 subfamilies has significantly decreased in the last 100 My. For A3Z1 and A3Z3, the linear dependence with time explains a large variation in the evolutionary rate, above 50% for both genes (Figure 6b). In the case of the A3Z2 this dependence is still significant but less obvious, and explains less than 15% of the total variability in evolutionary rate.

Discussion

A3 genes belong to a large superfamily of deaminases that edit nucleic acids and constitute the sister taxa to AICDA, as previous studies suggested [37,54]. Certain members of the deaminase family, such as A1, target RNA as a substrate, but it has been proposed that the ancestral activity may have been to target DNA [54]. Our results show that the MRCA of AICDA and A3, ca 420 Mya, predates the split between the lineages of zebra fish and humans. The subsequent duplication events of the ancestral A3 gene to generate the three extant A3Z1, A3Z2 and A3Z3 genes could not be dated with precision and overlap largely. The first one occurred ca. 247 Mya (95% HPD 193–351 Mya) and the second one ca. 199 Mya (95% HPD 164–251). Representatives of Laurasiatheria, Euarchontoglires and Afrotheria are found in all A3Z1, A3Z2 and A3Z3 subtrees. For Afrotherians, a few sequences were found basal to the A3Z1 subgroup (only sequences from the African bush elephant *Loxodonta Africana*), to the A3Z2 subgroup (the N-terminus from *L. Africana*) or to the A3Z3 subgroup (an *in silico* prediction from genomic DNA from the rock hyrax *Procavia capensis* and the C-terminus from one *L. africana* sequence). On the view of the evolutionary relationships and time divergences, we can conclude that the duplication events that originated the three ancestral A3 loci had already occurred well before diversification within placental mammals, which took place at some point between 95–120 Mya [55]. The presence of A3Z1 and A3Z3 representatives in Afrotheria and of A3Z1, A3Z2 and A3Z3 in Boreoeutheria sustains also this view. The estimated appearance of the most recent common ancestor of all A3s predates the split between placental mammals and marsupials 125–150 Mya, and also the split of monotremes at the base of the crown clade of

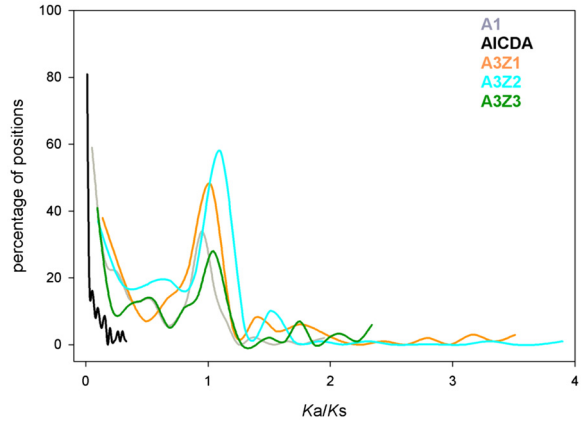


Figure 4 Histogram showing the distribution of the percentages of *Ka/Ks* values for each position for the genes A1, AICDA, A3Z1, A3Z2 and A3Z3 (colour code in the inset).

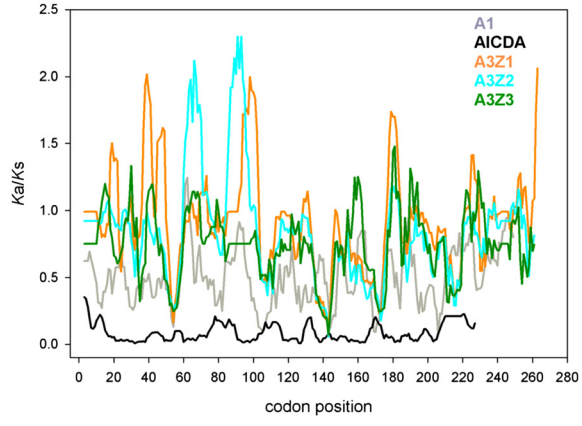


Figure 5 Slide window analysis showing the variation of *Ka/Ks* along the corresponding sequence (colour code in the inset). Values have been computed using a window of five positions and a step of one.

Table 2 Evolutionary rates for the different large clades described measured as substitutions per site per million of years, as inferred from the Bayesian time analyses

	Evolutionary rate (substitutions per site per My)				Evolution of the evolutionary rate (Δ evolutionary rate per My)	
	Huber M-estimator	Median absolute deviation	Mean	95% confidence interval of the mean	Slope	Standard error
all	$1.62\ 10^{-3}$	$1.07\ 10^{-3}$	$1.97\ 10^{-3}$	$4.31\ 10^{-4}$ – $6.44\ 10^{-3}$	NA	NA
A1	$1.11\ 10^{-3}$ A,B	$5.86\ 10^{-4}$	$1.15\ 10^{-3}$	$4.49\ 10^{-4}$ – $2.29\ 10^{-3}$	$5.15\ 10^{-6}$	$1.7\ 10^{-6}$
AICDA	$7.41\ 10^{-4}$ A	$2.74\ 10^{-4}$	$7.99\ 10^{-4}$	$4.06\ 10^{-4}$ – $1.53\ 10^{-3}$	NS	NS
A3Z1	$1.79\ 10^{-3}$ B,C	$9.88\ 10^{-4}$	$2.14\ 10^{-3}$	$7.49\ 10^{-4}$ – $6.50\ 10^{-3}$	$4.15\ 10^{-5}$	$4.4\ 10^{-6}$
A3Z2	$2.06\ 10^{-3}$ C	$1.28\ 10^{-3}$	$2.60\ 10^{-3}$	$4.52\ 10^{-4}$ – $7.61\ 10^{-3}$	$3.34\ 10^{-5}$	$9.4\ 10^{-6}$
A3Z3	$2.13\ 10^{-3}$ C	$1.38\ 10^{-3}$	$2.15\ 10^{-3}$	4.17 10-4 – $4.40\ 10^{-3}$	$2.61\ 10^{-5}$	$4.7\ 10^{-6}$

Evolution of the evolutionary rate measured as variation of the substitution per site per (million of years)^2. *NA* not analysed. *NS* correlation was not significant. Values connected by the same letter are not significantly different after a Tukey-Kramer honestly significant difference test.

modern mammals, debated between 160–200 Mya [56], and 203–238 Mya [57]. Thus, our data support an evolutionary scenario where the genome of the MRCA of all mammals already encoded for at least an ancestral A3, and where the genome of the MRCA of all placental mammals already encoded for a A3Z1-A3Z2-A3Z3 arrangement. This genomic arrangement is conserved in all placental mammals, with the A3 genes flanked by the CBX6 and CBX7 genes, as in chromosome 22 q13.1 in humans. This conserved chromosomic location could be used for targeting and sequencing of the A3 locus of selected species in the future. Unfortunately, the sequence gap between CBX6 and CBX7 is not resolved with certainty in the opossum and in the platypus genomes, and we cannot confirm whether the last common ancestor of Eutheria and Metatheria already possessed a single ancestral A3, the intermediate stage with two loci, or already three loci, as the ancestor of all placental mammals. A note of caution should be nevertheless stated here, since our results cannot go beyond the sequence dataset that we have been able to gather. In a number of genomes we have not been able to recover any relative of the A3 sequences. This holds true for Monotremes, Marsupials and Xenarthra. For other organisms we have not been able to detect certain A3 genes, such as the A3Z1 gene in the cat genome and in Rodentia, or the A3Z2 gene in the dog genome. Since for humans the existence of copy number variation in the A3A and A3B genes is documented in a good number of cases (Additional file 4: Figure S3), a similar situation can be detected in the future in other species. It is important to note here that obtaining genomic sequences of the A3 locus is not trivial, as it evolves under strong selective pressures [45]. In our current analyses we have opted for the most parsimonious explanation: e.g. the presence of an A3Z2 gene in Euarchontoglires and in Felidae and the absence in Canidae has been interpreted as the Boroeutherian ancestor carrying a copy of the ancestral A3Z2 gene, which may had got lost in the Canidae lineage. The hypothetical finding in the future of an A3Z2 gene in Canidae would not affect our timing results as all calibrations have been chosen on subtrees

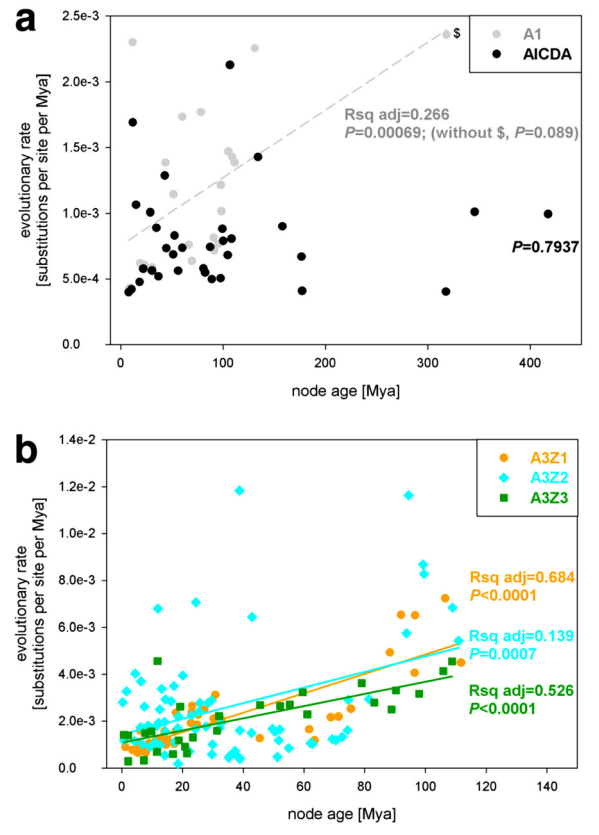


Figure 6 Graph plotting the evolutionary rate against the node age for (a) A1 and AICDA sequences and (b) A3Z1, A3Z2 and A3Z3 sequences. Colour code in the inset. Values for the goodness and significance of the fit to a straight line are provided. For the A1 genes, the values with and without the indicated point ($) are given. When this value is not included in the analysis the trend is not statistically significant.

whose structure matched that of the corresponding species (Additional file 5: Table S2).

The results here presented show that the A3 loci in mammalian genomes are extremely variable and have undergone independent events of gene duplication followed by fixation, gene loss and gene fusion. The outcome of these events has lead to parallel evolution, as is the case of the loss of A3Z1 gene in rodents, dogs and pigs. The most striking result, however, is that the amplification of A3 genes and the fixation of duplicated genes has been selected and expanded in the population, independently in several taxa, such as the A3Z2 gene in Primates (Figure 2), horses (Figure 3a) and cats (Figure 3b) and the A3Z1 gene in Primates and horses. Further, an increase in the frequency of duplication events in the last 50 Mya can be observed in these three lineages, i.e. human, horse, cat (labelled with star symbols in Figure 1). Gene duplications have occurred in both A3Z1 and A3Z2, but not in the A3Z3 clade. Duplication of the A3Z3 gene seems thus to be unfavoured, possibly reflecting the trade-off between the advantages of an increased antiviral repertoire and the increased self-toxicity by gene dosage effect [18]. Also pointing in this direction, there is evidence for independent events of A3Z3 activity loss in human alleles [58]. Finally, further data that exemplify the broad variability in the A3 locus come from copy number variation studies in humans, with several reports on the loss of the A3A (A3Z1a) and/or the A3B (A3Z2a-A3Z1b) genes (Additional file 4: Figure S3).

Different taxa have selected similar solutions through independent events in the A3 locus in terms of gene duplications and/or deletions. Such parallel evolution may be the response to similar environmental changes, e.g. increased retroviral activity either exogenous or endogenous, which could account for the parallel expansion of the A3 family in different mammalian clades. The fixation of similar duplication events in different lineages suggests that the preservation of the duplicated genes has a positive effect in fitness. Further, the main activity of all A3 genes seems to be ssDNA editing, and the differences among them are rather related to substrate specificity [59,60], expression pattern [61,62], and virus specificity [2]. The fate of new A3 copies may thus be sub-functionalisation by broadening the spectrum of molecular targets that can be edited by the A3 proteins, and that this fine tuning of the edited DNA substrates results in an increase in fitness. Such shift in substrate specificity is also supported by the evidence of positive selection in A3s in primates, rodents, felids, horses [16,50-53,63], and our results confirm as well the presence of residues under positive selection in all A3Z1, A3Z2 and A3Z3 clades (Table 1). However, the antiviral relevance of the differential deaminase activity as a function of the ssDNA sequence still needs to be explored.

Our results show that residues evolving under neutrality co-occur with residues under positive and under purifying selection in the A1 and in the A3 genes, and that this holds true for all sequences considered together even after accounting for the excess of Primates sequences (Table 1 and Figure 4). For all these genes, there is a clear peak and a substantial proportion of sites evolving close to neutrality. Additionally, a substantial amount of area below the curve is located in the two tails of the multimodal distribution, corresponding to positions evolving under strict purifying selection and intermediate values of $0.5 < Ka/Ks < 0.8$, and under positive selection. Previous descriptions had rather focused on specific genes in specific taxa, i.e. primates, rodents, horses, felids and pigs [16,50-53,63]. Our analyses have addressed the three A3 gene clusters in all mammals and show that multiple evolutionary pressures coexist in the same gene sequence. Similar cases of positive selection observed as an increase of *Ka/Ks* during subfunctionalisation of duplicated genes has been also documented in other gene families [64]. The biological interpretation of the significance of individual residues under positive selection needs to be analysed in the context of the evolution of the gene in which they reside and in the biochemical context of the exon combination present in the actually expressed proteins. Methods for detecting positive selection based on *Ka/Ks* ratios may result in false positives, as a relaxation of evolutionary constraints could also lead to an increase in this parameter [65]. Such constraint relaxation is specially expected to occur in genes that have undergone duplication events, as it is the case in the A3 genes. Thus, the actual meaning of sites under positive selection in the three A3Z1, A3Z2 and A3Z3 clades requires experimental confirmation. Additionally, gene duplication could be the result of the genomic arms race between target viruses and their hosts, as novel A3 proteins can counteract anti-antiviral activity raised in a virus that evolves anti-A3 mechanisms, such as the Vif or the Bet protein [19-22,24]. This combination could correspond to certain models among those proposed by Innan and Kondrashov [40] to classify the evolution of duplicated genes, such as modified duplication or diversifying selection. The evolution of the A3Z2 locus in primates, horse and cat, with several rounds of successive amplification events, suggests however that gene duplication may have occurred here through an adaptive radiation model, as proposed by Francino [66]. The new A3 copies could explore the sequence space that expands the family of substrates for the action of deaminases, typically by modifying the sequence context of the C->U edition. Similarity among duplicated genes could also allow for gene conversion and/or recombination [67], further enlarging the repertoire of substrates. In analogy to gene conversion at the DNA level, similarity in DNA sequence between tandem duplicated genes may be responsible for the read-through mRNA species that

encompass exons from different genes, and that have been described in cats [16,68], horses [27] and pigs [15,47] and have also been communicated in humans [69] (Additional file 6: Figure S4). Currently more than 500 ESTs in the databases map onto the A3 human loci, with a number of them containing spliced exons from different A3 genes (Additional file 6: Figure S4). This additional generation of diversity adds to the high number of alternatively spliced mRNAs originated from each individual A3 gene, especially of those composed of two fused "mono" A3s, either A3Z2-A3Z1 or A3Z2-A3Z2. In these cases we face fused genes that retain their individual coding capacity, can generate multiple splice alternatives and show potential to encode for read-through mRNAs. Finally, an additional level of spatiotemporal complexity to the regulation of this antiviral activity arises from A3 expression heterogeneity linked to cell type, tissue, developmental time or exposure to foreign DNA.

The architecture of several tandem copies of paralogs in a genome facilitates non-homologous recombination among paralog copies resulting in gene conversion, as has been suggested for the A3G gene in humans [67]. A high degree of gene conversion is expected to result in an increased degree of sequence similarity through concerted evolution [70,71], with the undesired outcome of rendering younger divergence times [72]. Our dataset is unfortunately not suited for an in-depth analysis of gene conversion, since it includes well-characterised mRNAs, but also ESTs that may contain exons that originate from different genes and putative mRNAs inferred from genomic sequences. A proper analysis will need good quality genomic sequences with enough sampling of individuals if variations found in the human A3 locus in terms of gene copy number and indels (Additional file 4: Figure S3 and Additional file 6: S4) appear also in other species. Such analyses will need to address gene order within the locus, but also exon-intron order within the genes, and alternative splicing and mRNA read-through. Nevertheless, certain cases can be considered. The confounding role of gene conversion may have disturbed the phylogenetic reconstructions for the A3Z2 genes. In Primates, the A3Z2a, A3Z2c and A3Z2e A3Z2g paralogs have appeared after two duplication events from their common ancestor around 73 Mya (Figure 2). The topology for the A3Z2g subtree matches well the phylogeny of the corresponding species (Additional file 2: Figure S1), and gene conversion may have had a limited impact here. The evolution of the A3Z2a, A3Z2c and A3Z2e genes however cannot be reconstructed with confidence, and appears as a series of small branches with small support values, which could be interpreted as a signature of gene conversion. The same holds true for the rest of the A3Z2 paralogs, A3Z2b/d/f. Finally, a second candidate for gene conversion to have occurred is the A3Z2a/b gene tandem in the cat genome [16].

The AICDA genes show the lowest evolutionary rate among the five clades studied, 1.11 10^{-3} substitutions per site per My, significantly different from the values inferred for the three A3 genes, which are around three times higher (Table 2). Very interestingly, we have found that these evolutionary rates have decreased for the A3 genes in the last 100 My, whereas for the AICDA and the A1 genes there are no significant variations in the evolutionary rate with time (Table 2). We interpret the different outcome for the A3 genes and for their sister taxa as an evidence for our results being genuine rather than an artefact from the evolutionary inference. The simultaneous identification of positions evolving under positive selection in the three A3 genes and the finding of a trend towards decrease with time of the evolutionary rate in the same genes are not contradictory. Our calculations for *Ka/Ks* have been performed separately for each position, while the values inferred for the evolutionary rates refer to the corresponding nodes in the phylogenetic reconstruction. In our scenario of gene duplication and gene family expansion, we interpret that the episodes of gene duplication have lowered the restrictions on the duplicated copies of the genes, thus allowing for increased evolutionary rates. Subfunctionalisation of the sister copies may have yielded one population of conserved sites evolving under purifying pressures (e.g. those sites that are indispensable for the deaminase activity), one population of sites positively selected sites (e.g. those that are responsible for the differential substrate recognition) and a third large majority of sites evolving close to neutrality. The decrease in evolutionary rate possibly reflects a plateau in the fixation of the novel function. In the A3Z2 genes, which have undergone the largest number of duplication events, the large variation in evolutionary rate for the different nodes (Table 2) and the lower proportion of the decrease in evolutionary rate that is explained by the independent variable time alone (less than 15%; Figure 6b) supports further the idea that gene duplication fosters a transient increase in evolutionary rate.

We have not been able to identify in the databases sequences that could be orthologs of the A3 genes in Marsupials or in Monotremes. We have dated the MRCA of the A3 genes around 246 Mya (95% HPD 192–351 Mya), and the second split that generated the MRCA of A3Z1 and A3Z2 around 199 Mya (95% HPD 164–250). The timing for the crown clade of mammals (i.e. the split between Monotremata and Theria) as well as the timing for the split between marsupials and placentals are controversial, ranging between 160 and 240 Mya [55-57]. Three explanations could thence account for the absence of extant A3 genes in monotremes and marsupials: i) we simply lack information and further sequencing will provide us with the missing genes; ii) members of one or of both groups may indeed have lost the A3 genes; and iii) genes may be

exclusive to placentals if the MRCA of placental mammals predates the appearance of the MRCA of all A3s. Our interpretation implies in any case that the genome of the placental ancestor already encoded for an A3 locus with the arrangement A3Z1-A3Z2-A3Z3. The absence of A3 genes in Xenarthra, the fourth large clade within placental mammals, must therefore imply either gene loss or incomplete coverage of the A3 locus in the two species analysed. Since we have described that loss of certain A3 genes has occurred in parallel in different lineages, as in Rodents and in Artyodactyla, it is conceivable that in certain lineages the loss of all A3 genes may have been selected. The adaptive value of an enlarged armoury against viruses is obvious, and evidences supporting positive selection of the A3 genes in different branches of the mammalian tree are strong. The intriguing hypothesis of the total local loss of A3 genes might imply that the constraints and pressures imposed by viral infections can largely vary among different taxa. Certain host lineages may thus either be less exposed to (certain) viruses, and/or may have evolved alternative antiviral strategies. An exciting question that arises from our dating results is the relative coincidence in time between speciation events and gene duplication events, as exemplified in Figures 2 and 3. It could be speculated that gene duplication of the A3 genes may trigger speciation, possibly through the differential fitness against viral infections that the additional A3 gene copy provides. Further genetic and functional research will be required to elucidate the fitness landscapes integrating viral pressures, expansion of the A3 repertoire and concomitant risks for the own genetic information.

Conclusions

The A3 gene family appeared together with the ancestral mammals. Independently in certain lineages, the A3 locus was expanded through a series of tandem duplications, best exemplified in Primates. The repertoire of A3 proteins is additionally expanded through splice alternatives and read-through mechanisms, resulting in broader substrate specificity and finer regulation of DNA modification potential. We have shown that this diversity has been generated by series of tandem duplication in the A3 locus probably followed by positive selection and/or relaxation of constraints and resulting in sub/neo-functionalisation. Such evolutionary solution has been independently selected in several lineages: Primates, felids and equids. Our findings constitute a paradigm of genomic parallel evolutionary solutions in the framework of the arms race between viruses and their hosts.

Methods

Dataset

Annotated sequences were retrieved from GenBank and Ensembl Genome Browser, the data set was completed with cDNA sequences and with genomic sequences putatively encoding for products similar to A3 after iterative BLAT, tBLASTn and PSI-BLAST searches. Our final dataset included sequences from within Laurasiatheria (seven Carnivora, seven Cetartiodactyla, one Perissodactyla and one Chiroptera species), Euarchontoglires (three Rodentia and seventeen Primates species) and Afrotheria (one Hyracoidea and one Proboscidea species). Fused genes as in human A3Z2a-A3Z1b were split and analysed as two sequences.

A selection of 24 APOBEC1 (A1) sequences (21 Eutheria species, two Metatheria and one Aves species) and 34 Activation induced cytidine deaminases (AICDA) sequences (23 Eutheria species, three Metatheria, one Prototheria, two Aves, two Amphibia and three Actinopterygii species) and were used as outgroups. The final sequence set comprised 202 sequences. A list of species and accession numbers is given in Additional file 1: Table S1.

Phylogenetic inference

Sequences were aligned with MAFFT (http://mafft.cbrc.jp/alignment/software/) at the amino acid level and back-translated into nucleotides for phylogenetic analyses. This matrix is available as Additional file 7 and from IGB on request. Maximum likelihood (ML) inference was performed with RAXML_v7.2.8 (http://www.exelixis-lab.org/) [73] using 5000 bootstrap cycles with the GTR + G4 model and introducing three partitions, one per codon position (number of patterns 290, 290 and 298, respectively) and allowing for different total tree length for each partition. Bayesian inference was performed with PHYLOBAYES v3.3 (http://www.phylobayes.org) [74] using the same dataset, under the GTR model, removing constant sites, running two independent chains and checking for convergence comparing discrepancies among partitions. Trees obtained after ML and Bayesian reconstructions were compared regarding topological congruence as well as pairwise distances using K-TreeDist [75], with the following result: Robinson-Foulds distance, 67/401 (16.7%); K-score, 1.07; scale factor, 0.542. Both ML and Bayesian topologies rendered A3Z1, A3Z2 and A3Z3 as monophyletic, and A1 as the outgroup to A3s and AICDA. Twenty-five local topologies, highly supported in both analyses (above 90% bootstrap support and above 0.98 Bayesian posterior probability) and consistent with the phylogenetic relationships of the hosts were identified and used for molecular dating, using truncated uniform priors based on fossil calibration dates, as suggested by Benton and Donaghue [56,76]. Calibration dates are listed in Additional file 5: Table S2. All discussion on the results obtained refers to the dates and analyses proposed by TimeTree (http://www.timetree.org/index.php) [77]. The

backbone tree for time inference was constructed with RAxML v7.2.8 with the same settings as described above, further forcing monophyly for each of the nodes used for calibration (–*g* option), plus for the sequences belonging to the crowngroups Afrotheria, Laurasiatheria and Euarchontoglires within each of the clades A1, AICDA, A3Z1, A3Z2 and A3Z3, respectively. This tree was not significantly worse than the original one, under the maximum-likelihood framework, as evaluated with a Shimodaira-Hasegawa test [78] implemented in RAxML v7.2.8. The comparison between the unconstrained and the constrained maximum likelihood trees rendered the following values: Robinson-Foulds distance, 71/401 (17.7%); K-score; 0.780; scale factor, 0.823. Using this tree, Bayesian time inference was performed with PHYLOBAYES v3.3 using a discrete gamma distribution with eight categories under the GTR matrix of exchange rates, using a log-normal autocorrelated relaxed clock together with a uniform prior on divergence times. A gamma prior of mean 550 and standard deviation 200 Mya was specified for the age of the root. The results from three 50-million steps independent chains were combined and analysed.

Positive selection analysis

Evidence for positive selection was analysed in a Bayesian framework using SELECTON V2.4 (http://selecton.tau.ac.il/) and DATAMONKEY (http://www.datamonkey.org/). The alignments for the A1, AICDA, A3Z1, A3Z2and A3Z3 genes were analysed separately for the whole sequence set, and those for A3Z1, A3Z2 and A3Z3 were additionally analysed after excluding the Primates sequences, to discard that the results were driven by the behaviour of sequences from this clade. With SELECTONV2.4, the Mechanistic Empirical Model (MEC) was tested against the M8a model [79]. In both cases the topology of the best-known maximum likelihood tree previously obtained was used as scaffold for the calculation of synonymous *Ks* and non-synonymous *Ka* mutations. Briefly, MEC expands an empirical amino acid replacement matrix –in our case the WAG matrix [80], the best scoring one as determined by ProtTest [81] among those available in the SELECTON algorithm- into a codon replacement matrix. This way, the chemical similarity between amino acids is incorporated when calculating non-synonymous substitutions [79]. The likelihood of this model was tested against the M8a model, which does not allow for positive selection, using the Akaike information criterion [82]. For all of the sequence sets the MEC model was preferred over the M8a. For each position, the value for Ka/Ks was calculated. The variation of *Ka/Ks* along the corresponding sequence was analysed using a slide-window analysis of width five and step one. With DATAMONKEY, the Random Effects Likelihood (REL) method was applied, which assumes an underlying nucleotide substitution model and allows for rate variation in both *Ka* and *Ks* substitution rates [83]. All analyses were repeated after excluding the Primates sequences from the dataset, in order to exclude putative biases inherent to the overrepresentation of this clade in terms of sequence composition and presence of the different genes in the genome. Statistic analyses were performed with R v1.40 and with JMP v7.0.2.

Additional files

Additional file 1: Table S1. Taxonomy and sequence accession numbers of the sequences used in this study.

Additional file 2: Figure S1. Best-known maximum likelihood tree for the A3 genes analysed, AICDA and for the outgroup A1. Colour code describes mammalian taxa, as in Figure 1. Values in the nodes depict bootstrap support.

Additional file 3: Figure S2. mRNAs deposited in the databases originating from the human A3 locus, after the USCS Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTracks), showing human chromosome 22, positions 39,250,000 to 39,550,000, accessed on December 13th 2011.

Additional file 4: Figure S3. Bayesian dated tree for the A3 genes analysed, AICDA and for the outgroup A1. Bars around the nodes describe the 95% HPD for the inference of the node age.

Additional file 5: Table S2. Values for calibration used in this study.

Additional file 6: Figure S4. Copy number variation in the human A3 locus, after the Database of Genomic Variants (http://projects.tcag.ca/cgi-bin/variation/gbrowse/hg19/), showing human chromosome 22, positions 39,250,000 to 39,550,000, accessed on December 13th 2011.

Additional file 7: Final sequence matrix, codon-aligned.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

Study conception: CM and IGB. Data collection and analysis: AW and IGB. Data interpretation: IGB. Manuscript draft: CM, AW and IGB. All authors read and approved the final manuscript.

Acknowledgements

We thank Dieter Häussinger for continuous support. The authors thank Francisco Codoner for technical help with initial sequence retrieval. This work was supported by the disappeared Spanish Ministry for Science and Innovation (MICINN) [Programa Ramón y Cajal, BFU2009-06702-E/BMC and CGL2010-16713]; the Generalitat Valenciana [FPA/2011/002]; and the Heinz-Ansmann Foundation for AIDS Research.

Author details

[1]Clinic for Gastroenterology, Hepatology and Infectiology, Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany. [2]Genomics and Health, Centre for Public Health Research (CSISP), Valencia, Spain. [3]Infections and Cancer, Catalan Institute of Oncology (ICO) | Bellvitge Institute of Biomedical Research (IDIBELL), Barcelona, Spain. [4]Infections and Cancer, Catalan Institute of Oncology (ICO), Avda. Gran Via, 199-203, L'Hospitalet de Llobregat, Barcelona 08908, Spain.

Received: 6 January 2012 Accepted: 1 May 2012
Published: 28 May 2012

References

1. Vartanian JP, Guetard D, Henry M, Wain-Hobson S: **Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions.** *Science* 2008, **320**(5873):230–233.
2. Chiu YL, Greene WC: **The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements.** *Annu Rev Immunol* 2008, **26**:317–353.

3. Chen H, Lilley CE, Yu Q, Lee DV, Chou J, Narvaiza I, Landau NR, Weitzman MD: **APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons.** *Curr Biol* 2006, **16**(5):480–485.
4. Tsuge M, Noguchi C, Akiyama R, Matsushita M, Kunihiro K, Tanaka S, Abe H, Mitsui F, Kitamura S, Hatakeyama T, *et al*: **G to A hypermutation of TT virus.** *Virus Res* 2008, **149**(2):211–216.
5. Vartanian JP, Henry M, Marchio A, Suspene R, Aynaud MM, Guetard D, Cervantes-Gonzalez M, Battiston C, Mazzaferro V, Pineau P, *et al*: **Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis.** *PLoS Pathog* 2010, **6** (5):e1000928.
6. Renard M, Henry M, Guetard D, Vartanian JP, Wain-Hobson S: **APOBEC1 and APOBEC3 cytidine deaminases as restriction factors for hepadnaviral genomes in non-humans in vivo.** *J Mol Biol* 2010, **400**(3):323–334.
7. Peng ZG, Zhao ZY, Li YP, Wang YP, Hao LH, Fan B, Li YH, Wang YM, Shan YQ, Han YX, *et al*: **Host apolipoprotein b messenger RNA-editing enzyme catalytic polypeptide-like 3 G is an innate defensive factor and drug target against hepatitis C virus.** *Hepatology* 2011, **53**(4):1080–1089.
8. Suspene R, Aynaud MM, Koch S, Pasdeloup D, Labetoulle M, Gaertner B, Vartanian JP, Meyerhans A, Wain-Hobson S: **Genetic editing of herpes simplex virus 1 and Epstein-Barr herpesvirus genomes by human APOBEC3 cytidine deaminases in culture and in vivo.** *J Virol* 2011, **85** (15):7594–7602.
9. Ferholz M, Kendl S, Prifert C, Weissbrich B, Lemon K, Rennick L, Duprex PA, Rima BK, Koning FA, Holmes RK, *et al*: **The innate antiviral factor APOBEC3G targets replication of measles, mumps, and respiratory syncytial virus.** *J Gen Virol* 2012. doi:10.1099/vir.0.038919-0. in press.
10. Pauli EK, Schmolke M, Hofmann H, Ehrhardt C, Flory E, Münk C, Ludwig S: **High level expression of the anti-retroviral protein APOBEC3G is induced by influenza A virus but does not confer antiviral activity.** *Retrovirology* 2009, **6**:38.
11. Kremer M, Suezer Y, Martinez-Fernandez Y, Münk C, Sutter G, Schnierle BS: **Vaccinia virus replication is not affected by APOBEC3 family members.** *Virol J* 2006, **3**:86.
12. Petit V, Vartanian JP, Wain-Hobson S: **Powerful mutators lurking in the genome.** *Philos Trans R Soc Lond B Biol Sci* 2009, **364**(1517):705–715.
13. Jarmuz A, Chester A, Bayliss J, Gisbourne J, Dunham I, Scott J, Navaratnam N: **An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22.** *Genomics* 2002, **79**(3):285–296.
14. LaRue RS, Andresdottir V, Blanchard Y, Conticello SG, Derse D, Emerman M, Greene WC, Jonsson SR, Landau NR, Löchelt M, *et al*: **Guidelines for naming nonprimate APOBEC3 genes and proteins.** *J Virol* 2009, **83**(2):494–497.
15. LaRue RS, Jonsson SR, Silverstein KA, Lajoie M, Bertrand D, El-Mabrouk N, Hotzel I, Andresdottir V, Smith TP, Harris RS: **The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals.** *BMC Mol Biol* 2008, **9**:104.
16. Münk C, Beck T, Zielonka J, Hotz-Wagenblatt A, Chareza S, Battenberg M, Thielebein J, Cichutek K, Bravo IG, O'Brien SJ, *et al*: **Functions, structure, and read-through alternative splicing of feline APOBEC3 genes.** *Genome Biol* 2008, **9**(3):R48.
17. Zaranek AW, Levanon EY, Zecharia T, Clegg T, Church GM: **A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing.** *PLoS Genet* 2010, **6**(5):e1000954.
18. Suspene R, Aynaud MM, Guetard D, Henry M, Eckhoff G, Marchio A, Pineau P, Dejean A, Vartanian JP, Wain-Hobson S: **Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism.** *Proc Natl Acad Sci U S A* 2011, **108**(12):4858–4863.
19. Sheehy AM, Gaddis NC, Choi JD, Malim MH: **Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein.** *Nature* 2002, **418**(6898):646–650.
20. Russell RA, Wiegand HL, Moore MD, Schafer A, McClure MO, Cullen BR: **Foamy virus Bet proteins function as novel inhibitors of the APOBEC3 family of innate antiretroviral defense factors.** *J Virol* 2005, **79**(14):8724–8731.
21. Mariani R, Chen D, Schröfelbauer B, Navarro F, König R, Bollman B, Münk C, Nymark-McMahon H, Landau NR: **Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif.** *Cell* 2003, **114**(1):21–31.
22. Löchelt M, Romen F, Bastone P, Muckenfuss H, Kirchner N, Kim YB, Truyen U, Rosler U, Battenberg M, Saib A, *et al*: **The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein.** *Proc Natl Acad Sci U S A* 2005, **102**(22):7982–7987.
23. Derse D, Hill SA, Princler G, Lloyd P, Heidecker G: **Resistance of human T cell leukemia virus type 1 to APOBEC3G restriction is mediated by elements in nucleocapsid.** *Proc Natl Acad Sci U S A* 2007, **104**(8):2915–2920.
24. Perkovic M, Schmidt S, Marino D, Russell RA, Stauch B, Hofmann H, Kopietz F, Kloke BP, Zielonka J, Strover H, *et al*: **Species-specific inhibition of APOBEC3C by the prototype foamy virus protein bet.** *J Biol Chem* 2009, **284**(9):5819–5826.
25. Okeoma CM, Huegel AL, Lingappa J, Feldman MD, Ross SR: **APOBEC3 proteins expressed in mammary epithelial cells are packaged into retroviruses and can restrict transmission of milk-borne virions.** *Cell Host Microbe* 2010, **8**(6):534–543.
26. Li MM, Emerman M: **Polymorphism in human APOBEC3H affects a phenotype dominant for subcellular localization and antiviral activity.** *J Virol* 2011, **85**(16):8197–8207.
27. Zielonka J, Bravo IG, Marino D, Conrad E, Perkovic M, Battenberg M, Cichutek K, Münk C: **Restriction of equine infectious anemia virus by equine APOBEC3 cytidine deaminases.** *J Virol* 2009, **83(15):**7547–7559.
28. Ross SR: **Are viruses inhibited by APOBEC3 molecules from their host species?** *PLoS Pathog* 2009, **5**(4):e1000347.
29. Berger A, Münk C, Schweizer M, Cichutek K, Schule S, Flory E: **Interaction of Vpx and apolipoprotein B mRNA-editing catalytic polypeptide 3 family member A (APOBEC3A) correlates with efficient lentivirus infection of monocytes.** *J Biol Chem* 2010, **285**(16):12248–12254.
30. Browne EP, Littman DR: **Species-specific restriction of apobec3-mediated hypermutation.** *J Virol* 2008, **82**(3):1305–1313.
31. Doehle BP, Schafer A, Wiegand HL, Bogerd HP, Cullen BR: **Differential sensitivity of murine leukemia virus to APOBEC3-mediated inhibition is governed by virion exclusion.** *J Virol* 2005, **79**(13):8201–8207.
32. Kobayashi M, Takaori-Kondo A, Shindo K, Abudu A, Fukunaga K, Uchiyama T: **APOBEC3G targets specific virus species.** *J Virol* 2004, **78**(15):8238–8244.
33. Low A, Okeoma CM, Lovsin N, de las Heras M, Taylor TH, Peterlin BM, Ross SR, Fan H: **Enhanced replication and pathogenesis of Moloney murine leukemia virus in mice defective in the murine APOBEC3 gene.** *Virology* 2009, **385**(2):455–463.
34. Okeoma CM, Petersen J, Ross SR: **Expression of murine APOBEC3 alleles in different mouse strains and their effect on mouse mammary tumor virus infection.** *J Virol* 2009, **83**(7):3029–3038.
35. Okeoma CM, Low A, Bailis W, Fan HY, Peterlin BM, Ross SR: **Induction of APOBEC3 in vivo causes increased restriction of retrovirus infection.** *J Virol* 2009, **83**(8):3486–3495.
36. Okeoma CM, Lovsin N, Peterlin BM, Ross SR: **APOBEC3 inhibits mouse mammary tumour virus replication in vivo.** *Nature* 2007, **445**(7130):927–930.
37. Conticello SG: **The AID/APOBEC family of nucleic acid mutators.** *Genome Biol* 2008, **9**(6):229.
38. Ohno S: *Evolution by gene duplication*. New-York: Springer; 1970.
39. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet* 2002, **18**(12):619–620.
40. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models.** *Nat Rev Genet* 2010, **11**(2):97–108.
41. Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, Hulihan M, Peuralinna T, Dutra A, Nussbaum R, *et al*: **Alpha-Synuclein locus triplication causes Parkinson's disease.** *Science* 2003, **302**(5646):841.
42. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, *et al*: **Diet and the evolution of human amylase gene copy number variation.** *Nat Genet* 2007, **39**(10):1256–1260.
43. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154**(1):459–473.
44. Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, *et al*: **A molecular phylogeny of living primates.** *PLoS Genet* 2011, **7**(3):e1001342.
45. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, *et al*: **Insights into hominid evolution from the gorilla genome sequence.** *Nature* 2012, **483**(7388): 169–175.
46. Price SA, Bininda-Emonds OR, Gittleman JL: **A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla).** *Biol Rev Camb Philos Soc* 2005, **80**(3):445–473.
47. Dörrschuck E, Fischer N, Bravo IG, Hanschmann KM, Kuiper H, Spotter A, Moller R, Cichutek K, Münk C, Tönjes RR: **Restriction of Porcine**

Endogenous Retrovirus by Porcine APOBEC3 Cytidine Deaminases. *J Virol* 2011, **85**(8):3842–3857.
48. Bogerd HP, Tallmadge RL, Oaks JL, Carpenter S, Cullen BR: **Equine infectious anemia virus resists the antiretroviral activity of equine APOBEC3 proteins through a packaging-independent mechanism.** *J Virol* 2008, **82**(23):11889–11901.
49. Johnson WE, Eizirik E, Pecon-Slattery J, Murphy WJ, Antunes A, Teeling E, O'Brien SJ: **The Late Miocene radiation of modern Felidae: a genetic assessment.** *Science* 2006, **311**(5757):73–77.
50. Zhang J, Webb DM: **Rapid evolution of primate antiviral enzyme APOBEC3G.** *Hum Mol Genet* 2004, **13**(16):1785–1791.
51. OhAinle M, Kerns JA, Malik HS, Emerman M: **Adaptive evolution and antiviral activity of the conserved mammalian cytidine deaminase APOBEC3H.** *J Virol* 2006, **80**(8):3853–3862.
52. Sawyer SL, Emerman M, Malik HS: **Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G.** *PLoS Biol* 2004, **2**(9): E275.
53. Sanville B, Dolan MA, Wollenberg K, Yan Y, Martin C, Yeung ML, Strebel K, Buckler-White A, Kozak CA: **Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove.** *PLoS Pathog* 2010, **6**:e1000974.
54. Severi F, Chicca A, Conticello SG: **Analysis of reptilian APOBEC1 suggests that RNA editing may not be its ancestral function.** *Mol Biol Evol* 2011, **28** (3):1125–1129.
55. Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A: **The delayed rise of present-day mammals.** *Nature* 2007, **446**(7135):507–512.
56. Benton MJ, Donoghue PC: **Paleontological evidence to date the tree of life.** *Mol Biol Evol* 2007, **24**(1):26–53.
57. Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, *et al*: **Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification.** *Science* 2011, **334**(6055):521–524.
58. OhAinle M, Kerns JA, Li MM, Malik HS, Emerman M: **Antiretroelement activity of APOBEC3H was lost twice in recent human evolution.** *Cell Host Microbe* 2008, **4**(3):249–259.
59. Langlois MA, Beale RC, Conticello SG, Neuberger MS: **Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities.** *Nucleic Acids Res* 2005, **33**(6):1913–1923.
60. Kohli RM, Maul RW, Guminski AF, McClure RL, Gajula KS, Saribasak H, McMahon MA, Siliciano RF, Gearhart PJ, Stivers JT: **Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification.** *J Biol Chem* 2010, **285**(52):40956–40964.
61. Koning FA, Newman EN, Kim EY, Kunstman KJ, Wolinsky SM, Malim MH: **Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets.** *J Virol* 2009, **83**(18):9474–9485.
62. Refsland EW, Stenglein MD, Shindo K, Albin JS, Brown WL, Harris RS: **Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction.** *Nucleic Acids Res* 2010, **38**(13):4274–4284.
63. Ortiz M, Bleiber G, Martinez R, Kaessmann H, Telenti A: **Patterns of evolution of host proteins involved in retroviral pathogenesis.** *Retrovirology* 2006, **3**:11.
64. Thornton K, Long M: **Excess of amino acid substitutions relative to polymorphism between X-linked duplications in Drosophila melanogaster.** *Mol Biol Evol* 2005, **22**(2):273–284.
65. Nozawa M, Suzuki Y, Nei M: **Reliabilities of identifying positive selection by the branch-site and the site-prediction methods.** *Proc Natl Acad Sci U S A* 2009, **106**(16):6700–6705.
66. Francino MP: **An adaptive radiation model for the origin of new gene functions.** *Nat Genet* 2005, **37**(6):573–577.
67. Henry M, Guetard D, Suspene R, Rusniok C, Wain-Hobson S, Vartanian JP: **Genetic editing of HBV DNA by monodomain human APOBEC3 cytidine deaminases and the recombinant nature of APOBEC3G.** *PLoS One* 2009, **4** (1):e4277.
68. Stern MA, Hu C, Saenz DT, Fadel HJ, Sims O, Peretz M, Poeschla EM: **Productive replication of Vif-chimeric HIV-1 in feline cells.** *J Virol* 2010, **84** (14):7378–7395.
69. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, *et al*: **Origins and functional impact of copy number variation in the human genome.** *Nature* 2010, **464** (7289):704–712.
70. Ohta T: **Some models of gene conversion for treating the evolution of multigene families.** *Genetics* 1984, **106**(3):517–528.
71. Hartl DL, Clark AG: **Molecular population genetics**. In *Principles of population genetics*. Massachusetts: Sinauer Associates Inc; 2007:317–384.
72. Teshima KM, Innan H: **The effect of gene conversion on the divergence between duplicated genes.** *Genetics* 2004, **166**(3):1553–1560.
73. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.
74. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25**(17):2286–2288.
75. Soria-Carrasco V, Talavera G, Igea J, Castresana J: **The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees.** *Bioinformatics* 2007, **23**(21):2954–2956.
76. Parham JF, Donoghue PC, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT, *et al*: **Best practices for justifying fossil calibrations.** *Syst Biol* 2011, **61**:346–359.
77. Hedges SB, Dudley J, Kumar S: **TimeTree: a public knowledge-base of divergence times among organisms.** *Bioinformatics* 2006, **22**(23): 2971–2972.
78. Shimodaira H, Hasegawa M: **Multiple comparison of log-likelihoods with application to phylogenetic inference.** *Mol Biol Evol* 1999, **16**:1114–1116.
79. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T: **Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach.** *Nucleic Acids Res* 2007, **35 (Web Server issue)**:W506–W511.
80. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Mol Biol Evol* 2001, **18**(5):691–699.
81. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21**(9):2104–2105.
82. Akaike H: **A new look at the statistical model identification.** *IEEE Trans Automatic Control* 1974, **119**:716–723.
83. Kosakovsky Pond SL, Frost SD: **Not so different after all: a comparison of methods for detecting amino acid sites under selection.** *Mol Biol Evol* 2005, **22**(5):1208–1222.

doi:10.1186/1471-2148-12-71
Cite this article as: Münk *et al.*: An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evolutionary Biology* 2012 **12**:71.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central