Immunogenetics (2019) 71:25–33
https://doi.org/10.1007/s00251-018-1079-x

ORIGINAL ARTICLE

# The evolution of S100A7 in primates: a model of concerted and birth-and-death evolution

Ana Águeda-Pinto[1,2] · Pedro José Esteves[1,2,3]

Received: 11 June 2018 / Accepted: 22 August 2018 / Published online: 29 August 2018

## Abstract

The human S100A7 resides in the epidermal differentiation complex (EDC) and has been described as a key effector of innate immunity. In humans, there are five *S100A7* genes located in tandem—S100A7A, S100A7P1, S100AL2, S100A7, and S100AP2. The presence of several retroelements in the S100A7A/S100A7P1 and S100A7/S100A7P2 clusters suggests that these genes were originated from a duplication around ~35 million years ago, during or after the divergence of Platyrrhini and Catarrhini primates. To test this hypothesis, and taking advantage of the high number of genomic sequences available in the public databases, we retrieved *S100A7* gene sequences of 12 primates belonging to the Cercopithecoidea and Hominoidea (Catarrhini species). Our results support the duplication theory, with at least one gene of each cluster being identified in both Cercopithecoidea and Hominoidea species. Moreover, given the presence of an ongoing gene conversion event between *S100A7* and *S100A7A*, a high rate of mutation in *S100A7L2* and the presence of pseudogenes, we proposed a model of concerted and birth-and-death evolution to explain the evolution of *S100A7* gene family. Indeed, our results suggest that *S100A7L2* most likely suffered a neofunctionalization in the Catarrhini group. Being S100A7 a major protein in innate defense, we believe that our findings could open new doors in the study of this gene family in immune system.

**Keywords** S100A7 (psoriasin) · Primates · Model of concerted and birth-and-death evolution

## Introduction

The S100 protein family is one of the largest subfamilies of the EF-hand type (helix E-loop-helix F motif) $Ca^{2+}$-binding proteins (Denessiouk et al. 2014). These small proteins exhibit a common structure consisting of two EF-hand-binding domains. The C-terminal EF-hand contains the canonical $Ca^{2+}$-binding loop, whereas the N-terminal EF-hand has a unique, non-canonical S100-specific domain with weaker calcium affinity (Ishikawa et al. 2000; Moroz et al. 2000). Although S100 proteins exhibit a high degree of structural similarity, they present highly specific expression patterns in mammals and a remarkable degree of cell and tissue specificity, which make them non-functionally interchangeable (Donato et al. 2013; Xia et al. 2018). This different spatial-temporal distribution of S100 proteins along with their ability to target different proteins (Santamaria-Kisiel et al. 2006) allows cells to transduce a $Ca^{2+}$ signal into a unique biological response. The human genome encodes more than 20 S100 protein families, 17 of which are assigned to chromosomal band 1q21, in the so-called epidermal differentiation complex (EDC), a cluster of about 25 genes involved in differentiation of the human epidermis (Engelkamp et al. 1993; Marenholz et al. 2004; Zimmer et al. 2013).

*S100A7*, also called psoriasin, is a member of the S100 gene family located on the EDC cluster, first identified as an upregulated protein in psoriatic skin lesions (Madsen et al. 1991). Outside cells, S100A7 proteins are considered to have an important role in inflammation processes, acting as chemotactic factors for $CD4^+$ T leukocytes and keratinocytes (Ruse et al.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00251-018-1079-x) contains supplementary material, which is available to authorized users.

✉ Pedro José Esteves
pjesteves@cibio.up.pt

[1] CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos genéticos, Universidade do Porto, Rua Padre Armando Quintas, 4485-661 Vairão, Portugal

[2] Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[3] Instituto de Investigação e Formação Avançada em Ciências e Tecnologias da Saúde (CESPU), Gandra, Portugal

2003; Tan et al. 1996). Moreover, it has also been implicated as a host-defense protein that adheres to and reduces *Escherichia coli* survival in a $Zn^{2+}$-dependent manner (Gläser et al. 2005). Indeed, mutagenesis studies showed that inactivation of the zinc-binding motif reduces antibacterial activity, while the calcium-binding EF-hand is not required for this function (Lee and Eckert 2007). In an intracellular context, S100A7 has been implicated in several cell tumor subtypes (Liu et al. 2015; Salama et al. 2008; Webb et al. 2005).

The S100A7 genomic region showed evidences of recent duplications (Zimmer et al. 2013) and, in humans, genomic analysis revealed a total of five gene copies dispersed by three distinct regions within a continuous 75-kb genomic region, with two of these copies (S100A7P1 and S100A7P2) being fragmented and therefore classified as pseudogenes (Kulski et al. 2003). Regions 1 and 3 contain both an intact gene and a pseudogene: S100A7A and S100A7P1 in region 1, and S100A7 S100A7P2 in region 3. These two duplicated regions are separated by an 11-kb intergenic region (region 2) that has only one S100A7-like gene (S100A7L2). The sharing of five Alu subfamily members between regions 1 and 3 might indicate that duplication occurred during or after AluS amplification, ~ 31–44 million years ago (Mya) (Kulski et al. 2003). This places the duplication of *S100A7* genes during or after the divergence of New World monkeys (Platyrrhini) and Catarrhini (~ 35 Mya) (Schrago and Russo 2003) and, for that reason, it is expected that Cercopithecoidea and Hominoidea species have all five *S100A7* copies. Gene duplication is a major mechanism through which new genetic material is generated, and for this reason, it can facilitate species adaptation (Magadum et al. 2013). Some gene functions had been acquired following gene duplication which has remarkably contributed for species adaptation, having been implicated, for example, in the evolution of the immune response and efficient protein synthesis (Otto and Yong 2002).

In this work, we analyzed *S100A7* gene sequences of Cercopithecidae (*Macaca mulatta*, *M. fascicularis*, *M. nemestrina*, *Papio anubis*, *Rhinopithecus bieti*, and *Colobus angolensis*) and Hominoidea (*Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo abelii*, and *Nomascus leucogenys*) from available databases using both comparative genomics and phylogenetic methods. We present a model of the molecular evolution of the *S100A7* primates' gene, which contributes to a better understanding of the mechanisms of evolution of the S100A7 gene family.

## Methods

### Sequences retrieval

All the available primate *S100A7*, *S100A7A*, and *S100A7L2* gene coding sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.gov) and Ensembl (http://www.ensembl.org/index.html) databases. Sequences were aligned with Clustal W (Thompson et al. 1994), implemented in BioEdit v7.2.6.1 (Hall 1999), followed by visual inspection. Nucleotide sequence translation into amino acids was also performed using BioEdit. The *S100A7* sequences collected belong to superfamily Cercopithecoidea (family Cercopithecidae: genera *Macaca*, *Papio*, *Colobus*, *Rhinopithecus*) and superfamily Hominoidea (family Hominidae: genera *Gorilla*, *Homo*, *Pan*, and *Pongo*; family Hylobatidae: genus *Nomascus*). In order to clarify the relationships within the obtained sequences, *S100A7* genes of two species of New World Monkeys (*Cebus capucinus* and *Saimiri boliviensis*) were included as an outgroup. The complete list of all *S100A7* sequences used in this study, together with gene names and accession numbers, is given in Table 1.

### Mapping of the *S100A7* pseudogenes

*S100A7P1* and *S100A7P2* were only found in human databases (Table 1). For this reason, *S100A7P1* and *S100A7P2* human pseudo exon sequences (NCBI Reference Sequences NG_009595.1 and NG_009592.1, respectively) were used to detect related sequences by performing a BLAST search in GenBank (NCBI, http://BLAST.ncbi.nlm.nih.gov/) and Ensembl (http://www.ensembl.org/Multi/blastview) databases against the Catarrhini species listed in Table 1.

For each species, only the fragment with the highest sequence similarity and the lower E-value was used to perform an alignment with *S100A7P1* and *S100A7P2* from humans (Supplementary Figs. 1 and 2). Nevertheless, sequences with less than 85% of sequence similarity and E-values > $1e^{-90}$ were not considered.

### Evolutionary analyses

In order to infer the phylogenetic relationships of *S100A7* genes in primates, evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016) using a maximum likelihood (ML) method based on the Tamura-Nei model (Tamura and Nei 1993). The reliability of the clusters was tested by performing the bootstrap test of phylogeny, with 1000 bootstrap replications. ML trees were displayed using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/). To increase the reliability of the phylogenetic analysis, partial gene sequences were excluded (Mamu_S100A7A and Poab_S100A7A), producing a final dataset of 31 sequences. Moreover, no pseudogenes were included in the phylogenetic analysis.

The number of amino acid differences per site between primates' S100A7, S100A7A, and S100A7L2 sequences was estimated using MEGA7 (Kumar et al. 2016). For that, four groups were created: Homininae_S100A7 (Hosa_S100A7, Patr_S100A7, Papa_S100A7, and

**Table 1** List of the sequences of the *S100A7* genes, available from NCBI and Ensembl databases and used in this study

| Common name | Species name | Gene | Accession no. | Abbreviation |
|---|---|---|---|---|
| Human | *Homo sapiens* | S100A7 | ENST00000368723 | Hosa_S100A7 |
| | | S100A7A | ENST00000368729 | Hosa_S100A7A |
| | | S100A7L2 | ENST00000368725 | Hosa_S100A7L2 |
| | | S100A7P1 | NG_009595.1 | Hosa_S100A7P1 |
| | | S100A7P2 | NG_009592.1 | Hosa_S100A7P2 |
| Chimpanzee | *Pan troglodytes* | S100A7 | ENSPTRG00000001350 | Patr_S100A7 |
| | | S100A7A | ENSPTRG00000001349 | Patr_S100A7A |
| | | S100A7L2 | ENSPTRG00000023699 | Patr_S100A7L2 |
| Bonobo | *Pan paniscus* | S100A7 | ENSPPAG00000042606 | Papa_S100A7 |
| | | S100A7A | ENSPPAG00000034917 | Papa_S100A7A |
| | | S100A7L2 | ENSPPAG00000040264 | Papa_S100A7L2 |
| Gorilla | *Gorilla gorilla* | S100A7 | ENSGGOG00000024101 | Gogo_S100A7 |
| | | S100A7A | ENSGGOG00000040685 | Gogo_S100A7A |
| | | S100A7L2 | ENSGGOG00000036463 | Gogo_S100A7L2 |
| Orangutan | *Pongo abelii* | S100A7A | ENSPPYG00000000812 | Poab_S100A7A |
| | | S100A7L2 | ENSPPYG00000000811 | Poab_S100A7L2 |
| Gibbon | *Nomascus leucogenys* | S100A7 | ENSNLEG00000010395 | Nole_S100A7 |
| | | S100A7A | ENSNLEG00000010398 | Nole_S100A7A |
| Olive baboon | *Papio anubis* | S100A7A | ENSPANT00000001109 | Paan_ S100A7A |
| | | S100A7L2 | ENSPANT00000024566 | Paan_ S100A7L2 |
| Rhesus macaque | *Macaca mulatta* | S100A7 | ENSMMUG00000047853 | Mamu_S100A7 |
| | | S100A7A | ENSMMUG00000004833 | Mamu_S100A7A |
| | | S100A7L2 | ENSMMUT00000006829 | Mamu_S100A7L2 |
| Crab-eating macaque | *M. fascicularis* | S100A7 | ENSMFAG00000033466 | Mafa_S100A7 |
| | | S100A7A | ENSMFAG00000035481 | Mafa_S100A7A |
| | | S100A7L2 | ENSMFAG00000036788 | Mafa_S100A7L2 |
| Pig-tailed macaque | *M. nemestrina* | S100A7A | ENSMNEG00000029655 | Mane_S100A7A |
| | | S100A7L2 | ENSMNEG00000027268 | Mane_S100A7L2 |
| Black snub-nosed monkey | *Rhinopithecus bieti* | S100A7 | ENSRBIG00000036259 | Rhbi_S100A7 |
| | | S100A7A | ENSRBIG00000043734 | Rhbi_S100A7A |
| Angola colobus | *Colobus angolensis* | S100A7L2 | ENSCANG00000041573 | Coan_S100A7L2 |
| Capuchin | *Cebus capucinus* | S100A7 | ENSCCAG00000028661 | Ceca_S100A7 |
| Bolivian squirrel monkey | *Saimiri boliviensis* | S100A7 | ENSSBOG00000018328 | Sabo_S100A7 |

Gogo_S100A7), Homininae_S100A7A (Hosa_S100A7A, Patr_S100A7A, Papa_S100A7A, and Gogo_S100A7A), Homininae_S100AL2 (Hosa_S100A7L2, Patr_S100A7L2, Papa_S100A7L2, and Gogo_S100A7L2), and Platyrrhini_S100A7 (Ceca_S100A7 and Sabo_S100A7).

Tajima's relative test was conducted in the primates' sequences used in the phylogenetic tree in order to evaluate statistical significance in molecular evolution of the duplicated genes. The statistical parameters were set as default using MEGA7 (Kumar et al. 2016). A *P value* < 0.05 was used to reject the null hypothesis of equal rates between the three lineages considered simultaneously (sequence A, sequence B, and outgroup).

## Results and discussion

### Data retrieving

A total of 31 *S100A7* genomic sequences (one *S100A7P1*, one *S100A7P2*, eight *S100A7*, eleven *S100A7A*, and ten *S100A7L2*) of 12 different Cercopithecoidea and Hominoidea species were retrieved (Table 1). We found five *S100A7* sequences in humans, two of which presenting premature stop codons, and therefore appear to be noncoding (data not shown). These findings are in agreement with previous studies that reported five copies of the *S100A7* genes in the human genome, with *S100A7P1* and *S100A7P2* being non-functional (Kulski et al. 2003). As a first step in establishing the evolutionary history of the *S100A7* gene family after the divergence of Platyrrhini and Catarrhini, S100A7P1 and S100A7P2 genes were searched in available databases for Catarrhini species. However, we could not find these genes described for any species (except for *H. sapiens*).

As a result of advances in genome sequencing technologies, new genomic data for the great apes, macaques, and other nonhuman primates become sequenced allowing a new opportunity to reconstruct the evolutionary history and the genetic divergence of *S100A7* genes among these evolutionary lineages (Rogers 2013; Rogers and Gibbs 2014). In the past years, an effort has been made in order to improve the genomic dataset of several primates, detecting highly fragmented areas that could result in several gaps. As a result, several genomes of primates have been sequenced and assembled at a high level of quality (Gordon et al. 2016; Kuderna et al. 2017; Norgren 2013; Yan et al. 2011). Most of these primates' genomes are based on a whole genome sequencing (WGS) technique and are available in online databases. In order to find evidences of *S100A7P1* and *S100A7P2* genes in Catarrhini species, we performed a BLAST search of *S100A7P1* and *S100A7P2* human sequences against these primates' genomes. Despite these genes were not described for Catarrhini species, we were able to map partial sequences of *S100A7P1* and *S100A7P2* in all Cercopithecoidea and Hominoidea species (Supplementary Figs. 1 and 2). No partial sequences for these two genes were found in any Platyrrhini species. Our results show that partial *S100A7P1* and *S100A7P2* sequences are mostly mapped in chromosome 1, within a continuous genomic sequence where *S100A7*-duplicated genes are located (Kulski et al. 2003); in *N. leucogenys*, all members of *S100A7* gene family are present in chromosome 12. Figure 1 resumes all the obtained *S100A7* genes for Primate species. Given the dataset collected for Catarrhini and Platyrrhini species, we believe that our results support the theory that a duplication event in *S100A7* genes occurred ~ 35 Mya.

## Phylogenetic analysis

In order to understand the relationships and evolutionary history of *S100A7* family members, we performed a phylogenetic analysis, using the maximum likelihood (ML) method (for details, see "Methods" section) (Fig. 2). Full-length coding sequences of primates *S100A7*-duplicated genes were analyzed together with known human *S100A7* genes (*S100A7*, *S100A7A*, *S100A7L2*). *M. mulatta* and *P. abelii* have partial gene sequences for *S100A7A* (Mamu_S100A7A and Poab_S100A7A), and for this reason, these incomplete sequences were not used.

From the ML tree, two major clusters can be distinguished, supported by high bootstrap values (Fig. 2). One of the clusters includes all *S100A7L2* sequences, with this gene appearing in a highly supported basal clade, matching the accepted primate phylogenetic relationships. The other cluster comprises the *S100A7* and *S100A7A* retrieved sequences and, despite they did not form clearly separated groups, it is possible to observe that *S100A7* and *S100A7A* clustered by gene in all Homininae species and by species in the remaining species. In fact, all Homininae *S100A7* and *S100A7A* sequences group in a gene-specific manner forming two well-separated groups (bootstrap values of 85), also matching the accepted primate phylogeny. However, this was not observed for the remaining species (Fig. 2). For example, *S100A7* and *S100A7A* sequences of *N. leucogenys* clustered with the well-supported *S100A7* group of Homininae. This was also the case for *M. mulatta*, *M. fascicularis*, and *R. bieti* genes, but here, *S100A7* and *S100A7A* sequences clustered along with the *S100A7A* Homininae group. The clustering of all *S100A7* and *S100A7A* genes of Cercopithecoidea, *N. leucogenys*, and *R. bieti* species by species and not by gene might be parsimoniously explained by an ongoing gene-conversion process between these two genes, which probably started after the divergence of Cercopithecoidea and Hominoidea families. From a closer look between S100A7 and

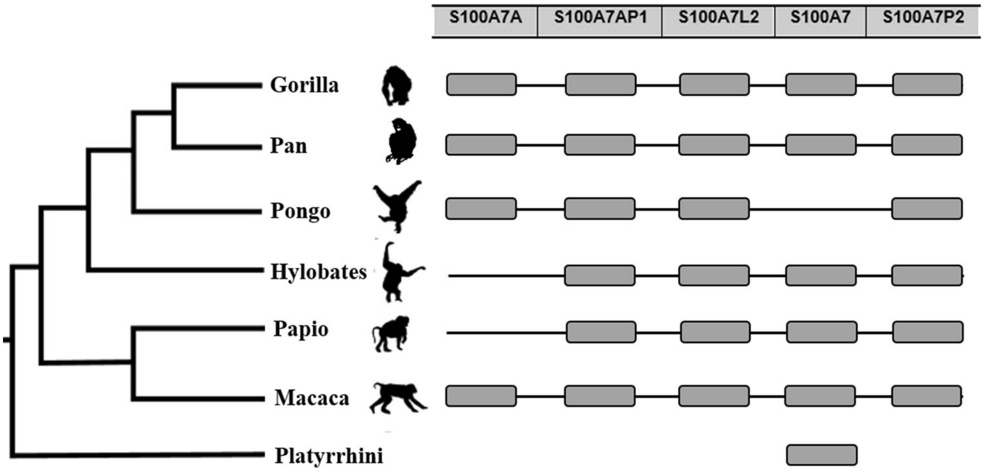

**Fig. 1** Resume of *S100A7* duplicated genes found in public databases for Catarrhini and Platyrrhini species

S100A7A sequences (Fig. 3, segment A and B), it is possible to observe at least two events of gene conversion between these genes. Interestingly, this is observed in both directions: in *N. leucogenys*, *S100A7* is converting the *S100A7A* gene, and for *M. mulatta*, *M. fascicularis*, and *R. bieti*, it is *S100A7A* gene that is converting *S100A7* (Fig. 3, segment A and B), helping us to understand why *N. leucogenys* genes clustered closer to S100A7 Homininae group and why the remaining Cercopithecoidea species clustered closer to S100A7A Homininae group. Moreover, the fact that these genes are located within a continuous 75-kb genomic region in the same chromosome prompts our hypothesis of gene conversion between them (Kulski et al. 2003).

Contrarily to what was expected, *S100A7* gene of Platyrrhini species (used as an outgroup) group in the same node as S100A7 and S100A7A sequences of Catarrhini species, leaving *S100A7L2* in a separated node (bootstrap values of 100). From this observation, two different hypotheses can be proposed to explain our results: *S100A7L2* gene can be a result of a duplication event of *S100A7* that occurred in ancestral primates and, for some reason, was lost/not found in New World Monkeys, or, on the other hand, this observation could indicate that *S100A7L2* is suffering a high rate of mutation, splitting this gene family in very distinct proteins. In the light of our current knowledge, the first hypothesis seems unlikely. The presence of several *Alu* members between the regions

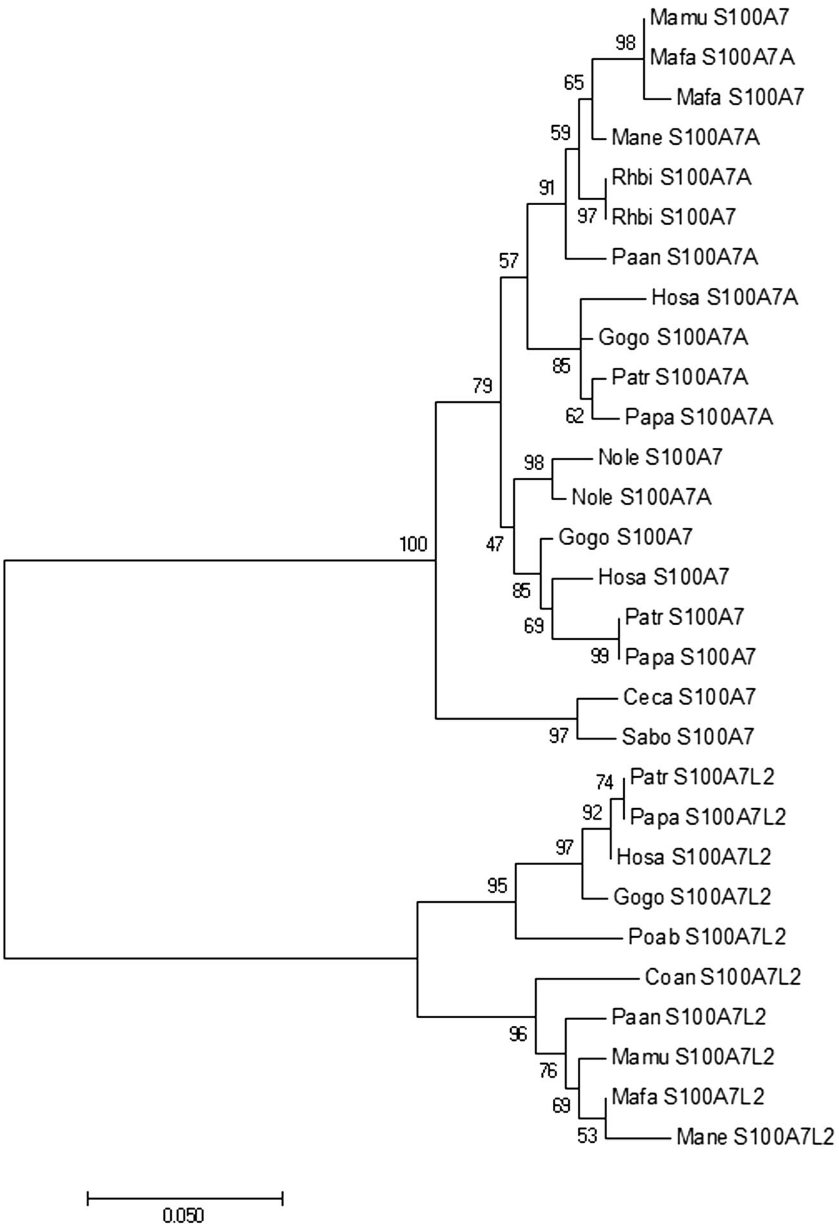

**Fig. 2** Maximum likelihood tree for the primate S100A7 genes. The analyses were performed with 1000,000 generations and 1000 bootstrap searches. Bootstrap values (%) are indicated on the branches. The abbreviations correspond to the following species: Hosa—*H. sapiens*; Patr—*P. troglodytes*; Papa—*P. paniscus*; Gogo—*G. gorilla*; Poab—*P. abelii*; Nole—*N. leucogenys*; Paan—*P. anubis*; Mamu—*M. mulatta*; Mafa—*M. fascicularis*; Mane—*M. nemestrina*; Rhbi—*R. bieti*; Coan—*C. angolensis*; Ceca—*Cebus capucinus*; Sabo—*Saimiri boliviensis*

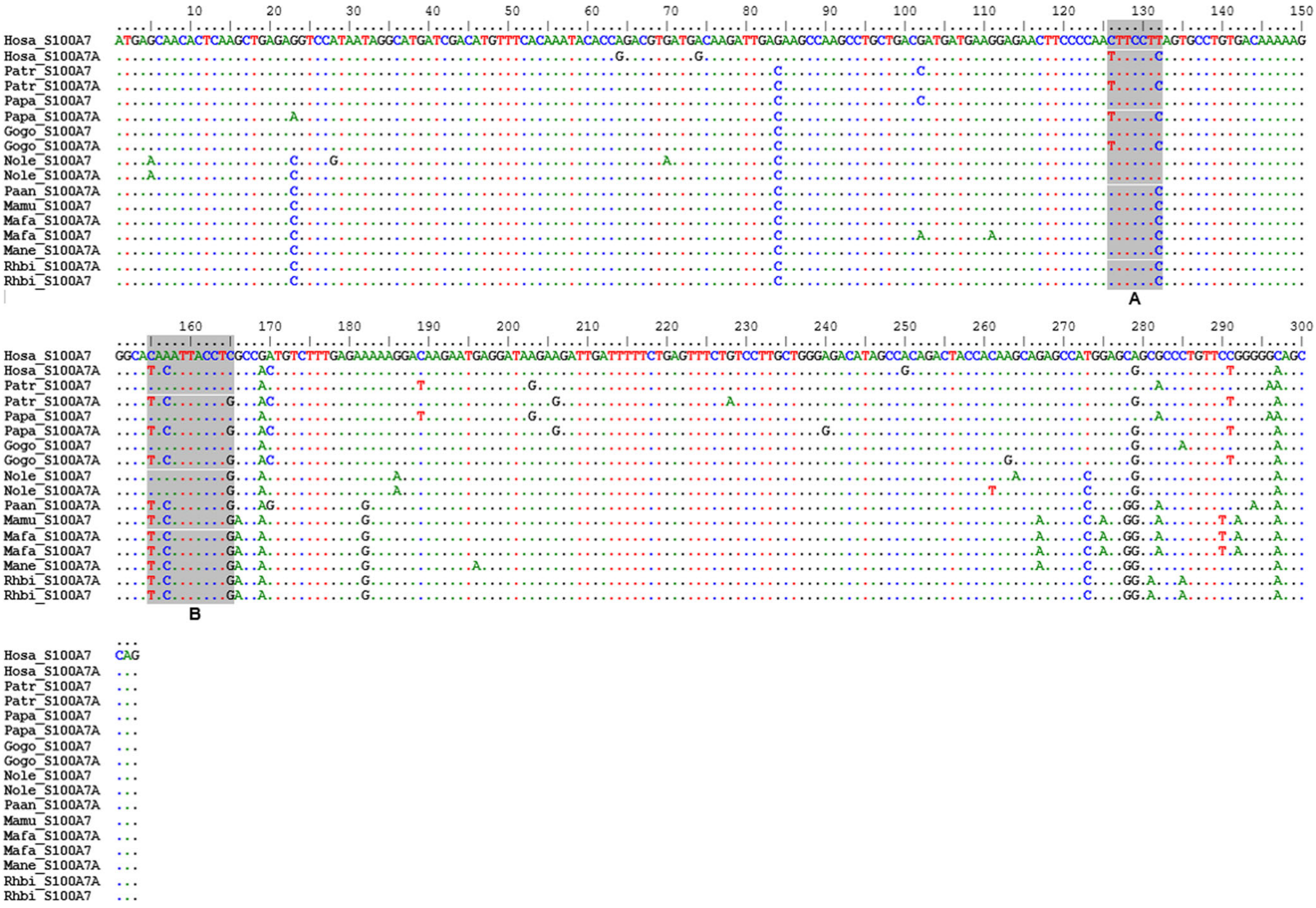

**Fig. 3** Nucleotide alignment of *S100A7* duplicated genes. The homologous region between S100A7 and S100A7A that may have resulted from gene conversion is highlighted in light gray (segment A and B). The abbreviations correspond to the following species: Hosa—*H. sapiens*; Patr—*P. troglodytes*; Papa—*P. paniscus*; Gogo—*G. gorilla*; Nole—*N. leucogenys*; Paan—*P. anubis*; Mamu—*M. mulatta*; Mafa—*M. fascicularis*; Mane—*M. nemestrina*; Rhbi—*R. bieti*. Dots = identity with Hosa_S100A7 sequence

where this gene family is located indicates that these genes were originated after the divergence of Platyrrhini and Catarrhini (~ 35 Mya). Moreover, from the phylogenetic tree (Fig. 2), even that Platyrrhini *S100A7* gene group closer to S100A7 and S100A7A from Catarrhini species, they are in a completely separate node (high bootstrap value, 97), suggesting that these genes diverged a long time ago.

The evolutionary divergence between these genes was studied and the number of amino acid differences per site between four different groups (Homininae_S100A7, Homininae_S100A7A, Homininae_S100AL2, and Platyrrhini_S100A7) was estimated (Table 2). For this analysis, we only used Homininae sequences since we did not detect evidences of an ongoing gene-conversion. The evolutionary divergence between Homininae_S100A7 and Homininae_S100A7A groups was very similar (3.4%) and, as expected, higher distance values were registered for Platyrrhini_S100A7 when compared to Homininae_S100A7 and Homininae_S100A7A (9 and 12%, respectively). Strikingly, in Homininae_S100AL2 proteins, the amino acid differences reached almost 50% when compared to S100A7 and S100A7A from Homininae species. These results suggest that in all Homininae primates, this gene is suffering a high rate of mutation. To test for statistical differences in the molecular evolution of this gene, a Tajima's relative rate test was performed (Table 3). In all analysis, the *P* values were found to be significant and the null hypothesis of equal mutation

**Table 2** Estimates of evolutionary divergence between S100A7 and S100A7A amino acid sequences

| Groups | | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Homininae_S100A7 | – | – | – | – |
| Homininae_S100A7A | 0.034 | – | – | – |
| Plathyrrhini_S100A7 | 0.090 | 0.116 | – | – |
| Homininae_S100A7L2 | 0.450 | 0.456 | 0.417 | – |

Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016). All positions containing gaps and missing data were eliminated. Options used: amino acid distance, complete deletion, and *p*-distance

**Table 3** Results of Tajima's relative rate test

| Sequence A | Sequence B | Outgroup | Divergent sites | Sequence A specific | Sequence B specific | Outgroup specific | $X^2$ test | *P* value |
|---|---|---|---|---|---|---|---|---|
| *Hosa_S100A7L2* | *Hosa_S100A7* | *Ceca_S100A7* | 2 | 62 | 11 | 12 | 35.63 | 0.0000 |
| *Patr_S100A7L2* | *Patr_S100A7* | *Ceca_S100A7* | 2 | 60 | 10 | 15 | 35.71 | 0.0000 |
| *Papa_S100A7L2* | *Papa_S100A7* | *Ceca_S100A7* | 2 | 60 | 10 | 15 | 35.71 | 0.0000 |
| *Gogo_S100A7L2* | *Gogo_S100A7* | *Ceca_S100A7* | 2 | 62 | 8 | 12 | 41.66 | 0.0000 |

The statistical parameters were set as default using MEGA7 (Kumar et al. 2016). A *P* value < 0.05 was used to reject the null hypothesis of equal rates between the 3 lineages considered simultaneously (sequence A, sequence B, and outgroup)

rates was rejected. From this simple test, it was possible to confirm that the substitution rate of *S100A7L2* gene is significantly high, with mutation rates being up to five times higher in comparison to *S100A7* sequences of Homininae. According to the neofunctionalization model, after a duplication event arise, the duplicated copy can achieve fixation in the genome and be free to evolve and potentially acquire new gene functions (Assis and Bachtrog 2013; Rastogi and Liberles 2005; Teshima and Innan 2008). From the obtained results, the high substitution rate that is present in *S100A7L2* gene may therefore indicate a new functionalization of this gene. Indeed, it has been shown that this gene is expressed in epidermal and 21 sites were found under positive selection, with 15 of them being mapped within EF-hand domain (Goodwin and de Guzman Strong 2017). The presence of positive selection within this domain may suggest an alteration of either calcium binding or alteration in the target proteins after calcium-binding conformation chances (Denessiouk et al. 2014; Santamaria-Kisiel et al. 2006). Nevertheless, additional studies will be needed in order to better understand the functional implications of these specific alterations in this protein.

As mentioned before, the sequence similarity found between S100A7 and S100A7A genes in some Catarrhini species can be explained by an ongoing event of gene conversion, resulting in a homogenization of these members of S100A7 multigene family. According to Wolf and collaborators (Wolf et al. 2006), the coexpression of S100A7 and S100A7A in psoriasis suggests that these proteins have similar roles in cells, being able to participate in keratinocyte maturation, proliferation, and/or skin inflammation (Wolf et al. 2006). At first sight, the sequence similarity found in these two genes could suggest a model of concerted evolution acting in this gene family (Mano and Innan 2007; Zimmer et al. 1980). However, this model of evolution fails to explain the presence of heterogeneous genes such as *S100A7L2* and *S100A7* pseudogenes. Therefore, taking in consideration the obtained results, we propose a concerted and birth-and-death model to better explain the evolution of *S100A7* gene family (Karev et al. 2006; Nei et al. 1997). According to this model, several duplicated genes were produced in the evolution of primates, with some being homogenized, some of them acquiring new functions, and others become pseudogenes as a result of deleterious mutations. As a result, these genetic events result in divergent groups of *S100A7* genes and highly homologous genes within each of these groups. In this study, given the high mutation rate that is present in *S100A7L2* genes of Catarrhini species, we suggest a neofunctionalization of this gene. Moreover, the presence of two pseudogenes in humans (S100A7P1 and S100A7P2) appears to be in line with the proposed model of evolution.

Multigene families, with important roles in the immune system (e.g., MHC and immunoglobulin genes), have evolved under the birth-and-death model of evolution (Esteves et al. 2005; Nei et al. 1997; Nei and Rooney 2005). In recent years, S100A7 has been pointed out as a major component in the innate immune system (Gläser et al. 2005; Lee and Eckert 2007; Schröder and Harder 2006). S100A7 protein has a major role in inflammation and keratinocyte differentiation, not only in psoriatic skin but also in atopic dermatitis and skin cancer (Gläser et al. 2009; Salama et al. 2008). In addition to its anti-*E. coli* activity, it has also been stablished its importance as a chemotactic agent and cytokine towards neutrophils and T cells (Hoffmann et al. 1994; Tan et al. 1996). The important role of these proteins in defending the host from invading pathogens might be a major force for gene diversification, with natural selection acting in favor of a high diversity of these genes (S100A7L2 neofunctionalization). Our study suggests that *S100A7* gene family is evolving rapidly in Catarrhini species, especially considering S100A7L2, and that this gene may present a different, but very important role in immune defense of these mammals. Nevertheless, further studies are needed to fully elucidate the functional role of S100A7L2 in innate immunity.

## Conclusion

Due to the presence of five *Alu* sequences in human genome, Kulski and collaborators (Kulski et al. 2003) hypothesized that S100A7/S100A7P1 and S100A7P2/S100A7A clusters were duplicated during or after the divergence of Platyrrhini and Catarrhini, around ~ 35 Mya. In this study, we have identified at least one gene of each cluster in both Cercopithecoidea and Hominoidea species, which supports the duplication theory suggested by these authors. Moreover, our results further suggest that after the duplication of these genes, ongoing gene conversion events between *S100A7* and *S100A7A* genes as well as a neofunctionalization of *S100A7L2* gene might be shaping the evolution of *S100A7* gene family.

**Funding** This work was funded by the National Funds through the FCT—Foundation for Science and Technology, Portugal, under the doctoral fellowship of AAP (ref. SFRH/BD/128752/2017). FCT supported the FCT Investigator grants of PJE (IF/00376/2015).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Assis R, Bachtrog D (2013) Neofunctionalization of young duplicate genes in *Drosophila*. Proc Natl Acad Sci U S A 110:17409–17414. https://doi.org/10.1073/pnas.1313759110

Denessiouk K, Permyakov S, Denesyuk A, Permyakov E, Johnson MS (2014) Two structural motifs within canonical EF-hand calcium-binding domains identify five different classes of calcium buffers and sensors. PLoS One 9:e109287. https://doi.org/10.1371/journal.pone.0109287

Donato R, Cannon BR, Sorci G, Riuzzi F, Hsu K, Weber DJ, Geczy CL (2013) Functions of S100 proteins. Curr Mol Med 13:24–57

Engelkamp D, Schafer BW, Mattei MG, Erne P, Heizmann CW (1993) Six S100 genes are clustered on human chromosome 1q21: identification of two genes coding for the two previously unreported calcium-binding protein S100D and protein S100E. Proc Natl Acad Sci U S A 90:6547–6551. https://doi.org/10.1073/pnas.90.14.6547

Esteves PJ, Lanning D, Ferrand N, Knight KL, Zhai SK, van der Loo W (2005) The evolution of the immunoglobulin heavy chain variable region (IgV H ) in Leporids: an unusual case of transspecies polymorphism. Immunogenetics 57:874–882. https://doi.org/10.1007/s00251-005-0022-0

Gläser R, Harder J, Lange H, Bartels J, Christophers E, Schroder JM (2005) Antimicrobial psoriasin (S100A7) protects human skin from *Escherichia coli* infection. Nat Immunol 6:57–64. https://doi.org/10.1038/ni1142

Gläser R, Meyer-Hoffert U, Harder J, Cordes J, Wittersheim M, Kobliakova J, Fölster-Holst R, Proksch E, Schröder JM, Schwarz T (2009) The antimicrobial protein psoriasin (S100A7) is upregulated in atopic dermatitis and after experimental skin barrier disruption. J Invest Dermatol 129:641–649. https://doi.org/10.1038/jid.2008.268

Goodwin ZA, de Guzman Strong C (2017) Recent positive selection in genes of the mammalian epidermal differentiation complex locus. Front Genet 7:227. https://doi.org/10.3389/fgene.2016.00227

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, Dunn C, Baker C, Armstrong J, Diekhans M, Paten B, Shendure J, Wilson RK, Haussler D, Chin CS, Eichler EE (2016) Long-read sequence assembly of the gorilla genome. Science 352:aae0344. https://doi.org/10.1126/science.aae0344

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: Nucleic acids symposium series, 1999. vol 41. [London]: Information Retrieval Ltd., c1979-c2000., pp 95–98

Hoffmann HJ, Olsen E, Etzerodt M, Madsen P, Thøgersen HC, Kruse T, Celis JE (1994) Psoriasin binds calcium and is upregulated by calcium to levels that resemble those observed in normal skin. J Invest Dermatol 103:370–375

Ishikawa K, Nakagawa A, Tanaka I, Suzuki M, Nishihira J (2000) The structure of human MRP8, a member of the S100 calcium-binding protein family, by MAD phasing at 1.9 A resolution. Acta Crystallogr Sect D Biol Crystallogr 56:559–566. https://doi.org/10.1107/s0907444900002833

Karev GP, Wolf YI, Koonin EV (2006) Birth and death models of genome evolution. In: Power laws, scale-free networks and genome biology. Springer US, Boston, MA, pp 65–85. https://doi.org/10.1007/0-387-33916-7_6

Kuderna LFK, Tomlinson C, Hillier LDW, Tran A, Fiddes IT, Armstrong J, Laayouni H, Gordon D, Huddleston J, Garcia Perez R, Povolotskaya I, Serres Armero A, Gómez Garrido J, Ho D, Ribeca P, Alioto T, Green RE, Paten B, Navarro A, Betranpetit J, Herrero J, Eichler EE, Sharp AJ, Feuk L, Warren WC, Marques-Bonet T (2017) A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan_tro_3.0). GigaScience 6:1–6. https://doi.org/10.1093/gigascience/gix098

Kulski JK, Lim CP, Dunn DS, Bellgard M (2003) Genomic and phylogenetic analysis of the S100A7 (psoriasin) gene duplications within the region of the S100 gene cluster on human chromosome 1q21. J Mol Evol 56:397–406. https://doi.org/10.1007/s00239-002-2410-5

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054

Lee KC, Eckert RL (2007) S100A7 (Psoriasin) - mechanism of antibacterial action in wounds. J Invest Dermatol 127:945–957. https://doi.org/10.1038/sj.jid.5700663

Liu GJ, Wu Q, Liu GL, Song XY, Zhang JH (2015) Psoriasin (S100A7) is a novel biomarker for lung squamous cell carcinoma in humans. Cancer Cell Int 15(9). https://doi.org/10.1186/s12935-014-0154-0

Madsen P, Rasmussen HH, Leffers H, Honoré B, Dejgaard K, Olsen E, Kiil J, Walbum E, Andersen AH, Basse B, Lauridsen JB, Ratz GP, Celis A, Vandekerckhove J, Celis JE (1991) Molecular-cloning, occurrence, and expression of a novel partially secreted protein psoriasin that is highly up-regulated in psoriatic skin. J Invest Dermatol 97:701–712. https://doi.org/10.1111/1523-1747.ep12484041

Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R (2013) Gene duplication as a major force in evolution. J Genet 92:155–161. https://doi.org/10.1007/s12041-013-0212-8

Mano S, Innan H (2007) The evolutionary rate of multigene family under concerted evolution. Genes Genet Syst 82:530–530

Marenholz I, Heizmann CW, Fritz G (2004) S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). Biochem Biophys Res Commun 322:1111–1122. https://doi.org/10.1016/j.bbrc.2004.07.096

Moroz OV, Antson AA, Dodson GG, Wilson KS, Skibshoj I, Lukanidin EM, Bronstein IB (2000) Crystallization and preliminary X-ray diffraction analysis of human calcium-binding protein S100A12. Acta Crystallogr Sect D Biol Crystallogr 56:189–191. https://doi.org/10.1107/s0907444999014936

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 94:7799–7806. https://doi.org/10.1073/pnas.94.15.7799

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152. https://doi.org/10.1146/annurev.genet.39.073003.112240

Norgren RB (2013) Improving genome assemblies and annotations for nonhuman primates. ILAR J 54:144–153. https://doi.org/10.1093/ilar/ilt037

Otto SP, Yong P (2002) The evolution of gene duplicates. In: Advances in genetics, vol 46. Elsevier, pp 451–483

Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5:28–28. https://doi.org/10.1186/1471-2148-5-28

Rogers J (2013) In transition: primate genomics at a time of rapid change. ILAR J 54:224–233. https://doi.org/10.1093/ilar/ilt042

Rogers J, Gibbs RA (2014) Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet 15:347–359. https://doi.org/10.1038/nrg3707

Ruse M, Broome AM, Eckert RL (2003) S100A7 (psoriasin) interacts with epidermal fatty acid binding protein and localizes in focal adhesion-like structures in cultured keratinocytes. J Invest Dermatol 121:132–141. https://doi.org/10.1046/j.1523-1747.2003.12309.x

Salama I, Malone PS, Mihaimeed F, Jones JL (2008) A review of the S100 proteins in cancer. Ejso 34:357–364. https://doi.org/10.1016/j.ejso.2007.04.009

Santamaria-Kisiel L, Rintala-Dempsey AC, Shaw GS (2006) Calcium-dependent and independent interactions of the S100 protein family. Biochem J 396:201–214. https://doi.org/10.1042/bj20060195

Schrago CG, Russo CAM (2003) Timing the origin of New World monkeys. Mol Biol Evol 20:1620–1625. https://doi.org/10.1093/molbev/msg172

Schröder JM, Harder J (2006) Antimicrobial skin peptides and proteins. Cell Mol Life Sci 63:469–486. https://doi.org/10.1007/s00018-005-5364-0

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

Tan JQ et al (1996) Psoriasin: a novel chemotactic protein. J Invest Dermatol 107:5–10

Teshima KM, Innan H (2008) Neofunctionalization of duplicated genes under the pressure of gene conversion. Genetics 178:1385–1398. https://doi.org/10.1534/genetics.107.082933

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

Webb M, Emberley ED, Lizardo M, Alowami S, Qing G, Alfia'ar A, Snell-Curtis LJ, Niu Y, Civetta A, Myal Y, Shiu R, Murphy LC, Watson PH (2005) Expression analysis of the mouse S100A7/psoriasin gene in skin inflammation and mammary tumorigenesis. BMC Cancer 5:16. https://doi.org/10.1186/1471-2407-5-17

Wolf R, Voscopoulos CJ, FitzGerald PC, Goldsmith P, Cataisson C, Gunsior M, Walz M, Ruzicka T, Yuspa SH (2006) The mouse S100A15 ortholog parallels genomic organization, structure, gene expression, and protein-processing pattern of the human S100A7/A15 subfamily during epidermal maturation. J Invest Dermatol 126:1600–1608. https://doi.org/10.1038/sj.jid.5700210

Xia C, Braunstein Z, Toomey AC, Zhong JX, Rao XQ (2018) S100 proteins as an important regulator of macrophage inflammation. Front Immunol 8:11. https://doi.org/10.3389/fimmu.2017.01908

Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, du H, Chen J, Chen R, Zhang P, Huang Z, Thompson JR, Meng Y, Bai Y, Wang J, Zhuo M, Wang T, Huang Y, Wei L, Li J, Wang Z, Hu H, Yang P, le L, Stenson PD, Li B, Liu X, Ball EV, An N, Huang Q, Zhang Y, Fan W, Zhang X, Li Y, Wang W, Katze MG, Su B, Nielsen R, Yang H, Wang J, Wang X, Wang J (2011) Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol 29:1019–1023. https://doi.org/10.1038/nbt.1992 https://www.nature.com/articles/nbt.1992#supplementary-information

Zimmer DB, Eubanks JO, Ramakrishnan D, Criscitiello MF (2013) Evolution of the S100 family of calcium sensor proteins. Cell Calcium 53:170–179. https://doi.org/10.1016/j.ceca.2012.11.006

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC (1980) Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proc Natl Acad Sci U S A 77:2158–2162