


# Molecular evolution of type II MAGE genes from ancestral MAGED2 gene and their phylogenetic resolution of basal mammalian clades

Marcos De Donato<sup>1,2</sup>  · Sunday O. Peters<sup>3</sup> · Tanveer Hussain<sup>1,4</sup> · Hectorina Rodulfo<sup>2</sup> · Bolaji N. Thomas<sup>5</sup> · Masroor E. Babar<sup>4</sup> · Ikhide G. Imumorin<sup>1,6</sup>

Received: 8 March 2017 / Accepted: 6 May 2017 / Published online: 17 May 2017  
© Springer Science+Business Media New York 2017

**Abstract** Type II melanoma-associated antigens (MAGE) are a subgroup of about a dozen proteins found in various locations in the genome and expressed in normal tissues, thus are not related to cancer as the type I MAGE genes. This gene family exists as a single copy in non-mammals and monotremata, but found as two copies in metatherians and occur as a diverse group in all eutherians. Our studies suggest MAGED2 as the ancestor of this subfamily and the most likely evolutionary history of eutherian type II MAGE genes is hereby proposed based on synteny conservation, phylogenetic relations, genome location, homology conservation, and the protein and gene structures. Type II genes can be divided into two: those with 13 exons (MAGED1, MAGED2, TRO, and MAGED4) and those with only one exon (MAGEE1, MAGEE2, MAGEF1, NSMCE3, MAGEH1, MAGEL2, and NDN) with different evolutionary patterns. Our results suggest a need to change the gene nomenclature to MAGE1 (the

ancestral gene), currently designated as LOC103095671 and LOC100935086, in opossum and Tasmanian devil, respectively, and MAGE2 (the duplicated one), currently designated as LOC100617402 and NDNL2, respectively, to avoid confusion. We reconstructed the phylogenetic relationships among 23 mammalian species using the combined sequences of MAGED1, MAGED2, MAGEL2, and NDN, because of their high divergence, and found high levels of support, being able to resolve the phylogenetic relationships among Euarchontoglires, Laurasiatheria, Afrotheria, and Xenarthra, as an example that small, but phylogenetically informative sequences, can be very useful for resolving basal mammalian clades.

## Introduction

Melanoma-associated antigens (MAGE) are a family of genes, consisting of 38 genes and 18 pseudogenes annotated in humans, whose members have been divided into two big subfamilies: type I and II, based on differences in tissue-specific expression and gene structure (van der Bruggen et al. 1991). Type I members are large in number, with no introns and all located on the X chromosome, in subtype-specific clusters known as A, B, and C, and they are mainly expressed in testis and cancer cells (Chomez et al. 2001). Since the expression of MAGE type I proteins is associated with malignancy, they are being studied as targets for cancer vaccine development (Atanackovic et al. 2004), as well as the use of their expression pattern in cancer cells for diagnostic or prognostic purposes (Sang et al. 2011).

The second group, known as type II MAGE genes, has more heterogeneous structures and are dispersed in different locations of the genome. They are expressed in normal

✉ Ikhide G. Imumorin  
igi2@cornell.edu

<sup>1</sup> Animal Genetics and Genomics Laboratory, International Programs, College of Agriculture and Life Science, Cornell University, Ithaca, NY 14853, USA

<sup>2</sup> Escuela de Bioingenierías, Instituto Tecnológico y de Estudios Superiores de Monterrey, Queretaro, Mexico

<sup>3</sup> Department of Animal Science, Berry College, Mount Berry, GA 30149, USA

<sup>4</sup> Department of Molecular Biology, Virtual University of Pakistan, Lahore 54000, Pakistan

<sup>5</sup> Department of Biomedical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA

<sup>6</sup> School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA

tissues, with no relation to cancer (Osterlund et al. 2000). Although the physiological functions of type II MAGE proteins remain largely uncharacterized (Sang et al. 2011), they are increasingly important due to their roles in the regulation of cell cycle progression and cell differentiation. For example, MAGED1, D2, TRO, E1, E2, F1, NSMCE3, H1, and NDN are involved in the early process of neurogenesis, and MAGEL2 is connected with maintenance of pluripotency of stem cells (Liu et al. 2012). In addition, they may function in the ubiquitination cascade mediated by RING proteins, since they have been identified as binding partners for both type I and II MAGE proteins (Feng et al. 2011).

The only region of homology conserved in all of the members of the family is a stretch of ~200 amino acids known as MAGE homology domain (MHD), usually located close to the COOH termini of the proteins, which has been proposed to interact with p75 neurotrophin or related receptors in some of the members of the family (Chomez et al. 2001). However, despite the sequence and structural similarities of the MHD for the different proteins, mounting evidence suggests that MHDs are more versatile and complex than one might expect, and rather than recognizing and binding a common motif, MHDs confer binding specificity to multiple unique interaction motifs (Doyle et al. 2010; Lee and Potts 2017).

Due to the importance of type I genes in cancer, studies have investigated the evolutionary relationship within this group (Katsura and Satta 2011), but the relationships among type II genes and their evolutionary patterns of duplication and diversification remain unclear, even though these genes have important functions in the cell. Some studies have suggested that since MAGED genes (TRO, MAGED1, MAGED2, and MAGED4) in eutherians have multiple exons like the non-eutherian condition, they should represent the ancestral gene (Chomez et al. 2001; Katsura and Satta 2011). Other studies have suggested that NSMCE3 is more functionally related to the ancestral MAGE (Lee and Potts 2017). A proteomics study identified NSMCE3 as the human ortholog of yeast NSE3 and determined NSMCE3 and its cognate RING ligase, NSE1, to be essential components of the human SMC5/6 complex (Taylor et al. 2008). In addition, NSMCE3 shows highest sequence identity to the *Drosophila* MAGE protein (Nishimura et al. 2007). Moreover, the chicken MAGE protein and human NSMCE3 interact with E2F1 and the p75 neurotrophin receptor (López-Sánchez et al. 2007). Therefore, while genomic architecture points to MAGED genes, functional studies suggest that NSMCE3 may be most related to the ancestral MAGE (Lee and Potts 2017).

In this study, we carried out detailed phylogenetic analyses of the eutherian MAGE type II genes, comparing them with the MAGE genes in non-eutherian vertebrates to try to

infer the evolutionary history of the members of this subfamily. In addition, due to their rapid diversification and duplication events early in eutherian evolution, we assess the use of these genes as evolutionary markers to understand early eutherian diversification.

## Materials and methods

### Sequence retrieval and analysis

Protein sequence IDs of MAGE type II genes of *Homo sapiens* (human, taxon ID: 9606), *Macaca mulatta* (rhesus monkey, taxon ID: 9544), *Pongo abelii* (Sumatran orangutan, taxon ID: 9601), *Callithrix jacchus* (marmoset, taxon ID: 9483), *Mus musculus* (mouse, taxon ID: 10090), *Rattus norvegicus* (rat, taxon ID: 10116), *Marmota marmota* (marmot, taxon ID: 9993), *Oryctolagus cuniculus* (rabbit, taxon ID: 9986), *Bos taurus* (cattle, taxon ID: 9913), *Ovis aries* (sheep, taxon ID: 9940), *Lipotes vexillifer* (river dolphin, taxon ID: 118797), *Sus scrofa* (pig, taxon ID: 9823), *Vicugna pacos* (alpaca, taxon ID: 30538), *Pteropus alecto* (black flying fox, taxon ID: 9402), *Equus asinus* (donkey, taxon ID: 9793), *Equus caballus* (horse, taxon ID: 9796), *Ceratotherium simum simum* (white rhinoceros, taxon ID: 73337), *canis lupus familiaris* (dog, taxon ID: 9615), *Ailuropoda melanoleuca* (panda bear, taxon ID: 9646), *Acinonyx jubatus* (cheetah, taxon ID: 32536), *Trichechus manatus latirostris* (Florida manatee, taxon ID: 127582), *Dasypus novemcinctus* (nine-banded armadillo, taxon ID: 9361), and *Loxodonta africana* (elephant, taxon ID: 9785) retrieved for analyses are shown in Table 1. In addition, we used MAGE protein sequences for *Sarcophilus harrisi* (Tasmanian devil, MAGE1: XP\_003775163.1 and MAGE2: XP\_003771463.2, taxon ID: 9305), *Monodelphis domestica* (opossum, MAGE1: XP\_007507813.1 and MAGE2: XP\_003342056, taxon ID: 13616), *Ornithorhynchus anatinus*, (platypus, XP\_001510511.2, taxon ID: 9258), *Anolis carolinensis* (green anole, XP\_003215786.1, taxon ID: 28377), *Gallus gallus* (chicken, NP\_001098534.1, taxon ID: 9031), *Tangara guttata* (speckled tanager, NP\_001232572.1, taxon ID: 256443), *Xenopus tropicalis* (western clawed frog, NP\_001016930.1, taxon ID: 8364), *Gekko japonicus* (Japanese gecko, XP\_015279741.1, taxon ID: 146911), and *Python bivittatus* (Burmese python, XP\_007422261.1, taxon ID: 176946) in the analyses.

Protein sequences were analyzed using ScanProsite, a database of protein domains, families, and functional sites (Sigrist et al. 2010) at the ExPASy bioinformatics resource portal developed by the Swiss Institute of Bioinformatics. Sequence logo of the MHD region was built using WebLogo (Crooks et al. 2004), which shows a graphical representation of the amino acid conservation among

**Table 1** Protein accession numbers of the other class type II genes in the mammals used in this study

Species	MAGED1	MAGED2	MAGEL2	NDN	TRO	MAGED4	MAGED4B	MAGEE1	MAGEE2	MAGEF1	NSMCE3	MAGEHI
Armadillo	XP_004481832	XP_012385125	XP_012384027	XP_004467451								
Bat	XP_006915920	XP_006923751	XP_006907152	XP_006907151								
Cattle	NP_001039590.2	NP_001069133	XP_002696517	NP_001014982	XP_003588232	NP_001096781			NP_001070344	NP_001095519	NP_001071548	NP_001074197
Cheetah	XP_014931744	XP_014943434	XP_014919590	XP_014919582								
Dog	XP_538044.4	XP_851949	XP_003434387	XP_545810	XP_855479.3	XP_538047.3			XP_538082.3	XP_545233.2	XP_005618335	XP_549024
Dolphin	XP_007467471	XP_007446530	XP_007459954	XP_007459953								
Donkey	XP_014695406	XP_014709353	XP_014700972	XP_014700973								
Elephant	XP_010598207	XP_003420419	XP_010598486	XP_003420671	XP_003420437	XP_003420431			XP_003412819	XP_003412822	XP_003420657	XP_003420423
Horse	XP_001914968	XP_003365769	XP_014589560	XP_001492662	XP_014584151	XP_003365705			XP_001505033	XP_014588272	XP_014588955	XP_005614253
Human	NP_001005333	NP_055414.2	NP_061939.3	NP_002478	NP_001034794	NP_001092270	NP_001229291	NP_065983	NP_619648	NP_071432.2	NP_619649	NP_054780.2
Manatee	XP_004376878	XP_004389189	XP_004372734	XP_004372686								
Marmoset	XP_003735781	XP_002762958	XP_008996490	XP_002749089								
Marmot	XP_015362797	XP_015362033	XP_015362076	XP_015362074								
Mouse	NP_062765	NP_001186175	NP_038807.4	NP_035012.2	NP_001002272			NP_444431.3	NP_444436	Pseudogene	NP_075728	NP_076277
Orangutan	XP_002831705	XP_009233159	XP_009247884	XP_002825254								
Panda	XP_002930758	XP_011216411	XP_011231287	XP_002925795								
Pig	NP_001001860	XP_003135155	XP_001924925.2	NP_001116616	XP_003484162	XP_003135170				XP_003483331	XP_005659866	XP_003135159
Rabbit	XP_008246727	XP_008270880	XP_008267991	XP_002718331								
Rat	NP_445861	NP_536727.2	XP_001054803	NP_001008558	XP_001067331			NP_001073360	NP_001100411	Pseudogene	NP_001166998	NP_001013268
Rhesus	XP_002808576	NP_001244543	XP_001114423.2	NP_001165573	XP_001092862.2	XP_002806293		XP_001098743	XP_001098246	XP_001097423	NP_001244436	NP_001247554
Rhinoceros	XP_004443842	XP_004443279	XP_004431927	XP_004431926								
Sheep	XP_004022123	XP_012028595	XP_014957332	XP_004017754								
Vicugna	XP_015107724	XP_015107882	XP_015102456	XP_015102439								

the proteins containing this domain in the UniProtKB/Swiss-Prot databases. The map of the region containing the MAGE gene in the green anole and platypus, as well as the region containing the MAGED2 and TRO genes in human, mouse, cattle, dog, horse, and pig was constructed using the gene structure and location from the MapView at NCBI.

A comparison of the genomic sequence of human MAGED4 and MAGED4B genes was carried out using mVISTA program suite (Dubchak and Ryaboy 2006, <http://genome.lbl.gov/vista/index.shtml>) by the alignment of the region NC\_000023.10:51723109-52069309 in the sense orientation with its antisense sequence. The composition of the repetitive elements in this region was determined using RepeatMasker (A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.3.0, <http://www.repeatmasker.org>).

### Phylogenetic analysis

Phylogenetic analyses were carried out using three approaches. First, the maximum likelihood method was used based on the JTT matrix-based model (Jones et al. 1992) using MEGA6 (Tamura et al. 2013). Initial tree(s) for the heuristic search were obtained by applying the neighbor-joining method to a matrix of pairwise distances estimated using a JTT model. A discrete gamma distribution was used to model evolutionary rate differences among sites. Second, the maximum likelihood method was used with a 4-matrix model (LG4X) using PhyML software (Phylogenetic Maximum Likelihood, version 3.0, Guindon et al. 2010), where sites are categorized depending on their evolutionary rate, and different replacement matrices were used for each site category of the gamma distribution assumption (Le et al. 2008). Third, a Bayesian phylogenetic analysis was conducted using Mr.Bayes, v 3.2.1 (Ronquist et al. 2012), implementing the mixed model setting to estimate the amino acid rate change using different models of fixed rate matrices.

We studied the evolutionary patterns of the human proteins using the protein sequences of the MHD (because the rest of the protein sequence showed poor conservation) for all the members of the type II genes, except for MAGEH1 because it contains only a partial domain, and the frog and platypus MAGE as outgroups. Additionally, we studied the evolution of MHD for all jawed vertebrates (Gnathostomata) with annotated sequences and the human MAGED2 for eutherian mammals. Finally, we reconstructed the phylogenetic relationships among 23 mammalian species using the combined complete protein sequences of MAGED1, MAGED2, MAGEL2, and NDN. These were selected because they were present in all the species and their protein sequences have been annotated. Additionally,

MAGEL2 and NDN show the highest divergence among all type II MAGE proteins, being highly informative to help define the evolutionary pattern of the mammalian clades.

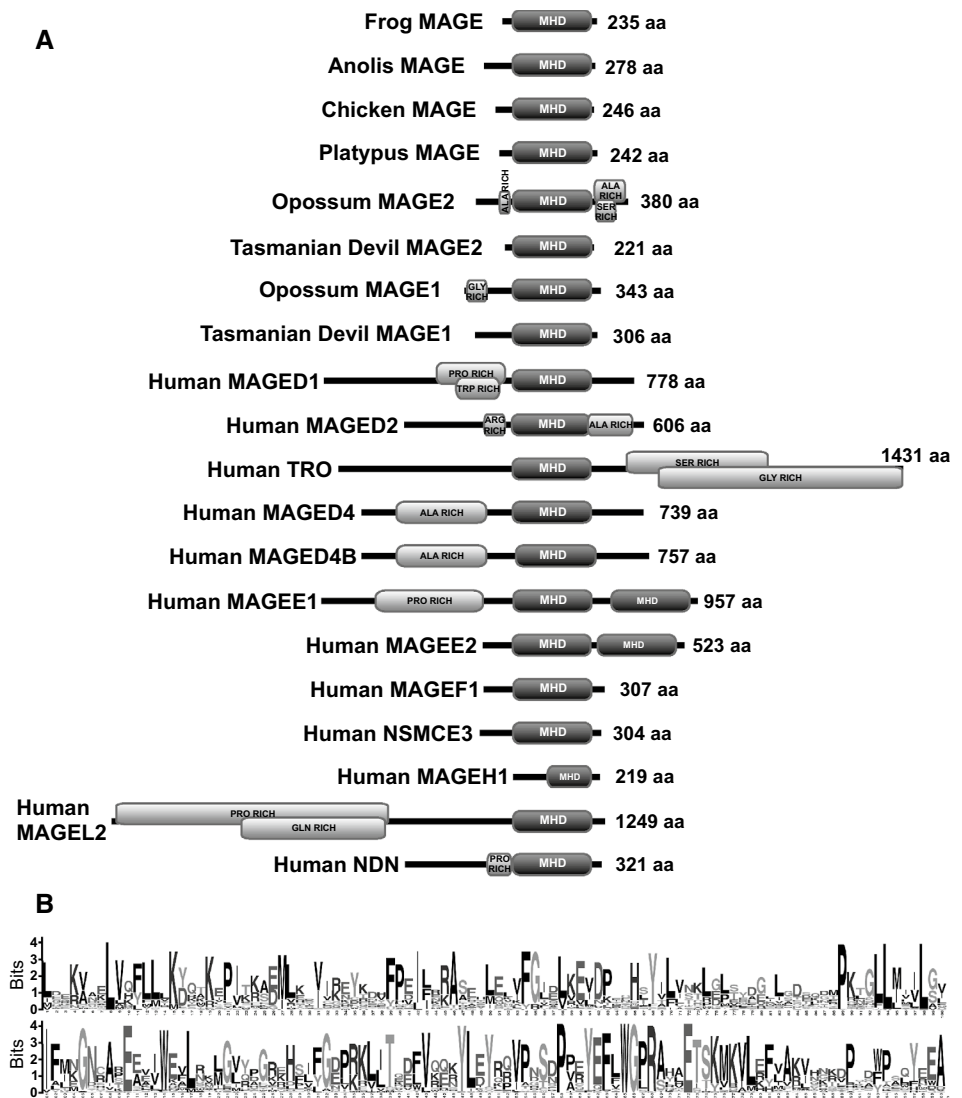
### Results

The structure of MAGE proteins in frog, lizard, chicken, platypus, opossum, and Tasmanian devil was similar, but the structure of MAGE genes in humans shows great variation and larger sequences for most of the members (Fig. 1a). The logo of the MAGE homology domain (MHD) shows that many of the amino acids are highly conserved (Fig. 1b), even though the sequence of the rest of the protein have changed significantly in different members of the family in eutherians.

The genomic region containing the lizard MAGE gene has been conserved in mammals, although an inversion has rearranged the order of the genes (Fig. 2). In eutherians, this region contains MAGED2 and TRO and these genes were involved in small inversions in cattle, while in dog the genes MAGED2 and ITIH5L are inverted and the last was duplicated into ITIH6. The size of the region in the lizard and platypus is similar, but it is twice the size in eutherians. There are significant differences in the size of the introns in the genes of this region among all the species. The localization of MAGED2 and TRO in the conserved region containing the ancestral MAGE gene suggests that one of these two genes is the candidate ancestor for all the eutherian MAGE genes.

Very similar topologies were showed by all three phylogenetic analyses (Figs. 3, 5, 6), but there were some differences in the trees. In most cases, the Bayesian trees showed smaller substitutions per site and larger for LG4X. Also, ML trees tended to have lower branch support, while the supports for Bayesian and LG4X branches were mostly similar. Two different groups of type II genes of the MAGE family can be established from the structure of the genes, with one group showing 13 exons (MAGED1, MAGED2, TRO, and MAGED4), like the ancestral MAGE, and the other with only one exon (MAGEE1, MAGEE2, MAGEF1, NSMCE3, MAGEH1, MAGEL2, and NDN). The phylogenetic relationships of the MAGE domain reveal different evolutionary patterns between these two groups of proteins (Fig. 3), with the 13-exon group being clearly monophyletic and of lower divergence compared to the other genes. All the trees showed that MAGED2 protein was ancestral to the 13-exon group and TRO as being more derived than MAGED2. Also, NDN and MAGEL2 from the group of genes with only one exon, were the closest to the ancestral MAGE protein, and were probably derived early in eutherian diversification. The group of 1-exon

**Fig. 1** Structure of type II MAGE protein. **a** Characterization of human type II MAGE protein sequences compared to the MAGE protein in the western clawed frog (*Xenopus tropicalis*), green anole (*Anolis carolinensis*), chicken (*Gallus gallus*), platypus (*Ornithorhynchus anatinus*), opossum (*Monodelphis domestica*), and the Tasmanian devil (*Sarcophilus harrisii*), using the ScanProsite. **b** Protein sequence logo of the MHD showing the most conserved amino acids in each position. The y axis is the relative frequency of the amino acids

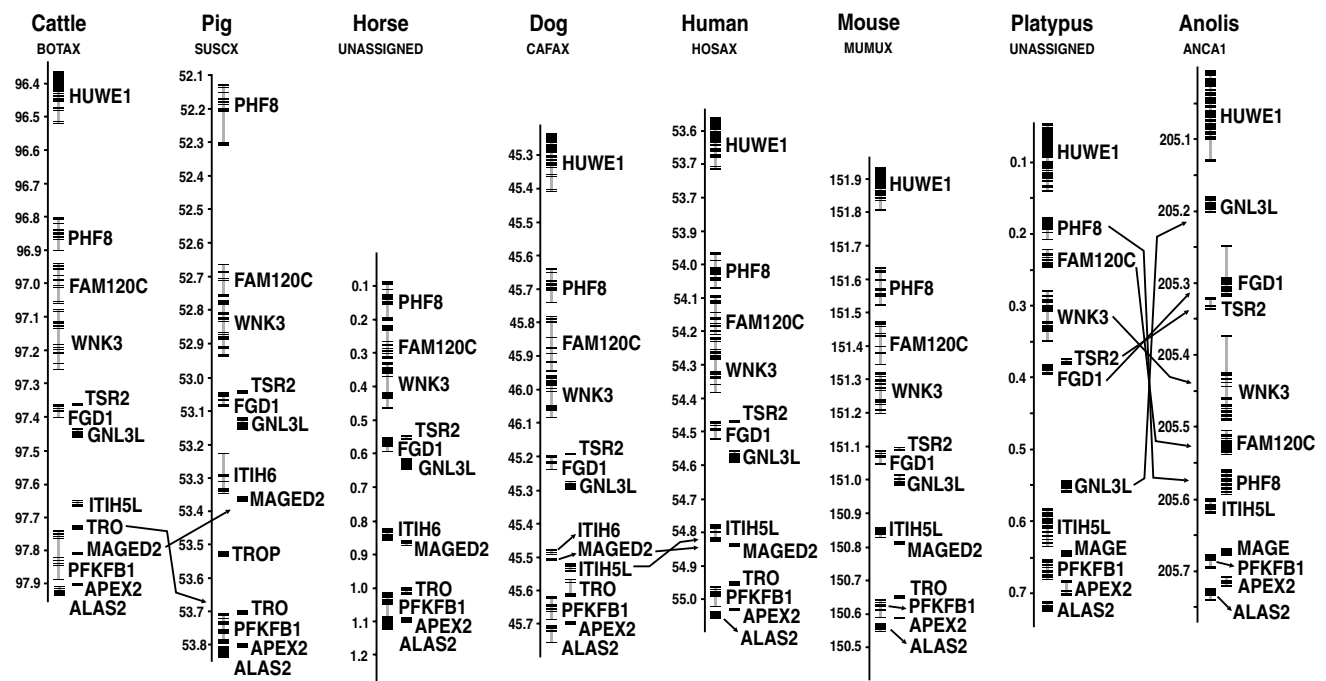


proteins shows higher divergence than the 13-exon proteins (Fig. 4), and they have probably gone through rapid duplication and diversification events during early eutherian evolution.

The phylogenetic pattern of the MAGE proteins in tetrapoda (Fig. 5) can be followed from the theoretical ancestor of frogs to all diapsids (lizards, snakes, crocodiles, and birds) and mammals. However, in the opossum and the Tasmanian devil, a second MAGE gene is present with high conservation of both genes in both species. The eutherian homologous region, where the opossum MAGE2 gene is located, does not show conservation to any of the eutherian MAGE genes, and the phylogenetic trees suggest that the duplication event in metatherian occurred after the split from the common ancestor to eutherian. The ancestral genes in these species are located on the X chromosome, while the derived gene is located on chromosome 8 in opossum and not yet localized in the Tasmanian devil.

Further, MAGE2 was most likely duplicated by retrotransposition, since it has no introns, differing from MAGE1 which has 10 exons.

The phylogenetic analyses of all the MAGE type II genes in human, rhesus monkey, mouse, rat, cattle, pig, dog, horse, and elephant follow a similar evolutionary pattern with few exceptions (Fig. 4). For example, MAGEF1 is present in mouse and rat as pseudogenes. Also, MAGED4 protein is not present in either species, but a BLAST search using the human mRNA sequences localizes a region containing segments that combined have 71.4% identity in a 1614 bp sequence in the mouse, about 90 kb apart from MAGED1, and 67.4% identity in a 1039 bp sequence in the rat, about 120 kb apart from MAGED1, showing the loss of this gene in rodents. In humans, MAGED1 and MAGED4 are about 400 kb apart from each other. Alternatively, MAGEE1 protein is only present in primates and rodents. It is evident



**Fig. 2** Genomic region containing the ancestral MAGE gene. Maps of the regions containing the MAGE gene in the green anole (*Anolis*) and platypus, compared to the regions containing MAGED2 and TRO in mouse, human, dog, horse, pig, and cattle. Physical distances (Mb) are drawn to scale in the eutherian mammals, but in the green anole and platypus are drawn in a different scale. The mouse genome

is drawn in a reverse orientation for better comparison. *Black, wide boxes* represent the exons and the *gray, thin boxes* represent introns. Genes to the left are in antisense orientation and those to the right are in a sense orientation. Chromosome assignment is located below the common name, being all known eutherian on the X chromosome and in the platypus on chromosome 1

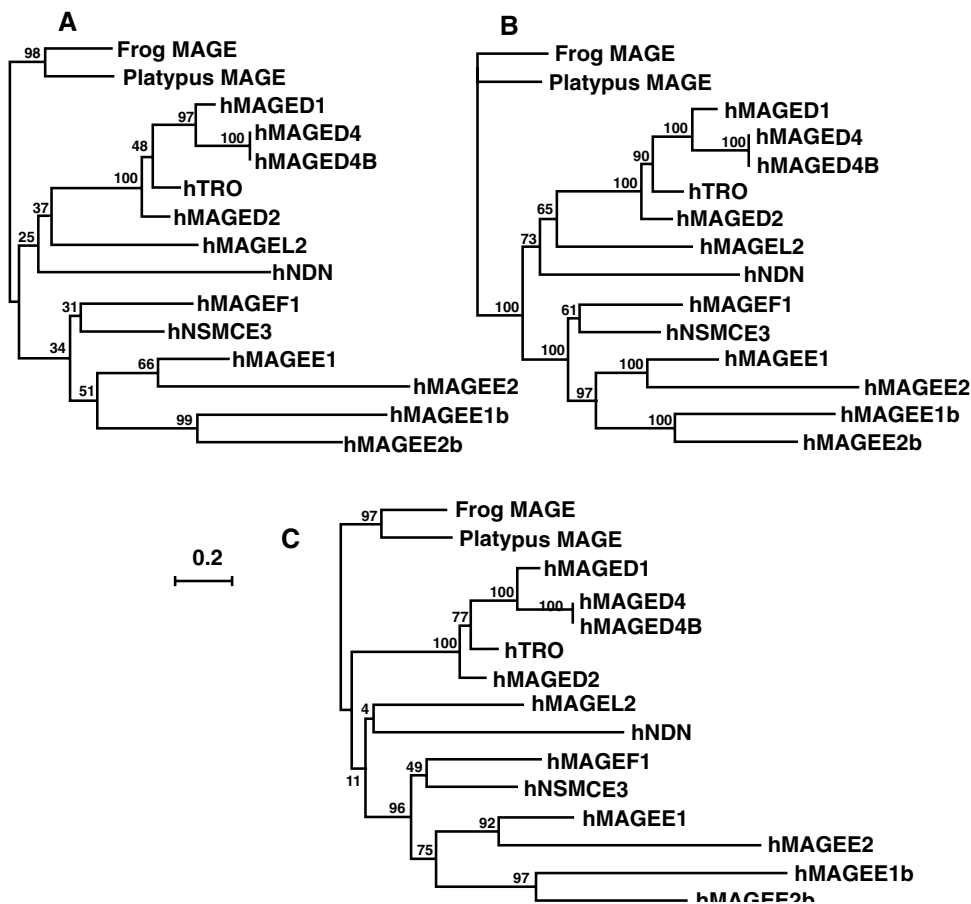
that MAGED1, MAGED2, MAGED4, and NDN are the least divergent of all the proteins (Fig. 4), even though all of these genes but NDN are located on the X chromosome. On the other hand, TRO, MAGEL2, MAGEF1, and MAGEE1 are the most divergent.

The phylogenetic trees of the eutherian species built using the MHD sequences of MAGED1, MAGED2, MAGEL2, and NDN proteins show that most branches have very high levels of support (Fig. 6), with good differentiation among the basal groups used. Carnivora and Perissodactyla form a single cluster, as well as Cetartiodactyla and Chiroptera (*P. alecto*) which also form a cluster, which agrees with Laurasiatheria, previously reported to contain these groups. Elephant and manatee cluster together next to armadillo. Rodents and Lagomorpha cluster together but show high divergence to the rest of the mammals, and are next to primates, as expected.

The human genome shows a duplication of MAGED4 and MAGED4B not present in the genomes of chimpanzee, orangutan, or any other primate. The sequence alignment of both genes shows an inverse orientation between them and with sequence identity of 100% of the coding and intergenic regions but distributed in blocks separated by different repetitive elements, mostly LINEs (Fig. 7). At the edge of the two inverted regions, two LINE1

elements (L1PA3) are located in inverse orientation and this might be the breakpoint of duplication of the entire region.

The most likely evolutionary history of the eutherian type II MAGE genes, based on their phylogenetic relations, genome location, homology conservation, and the protein and gene structures, is shown in Fig. 8. In eutherians, TRO most likely originated by duplication from MAGED2, and it is probable that this whole segment was duplicated to originate MAGED4 and MAGED1, respectively, since the MAGED2–MAGED1 and TRO–MAGED4 pairs are closely related in all eutherian mammals analyzed. The genes of the 1-exon group were most likely derived from either MAGEL2 or NDN, since they are basal to the 13-exon group of MAGE type II. Because MAGEL2, NDN, and NSMCE3 are located in the same chromosomes in all the species, it is very likely that they originated by duplication events from MAGEL2 or NDN, while MAGEF1 was most likely duplicated by transposition. Due to the short sequence of MAGEH1, it is very difficult to infer where it comes from, but it may have been derived from a MAGE gene of the 1-exon group by transposition, since there is a strong phylogenetic relationship to these genes.



**Fig. 3** Phylogenetic analysis of the human MAGE homology domains (MHD). Phylogeny of all the members of the type II genes, except for MAGEH1, in the human genome, as well as for the MAGE proteins in the frog (*Xenopus tropicalis*) and platypus (*Ornithorhynchus anatinus*). Phylogenetic analyses were carried out using the maximum likelihood method based on the JTT matrix-based model (a), a Bayesian analysis implementing the mixed model setting to estimate the amino acid rate change (b), and a ML with a 4-matrix

model where sites are categorized depending on their evolutionary rate (c). The trees are drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 15 amino acid sequences, with a total of 203 positions in the final dataset. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Frog and platypus were used as outgroups since they are the closest to the ancestral MAGE sequence

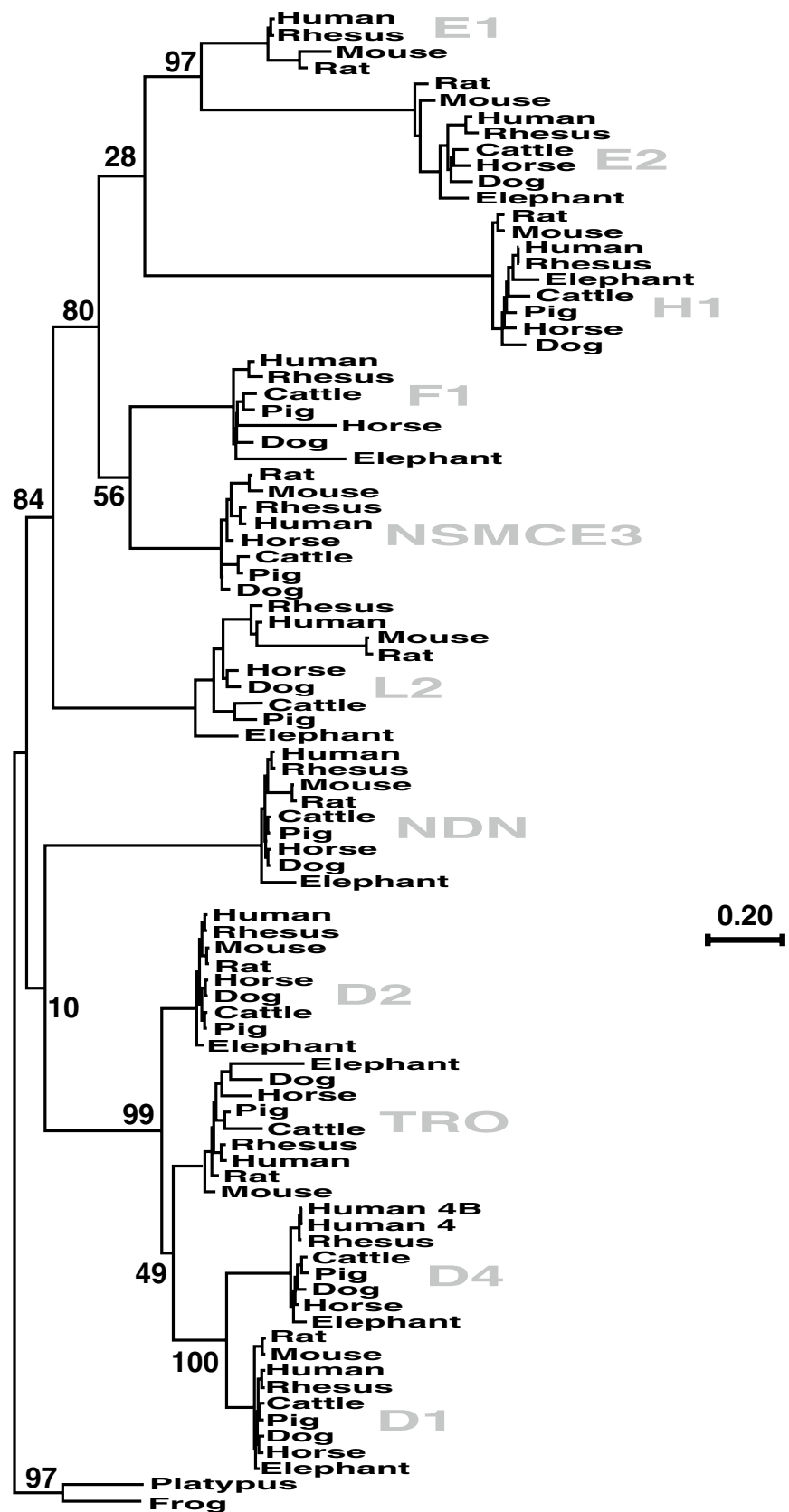
## Discussion

López-Sánchez et al. (2007) analyzed the sequence similarity of MHD among all eukaryotes and found that this domain has been well conserved in all the metazoans they studied, as well as with some degree of conservation in plants and protozoans, suggesting that MAGE proteins are ancient in eukaryotes. They also suggest that the ancestral gene may have lacked introns, since the MAGE genes in *Entamoeba histolytica* and *Drosophila melanogaster* do not contain introns. Katsura and Satta (2011) identified 2 genes in the opossum, the American marsupial, showing a monophyletic tree. This is consistent with our result. We also identified two MAGE genes from the Tasmanian devil, the Australian marsupial, forming a monophyletic tree with high similarity between these species for

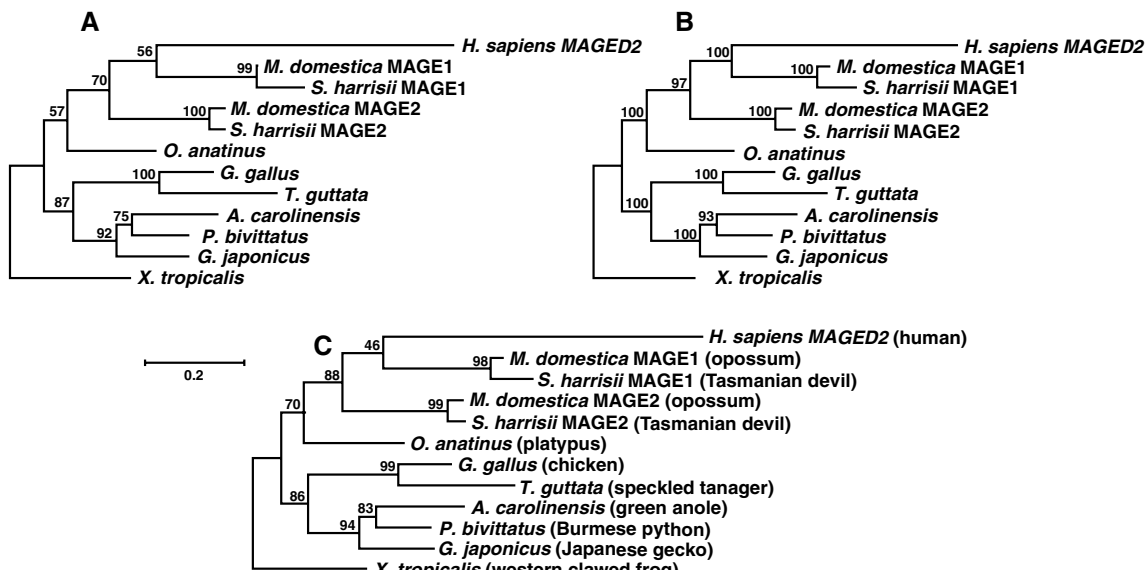
each gene, suggesting a duplication event before the split of these two groups. We therefore suggest changes in the gene nomenclature in these species as MAGE1 (the ancestral gene), currently designated as LOC103095671 and LOC100935086, in opossum and Tasmanian devil, respectively, and MAGE2 (the duplicated one), currently designated as LOC100617402 and NDNL2, respectively, in order to avoid confusion.

Even though the region containing the MAGE gene in platypus has not been assigned to a chromosome yet, the gene HUWE1 in that region has been localized on chromosome 6 by in situ hybridization (Delbridge et al. 2009), and this chromosome has been shown to be homologous to the autosomal ancestor of the sex chromosome in metatherians and eutherians. The monotremata sex chromosomes have no homology with the therian sex chromosome, instead

**Fig. 4** Phylogenetic analysis of all MAGE type II genes from 9 mammalian species. Phylogeny of all the members of the type II genes, except for MAGEH1, using the MHD from 9 mammalian species, as well as for the MAGE proteins in the frog (*Xenopus tropicalis*) and platypus (*Ornithorhynchus anatinus*). The phylogenetic analysis was carried out using a Bayesian analysis implementing the mixed model setting to estimate the amino acid rate change. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 92 amino acid sequences, with a total of 218 positions in the final dataset. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown only next to the branches separating the genes but not the species' sequences. Frog and platypus were used as outgroups since they are the closest to the ancestral MAGE sequence







**Fig. 5** Phylogenetic analysis of the MAGE homology domains (MHD) in tetrapoda. Phylogeny of representatives of the main groups of living tetrapoda. Phylogenetic analyses were carried out as described in the legend of Fig. 3 and in methods. The analysis

involved 12 amino acid sequences, with a total of 203 positions in the final dataset. The frog sequence was used as an outgroup since it should be the closest to the ancestral sequence of the diapsids (lizards, snakes, crocodiles, and birds) and mammals

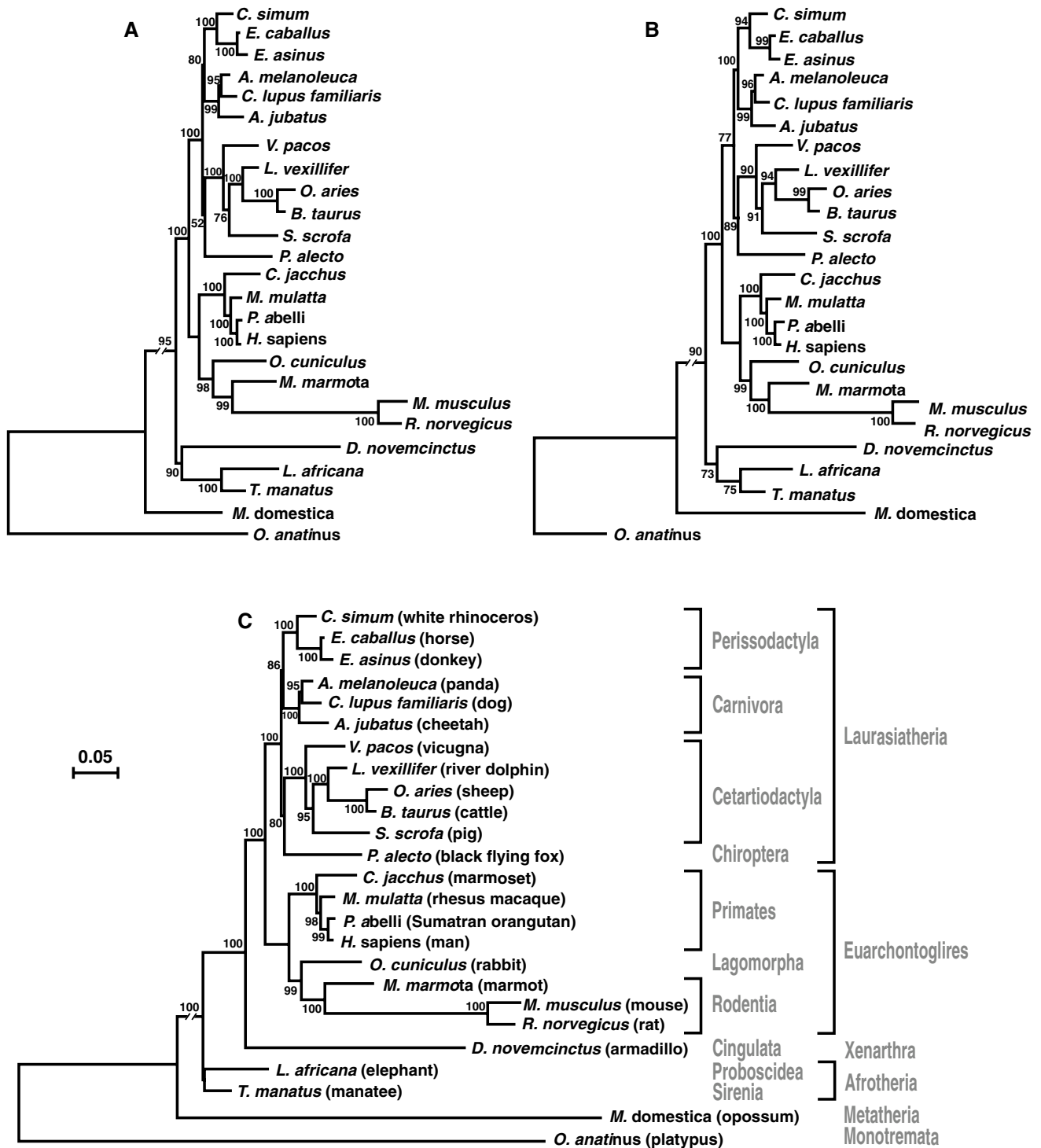
share homology with the Z chromosome of birds (Veyrunes et al. 2008).

In monotremata and metatherians, the ancestral region contains only one MAGE gene, but in the eutherians, there are two: MAGED2 and TRO. Studies have found that chorionic gonadotropin secreted from the preimplantation embryo up-regulates TRO expression by the uterine epithelium, in preparation for the attachment reaction (Sugihara et al. 2008). Furthermore, homophilic binding of TRO during the attachment reaction initiates downstream signaling that differentially alters the physiological state of each cell type, setting the stage for subsequent invasion of the uterus by EvT cells (Armant 2011). The important function of TRO in the implantation of the embryo and the formation of placenta (Tamura et al. 2011) suggests that the TRO function has been derived more recently in eutherian evolution, as suggested by Katsura and Satta (2011). This, and the fact that our phylogenetic analysis consistently showed MAGED2 as being closer to the ancestral MAGE gene, makes this gene the most likely ancestral candidate for all the type II genes. Katsura and Satta (2011) suggested that MAGED2 is probably the ancestor of eutherian MAGE genes, but their analysis could not conclude whether MAGED2 or TRO was the ancestral gene.

Consistent with our results here, Katsura and Satta (2011) have shown that there is significant sequence similarity within and between the type I and II genes but only in the CDS regions, suggesting that retrotransposition events have produced many of the genes. However, the

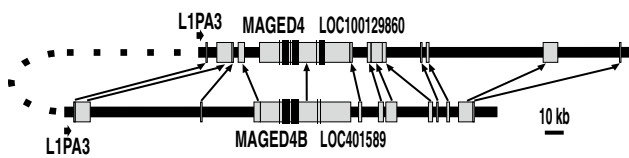
considerable breakage of recently duplicated MAGED4 and MAGED4B genes in humans suggests that the non-coding regions surrounding MAGE genes can undergo rapid diversification, making it very difficult to differentiate between retrotransposition and other types of duplication events, just by the divergence of the surrounding sequences. In this regard, when searching for MAGED4 and MAGED4B in the genome of the ancient Denisovan hominid (Meyer et al. 2012) and the Neanderthal (Prüfer et al. 2014), using the UCSC Genome Browser, we found both regions conserved, containing both genes, although the low coverage of these genomes do not allow for sequence comparison of the repetitive elements found. The divergence time of Neanderthals and Denisovans from modern humans from *Homo heidelbergensis* was estimated to have occurred 550–765 Kya (Prüfer et al. 2014). Thus, MAGED4/D4B were duplicated after the divergence of humans and chimpanzees but before the divergence of these hominoids.

Due to difficulties in studying phylogenetic relationships of major groups such as those underlying macroevolutionary processes and complex patterns of gene family evolution, new statistical analyses are moving away from simple parametric models and stepwise approaches towards integrative models (Lartillot 2015). Phylogenetic trees constructed using LG4X and Bayesian models produced higher levels of branch support, compared to simple parametric models such as neighbor-joining or UPGMA (not shown) and ML with gamma distribution. LG4X leaves aside the gamma distribution because it uses four different

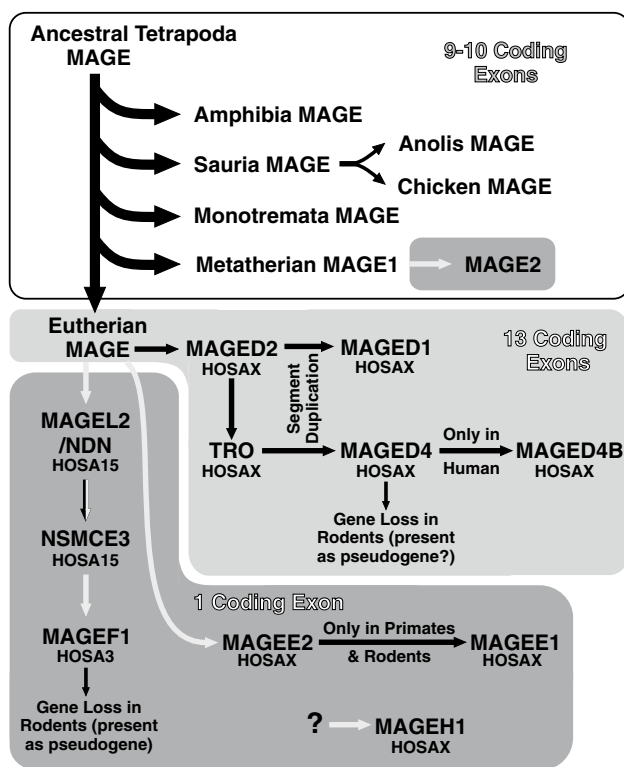


**Fig. 6** Mammalian phylogeny based on the complete sequence of 4 type II MAGE proteins. Phylogenetic relationship among 23 mammalian species, based on the entire, concatenated protein sequences combined of MAGED1, MAGED2, MAGEL2, and NDN. Phylogenetic analyses were carried out as described in the legend of Fig. 3 and in methods. The analysis involved 23 amino acid sequences,

with a total of 2864 positions in the final dataset. Platypus MAGE and opossum MAGE1 were used as outgroups since they are the most ancient mammalian MAGE genes. The branch distances in all trees for platypus and opossum, with respect to eutherians, were not drawn to scale because of their length



**Fig. 7** Sequence conservation in the human MAGED4–MAGED4B region. Sequence conservation of the region containing the MAGED4 and MAGED4B genes, using mVista software (<http://genome.lbl.gov/vista/index.shtml>). Gray boxes are regions showing >70% identity conservation, black, vertical lines represent annotated exons and those horizontal correspond to non-conserved genomic sequences. Black thick arrows show the direction of the LINE1 elements at the suggested duplication point, while the thin arrows show the homologous pair of conserved blocks. The dotted line is just showing the connection of both sides of the region



**Fig. 8** Evolutionary history of the type II MAGE gene family. Most likely evolution of the eutherian type II MAGE genes based on their phylogenetic relations, genome location, homology conservation, and the protein and gene structures. In eutherians and metatherians, the light gray arrows imply duplication by retrotransposition and the black arrows imply duplication by other means. Human chromosomes 3, 15, and X are denoted as HOSA3, HOSA15, and HOSA3, respectively

matrices, following a distribution-free scheme for the site rates (Le et al. 2008; Le and Gascuel 2010). The complexity of amino acid substitutions has been shown by analyzing different datasets, which makes flexible models such as LG4M and LG4X more suitable (Le et al. 2012), both

significantly outperforming single-matrix models, and providing gains of dozens to hundreds of log-likelihood units for most datasets. LG4X obtains substantial gains compared with LG4M, thanks to its distribution-free scheme for site rates. Le et al. (2012) suggests that LG4X is a good alternative to single replacement matrices, since it displays such advantages but require the same memory space and have comparable running times to standard models.

Mammalian phylogeny reconstruction inferred using protein sequences of the most informative MAGE genes is very much like the recently published studies that were inferred using large datasets (Prasad et al. 2008; Lindblad-Toh et al. 2011). In these studies, elephants diverged very early in eutherian diversification and rodents are the closest relatives to primates, even though they show very rapid diversification compared to other groups. In addition, both trees also show Perissodactyla closely related to carnivores and both to Cetartiodactyla, agreeing with the proposed group Ferungulata, which allegedly contains Pholidota, Carnivora, Perissodactyla, and Cetartiodactyla (Zhou et al. 2012). Prasad et al.'s (2008) results illustrate the difficulty in resolving some branches even with large amounts of data. An alternative to using large datasets is to use sequence data that are most informative in those problematic branches, such as the early divergent eutherian clades. Thus, the use of genes such as those of the type II MAGE family as evolutionary markers, which have diverged early in the history of eutherians, is an alternative to resolve the phylogenetic relationships among the basal mammalian clades such as Euarchontoglires, Laurasiatheria (within Boreoeutheria), Atlantogenata, and Metatheria. In this sense, Salichos and Rokas (2013) suggest that selecting genes with strong phylogenetic signals and demonstrating the absence of significant incongruence are essential for accurately reconstructing ancient divergences.

**Acknowledgements** This work was supported by the College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, and Pfizer Animal Health (now Zoetis, Inc.). Additional support by National Research Initiative Competitive Grant Program (Grant No. 2006-35205-16864) from the USDA National Institute of Food and Agriculture, USDA-NIFA Research Agreements (Nos. 2009-65205-05635, 2010-34444-20729), and USDA Federal formula Hatch funds appropriated to the Cornell University Agricultural Experiment Station are gratefully acknowledged. We thank the Higher Education Commission of Pakistan for a Visiting Fellowship awarded to TH.

## References

- Armant DR (2011) Life and death responses to trophinin-mediated adhesion during blastocyst implantation. *Cell Cycle* 10:574–575
- Atanackovic D, Altorki NK, Stockert E, Williamson B, Jungbluth AA, Ritter E, Santiago D, Ferrara CA, Matsuo M, Selvakumar A, Dupont B, Chen YT, Hoffman EW, Ritter G, Old LJ, Gnajatic

- S (2004) Vaccine-induced CD4 + T cell responses to MAGE-3 protein in lung cancer patients. *J Immunol* 172:3289–3296
- Chomez P, De Backer O, Bertrand M, De Plaen E, Boon T, Lucas S (2001) An overview of the MAGE gene family with the identification of all human members of the family. *Cancer Res* 61:5544–5551
- Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14:1188–1190
- Delbridge ML, Patel HR, Waters PD, McMillan DA, Marshall Graves JA (2009) Does the human X contain a third evolutionary block? Origin of genes on human Xp11 and Xq28. *Genome Res* 19:1350–1360
- Doyle JM, Gao J, Wang J, Yang M, Potts PR (2010) MAGE-RING protein complexes comprise a family of E3 ubiquitin ligases. *Mol Cell* 39(6):963–974
- Dubchak I, Ryaboy DV (2006) VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods Mol Biol* 338:69–89
- Feng Y, Gao J, Yang M (2011) When MAGE meets RING: insights into biological functions of MAGE proteins. *Protein Cell* 2:7–12
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321
- Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282
- Katsura Y, Satta Y (2011) Evolutionary history of the cancer immunity antigen MAGE gene family. *PLoS ONE* 6:e20365
- Lartillot N (2015) Probabilistic models of eukaryotic evolution: time for integration. *Philos Trans R Soc Lond B Biol Sci* 370(1678):20140338
- Le SQ, Gascuel O (2010) Accounting for accessibility to solvent and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol* 59:277–287
- Le SQ, Gascuel O, Lartillot N (2008) Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323
- Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol* 29:2921–2936
- Lee AK, Potts PR (2017) A comprehensive guide to the MAGE family of ubiquitin ligases. *J Mol Biol* 429:1114–1142
- Lindblad-Toh K, Garber M, Zuk O et al (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482
- Liu Y, Yang S, Yang J, Que H, Liu S (2012) Relative expression of type II MAGE genes during retinoic acid-induced neural differentiation of mouse embryonic carcinoma P19 cells: a comparative real-time PCR analysis. *Cell Mol Neurobiol* 32:1059–1068
- López-Sánchez N, González-Fernández Z, Niinobe M, Yoshikawa K, Frade JM (2007) Single mage gene in the chicken genome encodes CMage, a protein with functional similarities to mammalian type II Mage proteins. *Physiol Genomics* 30:156–171
- Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Dereviako AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paabo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226
- Nishimura I, Shimizu S, Sakoda JY, Yoshikawa K (2007) Expression of Drosophila MAGE gene encoding a necdin homologous protein in postembryonic neurogenesis. *Gene Expr Patterns* 7(3):244–251
- Osterlund C, Töhönen V, Forslund KO, Nordqvist K (2000) Mage-b4, a novel melanoma antigen (MAGE) gene specifically expressed during germ cell differentiation. *Cancer Res* 60:1054–1061
- Prasad AB, Allard MW, NISC Comparative Sequencing Program, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol* 25:1795–1808
- Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwillm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Dereviako AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49
- Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61:539–542
- Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–331
- Sang M, Wang L, Ding C, Zhou X, Wang B, Wang L, Lian Y, Shan B (2011) Melanoma-associated antigen genes—an update. *Cancer Lett* 302:85–90
- Sigrift CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 38(Database issue):D161–D166
- Sugihara K, Kabir-Salmani M, Byrne J, Wolf DP, Lessey B, Iwashita M, Aoki D, Nakayama J, Fukuda MN (2008) Induction of trophinin in human endometrial surface epithelia by CGbeta and IL-1beta. *FEBS Lett* 582:197–202
- Tamura N, Sugihara K, Akama TO, Fukuda MN (2011) Trophinin-mediated cell adhesion induces apoptosis of human endometrial epithelial cells through PKC-δ. *Cell Cycle* 10:135–143
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30:2725–2729
- Taylor EM, Copey AC, Hudson JJ, Vidot S, Lehmann AR (2008) Identification of the proteins, including MAGEG1, that make up the human SMC5-6 protein complex. *Mol Cell Biol* 28:1197–1206
- van der Bruggen P, Traversari C, Chomez P, Lurquin C, De Plaen E, Van den Eynde B, Knuth A, Boon T (1991) A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* 254(5038):1643–1647
- Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Grützner F, Deakin JE, Whittington CM, Schatzkammer K, Kremitzki CL, Graves T, Ferguson-Smith MA, Warren W, Marshall Graves JA (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res* 18:965–973
- Zhou X, Xu S, Xu J, Chen B, Zhou K, Yang G (2012) Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. *Syst Biol* 61:150–164