HAL open science

Chromosomal mapping, differential origin and evolution of the S100 gene family

Xuan Shang, Hanhua Cheng, Rongjia Zhou

To cite this version:

Xuan Shang, Hanhua Cheng, Rongjia Zhou. Chromosomal mapping, differential origin and evolution of the S100 gene family. Genetics Selection Evolution, 2008, 40 (4), pp.449-464. hal-00894638

HAL Id: hal-00894638

https://hal.science/hal-00894638

Submitted on 11 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Genet. Sel. Evol. 40 (2008) 449–464
© INRA, EDP Sciences, 2008
DOI: 10.1051/gse:2008013

Available online at:
www.gse-journal.org

Original article

Chromosomal mapping, differential origin and evolution of the *S100* gene family

Xuan SHANG, Hanhua CHENG*, Rongjia ZHOU*

Department of Genetics and Center for Developmental Biology, College of Life Sciences, Wuhan University, Wuhan 430072, P. R. China

(Received 13 October 2007; accepted 21 December 2007)

Abstract – S100 proteins are calcium-binding proteins, which exist only in vertebrates and which constitute a large protein family. The origin and evolution of the S100 family in vertebrate lineages remain a challenge. Here, we examined the synteny conservation of mammalian *S100A* genes by analysing the sequence of available vertebrate *S100* genes in databases. Five *S100A* gene members, unknown previously, were identified by chromosome mapping analysis. Mammalian *S100A* genes are duplicated and clustered on a single chromosome while two *S100A* gene clusters are found on separate chromosomes in teleost fish, suggesting that *S100A* genes existed in fish before the fish-specific genome duplication took place. During speciation, tandem gene duplication events within the cluster of *S100A* genes of a given chromosome have probably led to the multiple members of the *S100A* gene family. These duplicated genes have been retained in the genome either by neofunctionalisation and/or subfunctionalisation or have evolved into non-coding sequences. However in vertebrate genomes, other *S100* genes are also present *i.e. S100P*, *S100B*, *S100G* and *S100Z*, which exist as single copy genes distributed on different chromosomes, suggesting that they could have evolved from an ancestor different to that of the *S100A* genes.

chromosome mapping / *S100* / genome duplication / synteny / vertebrate

1. INTRODUCTION

S100 proteins constitute the largest gene family within the EF-hand protein super-family. In 1965, Moore isolated from bovine brain the first protein members of the S100 family: S100A1 and S100B [17]. In the following years, many other members of the S100 family were identified based on sequence homology and similar structural properties. For example, the human S100 family includes 20 members, which share 22% to 57% sequence identity [13]. S100 proteins are small acidic proteins (9–14 kDa) and contain two distinct EF-hand motifs. The C-terminal EF-hand contains a classical Ca^{2+}-binding motif, common to all

*Corresponding authors: rjzhou@whu.edu.cn; hhcheng@whu.edu.cn

Article published by EDP Sciences

EF-hand proteins while the N-terminal EF-hand differs from the classical EF-hand motif and constitutes a special characteristic of the S100 proteins.

S100 proteins exhibit a unique pattern of tissue/cell type specific expression and exert their intracellular effects by interacting with different target proteins that modulate their activity [5,23,31]. Two well-known pairs are S100A11-annexin A1 and S100A10-annexin A2 [9,20,24,25,27] and recently, interaction between S100A11 and annexin A6 has also been reported [3]. Until now, over 90 potential target proteins have been identified [23]. Many studies have observed an altered expression of various S100 proteins in a large number of diseases including cancer, depression, Down syndrome, Alzheimer disease and cystic fibrosis [1,13,14,26,28,29]. Therefore, S100 proteins could constitute important diagnostic markers as well as therapeutic targets of many diseases.

All known *S100* genes are found only in vertebrates and no *S100-like* sequences have ever been detected in invertebrates such as insects, nematodes and protozoa based on the analysis of available genome sequence information. This suggests that the genes encoding S100 proteins belong to a "young" gene family *i.e.* that originated during vertebrate evolution. Interestingly, because of the short phylogenetic history and the conservation of the *S100A* gene cluster in man and mouse [21], their origin in the vertebrate lineages remains a challenge. Moreover, in non-mammalian systems such as fish species, information on the *S100* gene family evolution and genomic organisation is very scarce and only a few *S100* gene members have been identified [7]. In this work, we analysed *S100* gene sequences of various vertebrates including mammals and fish from available databases using both comparative genomics and phylogenetic methods, and we present a model of the molecular evolution of the *S100* genes, which contributes to a better understanding of the mechanisms of evolution and biological functions of the *S100* gene family.

2. MATERIALS AND METHODS

2.1. Sequences and positions on the chromosomes or assembly scaffolds

A search in the GenBank and Ensembl databases (v39) provided 118 sequences of the *S100* gene family from seven mammals whose genomes have been sequenced. In addition, using human *S100* gene sequences as query sequences, orthologous sequences were found for three teleost fish, *Danio rerio*, *Takifugu rubripes* (Japanese pufferfish), *Tetraodon nigroviridis* (freshwater pufferfish). The complete list of the *S100* mammalian and fish sequences compiled in this study together with gene names and accession numbers are given in Table I.

Table I. Vertebrate *S100* genes available from NCBI and Ensembl databases.

Organism	Gene/code	Accession No.	Organism	Gene/code	Accession No.
Homo sapiens	*S100A1*	NP_006262	*Pan troglodytes*	*S100a11*	ENSPTRG00000001303
	S100A2	NP_005969		*S100a12*	ENSPTRG00000001346
	S100A3	NP_002951		*S100a13*	ENSPTRG00000022794
	S100A4	NP_002952		*S100a14*	ENSPTRG00000024364
	S100A5	NP_002953		*S100a15*	ENSPTRG00000001349
	S100A6	NP_055439		*S100a16*	ENSPTRG00000023848
	S100A7	NP_002954		*S100b*	ENSPTRG00000014026
	S100A8	NP_002955		*S100g*	ENSPTRG00000021699
	S100A9	NP_002956		*S100p*	ENSPTRG00000015887
	S100A10	NP_002957	*Danio rerio*[a]	*z55514*	ENSDARG00000055514
	S100A11	NP_005611		*z15543*	ENSDARG00000015543
	S100A12	NP_005612		*z25254*	ENSDARG00000025254
	S100A13	NP_005970		*a55589*	ENSDARG00000055589
	S100A14	NP_065723		*z36773*	ENSDARG00000036773
	S100A15	NP_789793		*z37425*	ENSDARG00000037425
	S100A16	NP_525127		*z09978*	ENSDARG00000009978
	S100B	NP_006263		*z38729*	ENSDARG00000038729
	S100G	NP_004048		*z57598*	ENSDARG00000057598
	S100P	NP_005971	*Takifugu rubripes*[a]	*f129020*	NEWSINFRUG00000129020
	S100Z	NP_570128		*f127285*	NEWSINFRUG00000127285
Pan troglodytes	*S100a1*	ENSPTRG00000001355		*f152973*	NEWSINFRUG00000152973
	S100a2	ENSPTRG00000001354		*f141424*	NEWSINFRUG00000141424
	S100a3	ENSPTRG00000001353		*f137581*	NEWSINFRUG00000137581
	S100a4	ENSPTRG00000001348		*f137599*	NEWSINFRUG00000137599
	S100a5	ENSPTRG00000001352		*f136068*	NEWSINFRUG00000136068

Table I. Continued.

Organism	Gene/code	Accession No.	Organism	Gene/code	Accession No.
	S100a6	ENSPTRG00000001351		*f159674*	NEWSINFRUG00000159674
	S100a7	ENSPTRG00000001350		*f163415*	NEWSINFRUG00000163415
	S100a8	ENSPTRG00000001347		*f159852*	NEWSINFRUG00000159852
	S100a9	ENSPTRG00000001345		*f156133*	NEWSINFRUG00000156133
	S100a10	ENSPTRG00000001302		*f165637*	NEWSINFRUG00000165637
Monodelphis domestica	*S100a1*	ENSMODG00000017368	*Mus musculus*	*S100b*	ENSMUSG00000033208
	S100a3	ENSMODG00000017395		*S100g*	ENSMUSG00000040808
	S100a4	ENSMODG00000017397		*S100z*	ENSMUSG00000021679
	S100a5	ENSMODG00000017400	*Tetraodon nigroviridis*[a]	*t44001*	GSTENG00033944001
	S100a8	ENSMODG00000017403		*t30001*	GSTENG00025230001
	S100a9	ENSMODG00000017406		*t25001*	GSTENG00005225001
	S100a10	ENSMODG00000018919		*t75001*	GSTENG00032575001
	S100a11	ENSMODG00000018920		*t87001*	GSTENG00032587001
	S100a12	ENSMODG00000017410		*t45001*	GSTENG00033945001
	S100a13	ENSMODG00000017387		*t22001*	GSTENG00013622001
	S100a14	ENSMODG00000017390		*t60001*	GSTENG00038360001
	S100a15	ENSMODG00000017402		*t74001*	GSTENG00032574001
	S100a16	ENSMODG00000017391		*t85001*	GSTENG00032585001
	S100g	ENSMODG00000017180		*t99001*	GSTENG00011699001
	S100p	ENSMODG00000002897	*Rattus norvegicus*	*S100a1*	ENSRNOG00000012410
	S100z	ENSMODG00000019747		*S100a3*	ENSRNOG00000012008
Mus musculus	*S100a1*	ENSMUSG00000044080		*S100a4*	ENSRNOG00000011821
	S100a3	ENSMUSG00000001021		*S100a5*	ENSRNOG00000011748
	S100a4	ENSMUSG00000001020		*S100a6*	ENSRNOG00000011647
	S100a5	ENSMUSG00000001023		*S100a8*	ENSRNOG00000011557
	S100a6	ENSMUSG00000001025		*S100a9*	ENSRNOG00000011483
	S100a8	ENSMUSG00000056054		*S100a10*	ENSRNOG00000023226

Table I. Continued.

Organism	Gene/code	Accession No.	Organism	Gene/code	Accession No.
	S100a9	ENSMUSG00000056071		*S100a11*	ENSRNOG00000010105
	S100a10	ENSMUSG00000041959		*S100a13*	ENSRNOG00000012393
	S100a11	ENSMUSG00000027907		*S100a15*	ENSRNOG00000033352
	S100a13	ENSMUSG00000042312		*S100a16*	ENSRNOG00000012053
	S100a14	ENSMUSG00000042306		*S100b*	ENSRNOG00000001295
	S100a15	ENSMUSG00000063767		*S100g*	ENSRNOG00000004222
	S100a16	ENSMUSG00000074457		*S100z*	ENSRNOG00000017998
Bos taurus	*S100a1*	ENSBTAG00000005163	*Canis familiaris*	*S100a1*	ENSCAFG00000017540
	S100a2	ENSBTAG00000000463		*S100a2*	ENSCAFG00000017547
	S100a4	ENSBTAG00000019203		*S100a3*	ENSCAFG00000017548
	S100a5	ENSBTAG00000000644		*S100a4*	ENSCAFG00000017550
	S100a6	ENSBTAG00000000643		*S100a5*	ENSCAFG00000017552
	S100a7	ENSBTAG00000008238		*S100a6*	ENSCAFG00000017553
	S100a8	ENSBTAG00000012640		*S100a8*	ENSCAFG000000175571
	S100a9	ENSBTAG00000006505		*S100a9*	ENSCAFG00000017558
	S100a10	ENSBTAG00000015147		*S100a13*	ENSCAFG00000017542
	S100a11	ENSBTAG00000015145		*S100a14*	ENSCAFG00000017544
	S100a12	NP_777076		*S100a15*	ENSCAFG00000017554
	S100a13	ENSBTAG00000021378		*S100a16*	ENSCAFG00000017545
	S100a14	ENSBTAG00000024437		*S100b*	ENSCAFG00000012228
	S100a15	ENSBTAG00000014204		*S100p*	ENSCAFG00000014333
	S100a16	ENSBTAG00000004777		*S100g*	ENSCAFG00000012583
	S100g	ENSBTAG00000017020			
	S100z	ENSBTAG00000020201			

[a]Codes of fish genes were defined by authors.

The chromosomal localisation of these genes is based on the Ensembl v39 genomic location data.

2.2. Gene prediction

In order to detect sequences that may contain unknown *S100* sequences, genomic sequences were aligned with the exons of homologous human genes by Vector NTI software and those identified were assembled into putative mRNA sequences. These mRNA sequences were translated into protein sequences, which were aligned with the corresponding human proteins to test the validity of the prediction.

2.3. Sequence alignment and construction of phylogenetic trees

Multiple alignments were performed with the Vector NTI software and Neighbour-Joining phylogenetic trees were built using the Phylip program (Joseph Felsenstein, Washington University). The reliability of the trees was measured by bootstrap analysis with 1000 replicates and the trees were edited and viewed by Treeview software.

3. RESULTS

3.1. Mammalian *S100A* genes are duplicated and clustered on one chromosome

The chromosomal organisation and location of the *S100A* genes identified in seven mammalian species *i.e.* man, chimpanzee, cow, dog, rat, mouse and opossum were determined using the Ensembl database. The results revealed that in each of these seven mammals the *S100A* genes are clustered on a single chromosome and comprise up to 16 members (Fig. 1 and Tab. I). Although these genes are located on a single chromosome, two subgroups (SGs) were identified: SG1 in which *S100A10* and *S100A11* are always tightly linked and SG2 in which other members (*S100A1–9* and *12–16*) are generally clustered together (Fig. 1). The distance between the two SGs covers several megabases, whereas only a few kilobases separates genes within each SG. Interestingly, the relative positions of the genes on the chromosomes are conserved among these mammalian species, which indicates a high level of conserved synteny (Fig. 1). In addition, other putative *S100A* gene members, previously unknown, were predicted from available genome sequence data based on information of conserved synteny and protein homology. Five genes were identified, *S100A3* and *S100A14*

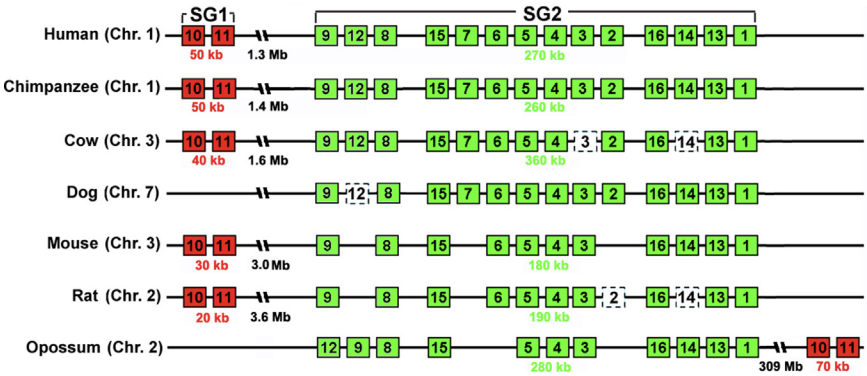

Figure 1. Conserved synteny and subgroup (SG) definition of the *S100A* gene cluster in mammals. The *S100A* genes from different mammalian species are clustered on a single chromosome and are divided into two subgroups (SG1 and SG2) based on their relative localisation on the chromosome. The gene distribution was analysed from data in the Ensembl database (http://www.ensembl.org). *S100A1–16* genes are indicated as two blocks of synteny by two colour boxes. Dashed boxes indicate the predicted genes. The name of the species and chromosome numbers are shown on the left.

in the cow, *S100A12* in the dog and *S100A2* and *S100A14* in the rat (Fig. 2 and Tab. II). Multiple protein sequence alignments with the corresponding human S100A proteins showed a high level of homology (Fig. 2). Thus, these sequences are not pseudogenes and corresponding expressed sequence tags (EST) are present in the EST databases (for details see legend of Fig. 2).

Differences in the arrangement of the *S100A* genes were observed between the opossum and the other species examined, *i.e.* SG1 (*S100A10* and *11*) together with *S100A1* is located at the 3′ end of opossum chromosome 2 and at the 5′ end of the corresponding chromosomes in the other species (Fig. 1). Also, in the opossum, the positions of *S100A9* and *S10012* are reversed comparatively to those in the other species. These discordances indicate that chromosomal rearrangements having occurred during mammalian speciation have disrupted the syntenic gene associations.

3.2. Two clusters of *S100A* genes in teleost fish

A phylogenetic tree was constructed to determine accurate predictions of orthology and paralogy relationships between fish and mammalian *S100A* genes (Fig. 3a). Fish S100A proteins are divided into two SGs as defined in Figure 1. SG1 includes *S100A10* and *S100A11* genes while SG2 contains all the other *S100A* genes. This distribution is supported by the data on gene organisation

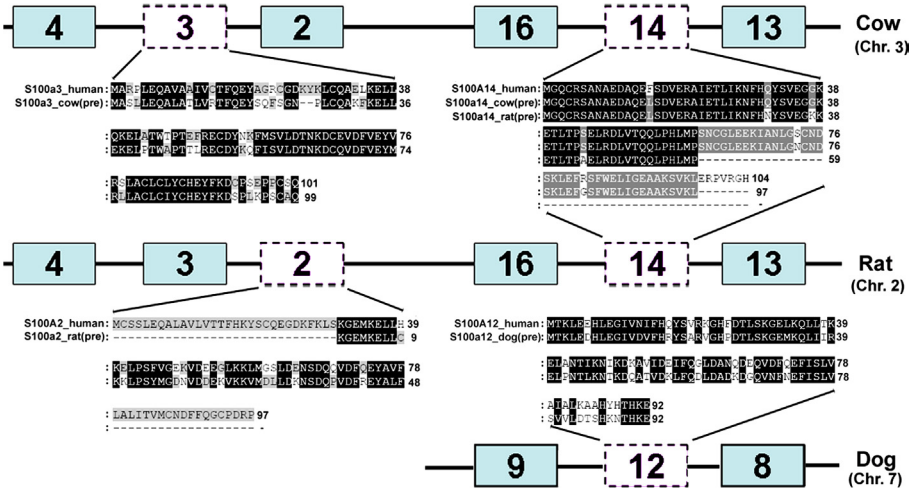

Figure 2. Five *S100A* predicted genes based on conserved synteny and homology. Predicted genes include bovine *S100a3* (complete CDS) and *S100a14* (partial CDS), rat *S100a2* (partial CDS) and *S100a14* (partial CDS) and dog *S100a12* (complete CDS). The multiple sequence alignments with the corresponding human S100 proteins are shown in the centre to confirm the identity of predicted genes. Two EST sequences (GenBank Accession Nos. XM_001063574 and NM_001079634) are similar to rat and bovine *S100a14*, especially in the CDS regions. More information is necessary to confirm that the two sequences correspond to gene *S100a14*. Two other EST: DR104796 (canine cardiovascular system biased cDNA, a *Canis familiaris* cDNA similar to that of Hs S100 calcium-binding protein A12) and DV924106 (*Bos taurus* cDNA clone IMAGE: 8232591 5′, mRNA sequence) may be the relevant bovine and rat genes, *S100a3* or *S100a2*, respectively.

for available fish genome assembly scaffolds and human chromosome 1 (Fig. 3b) although in some cases, gene members are only temporarily positioned on the scaffolds and their definite chromosome localisation needs to be confirmed. Seven zebrafish genes classified in the *S100A* category form two clusters on chromosome 16 and chromosome 19, respectively. Among the nine *takifugu* genes belonging to the *S100A* category, at least six form two clusters on scaffold 37 and scaffold 252, respectively. Furthermore, in *tetraodon*, a similar gene arrangement exists with four genes clustered on chromosome 21 and two other genes clustered on chromosome 8. Interestingly, in each synteny group, gene members of both SGs 1 and 2 are present. Thus overall, these results based on phylogenetic and comparative genomic analyses show the existence of two *S100A* gene clusters in fish genomes and only one in mammalian genomes.

Table II. Chromosome localisation and exon information of predicted *S100A* genes.

Name	Chromosome	Exons			
		No. exon	Start	End	Length (bp)
*S100A12*_dog	7	1	46 170 564	46 170 611	48
(complete CDS)		2	46 171 134	46 171 291	158
		3	46 171 670	46 171 945	276
*S100A3*_cow	3	1	11 224 336	11 224 412	77
(complete CDS)		2	11 225 308	11 225 447	140
		3	11 225 990	11 226 471	482
*S100A14*_cow	3	1	11 170 356	11 170 386	31
(partial CDS)		2	11 170 755	11 170 865	111
		3	11 171 303	1 171 449	147
		4 (partial)	11 171 672	1 171 785	115
*S100A14*_rat	2	1	182 799 278	182 799 757	30
(partial CDS)		2	182 800 085	182 800 202	118
		3	182 800 617	182 800 763	147
		4	–	–	–
*S100A2*_rat	2	1	–	–	–
(partial CDS)		2 (partial)	182 871 245	182 871 295	51
		3 (partial)	182 872 218	182 872 310	93

3.3. Presence of other single copy *S100* genes scattered in vertebrate genomes

Four other *S100* genes *i.e. S100P*, *S100B*, *S100G* and *S100Z* are present in the human genome and contrarily to the *S100A* genes clustered on chromosome 1 they are distributed on different chromosomes. A similar distribution pattern of the homologous genes is found in the genomes of the chimpanzee, cow, dog, rat, mouse and opossum. The absence of gene *S100P* could be due to the incomplete genome sequencing *e.g.* in the cow and the fish species examined here or to loss of the corresponding sequences during speciation *e.g.* in the mouse and rat (Fig. 4). Unlike the *S100A* genes, *S100P*, *B*, *G* and *Z* genes also exist as single copies in the three fish genomes according to the phylogenetic analysis.

4. DISCUSSION

We analysed all available information on *S100* genes in seven mammalian and three fish species and we determined their phylogenetic relationship and genomic organisation based on abundant sequence resources in databases.

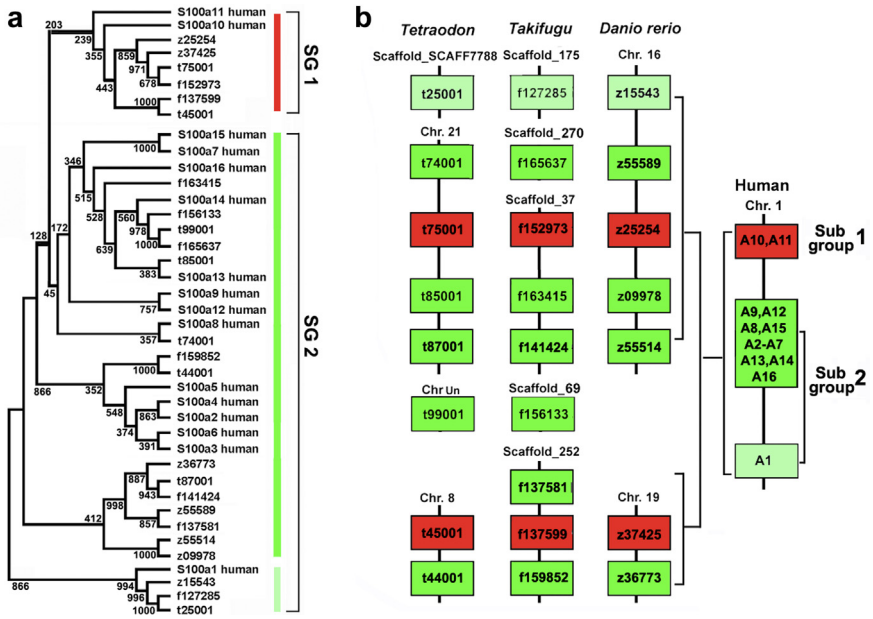

Figure 3. Analysis of the phylogenetic relationships and chromosome mapping of *S100A* genes in mammals and fish. (a) Phylogenetic tree of S100A proteins. The numbers on the branches represent the bootstrap values from 1000 replicates obtained using the (N-J) method. The tree shows two major subgroups of S100A proteins as in Figure 1. (b) Localisation of *S100A* genes on chromosomes or assembly scaffolds. At least two clusters are observed in fish species but only one in man. Genes are in the boxes and chromosome or scaffold numbers are shown at the top of each linkage group or gene. z09878 is an *S100* gene member, *ictacalcin* previously identified in zebrafish [7].

Until now, S100 proteins have been detected only in vertebrates, suggesting that they first appeared during vertebrate evolution. In the mouse and man [21], it has been previously shown that all *S100A* genes are present on a single chromosome but form two SGs, which agrees with our results on their genomic organisation and chromosomal localisation in other mammalian species *i.e.* the cow, dog, chimpanzee, rat and opossum (Fig. 1). We identified five new previously unknown *S100A* genes [18]. The structure of mammalian *S100A* genes is also highly conserved, generally, comprising three exons separated by two introns with the first exon untranslated [6]. The clustered localisations on a single chromosome, the highly conserved synteny and the similarity in exon/intron organisation suggest that gene duplication is responsible for the major expansion of this gene family.

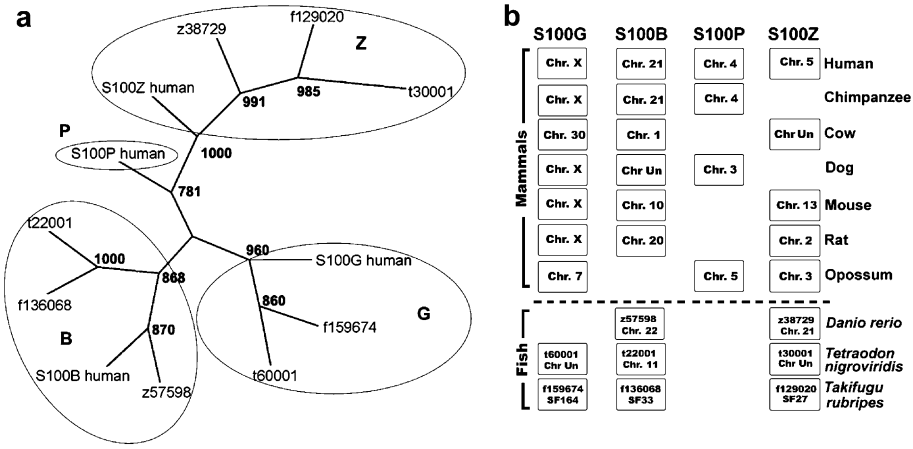

Figure 4. Phylogenetic tree and distribution of other S100 proteins in vertebrates. Mammalian homologous genes were found in NCBI and Ensembl databases. Fish genes were identified by searching the paralogue of the corresponding human *S100* gene. (a) Phylogenetic tree of S100B, S100G, S100P and S100Z proteins. The numbers on the branches represent the bootstrap values (%) from 1000 replicates obtained using the N-J method. Eight fish genes are classified into the *S100B*, *S100G* and *S100Z* subgroups. (b) Distribution of all known *S100B*, *S100G*, *S100P* and *S100Z* genes from seven mammals and three fish species. Chromosome numbers (for mammals) and chromosome/scaffold numbers with gene names are indicated in boxes (SF = scaffold, Un = unknown).

Furthermore, we analysed the organisation of *S100A* genes in three fish model species: zebrafish, *takifugu* and *tetraodon*. The phylogenetic tree shows that in these fish species the *S100A* genes are also subdivided into two major SGs as observed in mammalian species. However, in contrast to the existence of a single cluster in mammalian genomes, at least two clusters are present in fish genomes (Fig. 3). A comparison of the genomic architecture and arrangements between fish and mammalian *S100A* genes shows that they are remarkably consistent with the occurrence of the fish-specific genome duplication (FSGD or 3R) during vertebrate evolution. More and more studies propose that, during the evolution of vertebrates, two rounds (2R) of genome duplication occurred first and then later in the stem lineage of ray-finned fishes, not belonging to land vertebrates, a third genome duplication occurred (FSGD or 3R) [4,10,16]. Indeed, duplicated chromosomes and duplicated *S100A* genes are present in zebrafish *i.e.* chromosomes 16 and 19, in *tetraodon i.e.* chromosomes 8 and 21, and in *takifugu i.e.* scaffolds 37 and 252. In fact, previous studies have reported that *tetraodon* chromosomes 8 and 21 and zebrafish chromosomes 16

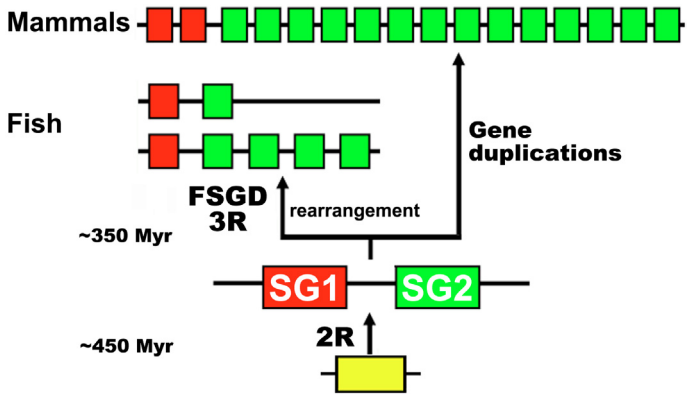

Figure 5. Model of the molecular evolution of *S100A* genes. The genomic architecture of fish and mammalian *S100A* genes is shown. The ancestral *S100A* gene was duplicated and formed two members defined as SG1 and SG2 during the 2R genome duplication about 450 Myr ago. Then, fish genomes (*e.g.* zebrafish, *tetraodon* and *takifugu*) underwent FSGD (3R), which generated two clusters of *S100A* genes on two different chromosomes about 350 Myr ago. Other rearrangements also took place during this process. Mammalian (*e.g.* human) *S100A* gene members have increased only by gene duplications on a single chromosome since a third round genome duplication (3R) did not occur in mammals.

and 19 originate from a common ancestral chromosome L. Furthermore, a high degree of conserved synteny between individual *tetraodon* chromosomes and zebrafish linkage groups has been observed and suggests a 1:1 chromosome correspondence in both species [8,30]. After the FSGD, interchromosomal rearrangement events (including chromosome fissions, fusions and translocations) probably occurred [10], which would explain our observations that duplicated *S100A* genes are asymmetrically distributed and that the gene positions in the two clusters are a little different.

We suggest that a single ancestral *S100A* gene was duplicated and led to the two gene member types defined as SG1 and SG2 during the 2R genome duplication event about 450 Myr (million years) ago. Then, fish genomes (*e.g.* zebrafish, *tetraodon* and *takifugu*) underwent FSGD (3R) and during fish speciation, two clusters of *S100A* genes appeared on two chromosomes about 350 Myr ago. In mammalian species, because of the absence of a 3R, only one cluster of *S100A* genes included in SGs 1 and 2 is present on a single chromosome. However, to adapt to diverse environmental conditions, mammals acquired multiple *S100A* genes by tandem gene duplications within the cluster on the one chromosome (Fig. 5) as, for example, the five copies of human gene *S100A7*

(*S100A7a–S100A7e*) present at the same locus [11,18]. These duplicated genes may have been retained in the genome by neofunctionalisation and/or subfunctionalisation mechanisms [12] or may lead to pseudogenes, such as *S100A7d* and *S100A7e* [18]. However, some genes have either not been duplicated or have been lost during speciation, for example, *S100A2*, *A7* or *A12*, which are not found in the mouse or in the rat, respectively [18].

In the case of the *S100P*, *B*, *G* and *Z* genes, the situation is different to that of the *S100A* genes. In vertebrate genomes, these genes are scattered on different chromosomes and exist as single copies in both mammalian and fish species. This suggests that they could have evolved from an ancestral gene different to that of the *S100A* genes. Differences in the mode of their interaction with target proteins support this hypothesis. Data on the crystal structure and protein interactions show that the structures of the S100A10/annexin A2 [19] and S100A11/annexin A1 [20] complexes are alike. However, the S100B protein can form a complex with a peptide derived from the C-terminal regulatory domain of p53 [22], or a TRTK-12 peptide existing in CapZ [15], or a peptide derived from Ndr-kinase [2] and the comparison of the structures of these complexes reveals differences in the orientation of the three peptides and in the type of interaction patterns with S100B protein. Moreover, the structure of the S100A10/annexin A2 or S100A10/annexin A1 complexes is different to that of all S100B/peptide complexes. These differences in structure indicate a large diversity of *S100A* and other *S100* genes. However, Marenholz has previously reported that *S100B*, *P* and *Z* genes are evolutionarily related to gene *S100A1*, which might point to a common ancestor of the *S100* gene family [13]. More information, *i.e.* whole genome comparisons with other fish species, is necessary to determine whether these two groups of *S100* genes have evolved from different ancestors or a common one. The analysis presented here is based on the current information available for whole genome sequences in public databases. Data on whole genome sequences increase daily and contig assemblies are frequently updated. With the completion of the current genome projects and the beginning of future genome projects of other vertebrate model systems new information will be provided, which will help understand the evolution and function of the *S100* gene family.

ACKNOWLEDGEMENTS

The work was supported by the National Natural Science Foundation of China, the National Key Basic Research project (2006CB102103), the Program for New Century Excellent Talents in University and the 111 project #B06018. No financial conflict of interest exists.

REFERENCES

[1] Abraha H.D., Noble P.L., Nicolaides K.H., Sherwood R.A., Maternal serum S100 protein in normal and Down syndrome pregnancies, Prenat. Diagn. 19 (1999) 334–336.

[2] Bhattacharya S., Large E., Heizmann C.W., Hemmings B., Chazin W.J., Structure of the Ca2+/S100B/NDR kinase peptide complex: insights into S100 target specificity and activation of the kinase, Biochemistry 42 (2003) 14416–14426.

[3] Chang N., Sutherland C., Hesse E., Winkfein R., Wiehler W.B., Pho M., Veillette C., Li S., Wilson D.P., Kiss E., Walsh M.P., Identification of a novel interaction between the Ca(2+)-binding protein S100A11 and the Ca(2+)- and phospholipid-binding protein annexin A6, Am. J. Physiol. Cell Physiol. 292 (2007) C1417–C1430.

[4] Christoffels A., Koh E.G., Chia J.M., Brenner S., Aparicio S., Venkatesh B., Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes, Mol. Biol. Evol. 21 (2004) 1146–1151.

[5] Donato R., S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles, Int. J. Biochem. Cell Biol. 33 (2001) 637–668.

[6] Heizmann C.W., Fritz G., Schafer B.W., S100 proteins: structure, functions and pathology, Front. Biosci. 7 (2002) d1356–d1368.

[7] Hsiao C.D., Ekker M., Tsai H.J., Skin-specific expression of ictacalcin, a homolog of the S100 genes, during zebrafish embryogenesis, Dev. Dyn. 228 (2003) 745–750.

[8] Jaillon O., Aury J.M., Brunet F., Petit J.L., Stange-Thomann N., Mauceli E., Bouneau L., Fischer C., Ozouf-Costaz C., Bernot A., Nicaud S., Jaffe D., Fisher S., Lutfalla G., Dossat C., Segurens B., Dasilva C., Salanoubat M., Levy M., Boudet N., Castellano S., Anthouard V., Jubin C., Castelli V., Katinka M., Vacherie B., Biemont C., Skalli Z., Cattolico L., Poulain J., De Berardinis V., Cruaud C., Duprat S., Brottier P., Coutanceau J.P., Gouzy J., Parra G., Lardier G., Chapple C., McKernan K.J., McEwan P., Bosak S., Kellis M., Volff J.N., Guigo R., Zody M.C., Mesirov J., Lindblad-Toh K., Birren B., Nusbaum C., Kahn D., Robinson-Rechavi M., Laudet V., Schachter V., Quetier F., Saurin W., Scarpelli C., Wincker P., Lander E.S., Weissenbach J., Roest Crollius H., Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype, Nature 431 (2004) 946–957.

[9] Johnsson N., Marriott G., Weber K., p36, the major cytoplasmic substrate of src tyrosine protein kinase, binds to its p11 regulatory subunit via a short amino-terminal amphipathic helix, EMBO J. 7 (1988) 2435–2442.

[10] Kasahara M., Naruse K., Sasaki S., Nakatani Y., Qu W., Ahsan B., Yamada T., Nagayasu Y., Doi K., Kasai Y., Jindo T., Kobayashi D., Shimada A., Toyoda A., Kuroki Y., Fujiyama A., Sasaki T., Shimizu A., Asakawa S., Shimizu N., Hashimoto S., Yang J., Lee Y., Matsushima K., Sugano S., Sakaizumi M.,

Narita T., Ohishi K., Haga S., Ohta F., Nomoto H., Nogata K., Morishita T., Endo T., Shin I.T., Takeda H., Morishita S., Kohara Y., The medaka draft genome and insights into vertebrate genome evolution, Nature 447 (2007) 714–719.

[11] Kulski J.K., Lim C.P., Dunn D.S., Bellgard M., 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 (2003) 397–406.

[12] Lynch M., O'Hely M., Walsh B., Force A., The probability of preservation of a newly arisen gene duplicate, Genetics 159 (2001) 1789–1804.

[13] Marenholz I., Heizmann C.W., Fritz G., S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature), Biochem. Biophys. Res. Commun. 322 (2004) 1111–1122.

[14] Marks A., Allore R., S100 protein and Down syndrome, Bioessays 12 (1990) 381–383.

[15] McClintock K.A., Shaw G.S., A novel S100 target conformation is revealed by the solution structure of the Ca2+-S100B-TRTK-12 complex, J. Biol. Chem. 278 (2003) 6251–6257.

[16] Meyer A., van de Peer Y., From 2R to 3R: evidence for a fish-specific genome duplication (FSGD), Bioessays 27 (2005) 937–945.

[17] Moore B.W., A soluble protein characteristic of the nervous system, Biochem. Biophys. Res. Commun. 19 (1965) 739–744.

[18] Ravasi T., Hsu K., Goyette J., Schroder K., Yang Z., Rahimi F., Miranda L.P., Alewood P.F., Hume D.A., Geczy C., Probing the S100 protein family through genomic and functional analysis, Genomics 84 (2004) 10–22.

[19] Rety S., Sopkova J., Renouard M., Osterloh D., Gerke V., Tabaries S., Russo-Marie F., Lewit-Bentley A., The crystal structure of a complex of p11 with the annexin II N-terminal peptide, Nat. Struct. Biol. 6 (1999) 89–95.

[20] Rety S., Osterloh D., Arie J.P., Tabaries S., Seeman J., Russo-Marie F., Gerke V., Lewit-Bentley A., Structural basis of the Ca(2+)-dependent association between S100C (S100A11) and its target, the N-terminal part of annexin I, Structure 8 (2000) 175–184.

[21] Ridinger K., Ilg E.C., Niggli F.K., Heizmann C.W., Schafer B.W., Clustered organization of S100 genes in human and mouse, Biochim. Biophys. Acta 1448 (1998) 254–263.

[22] Rustandi R.R., Baldisseri D.M., Weber D.J., Structure of the negative regulatory domain of p53 bound to S100B(betabeta), Nat. Struct. Biol. 7 (2000) 570–574.

[23] Santamaria-Kisiel L., Rintala-Dempsey A.C., Shaw G.S., Calcium-dependent and -independent interactions of the S100 protein family, Biochem. J. 396 (2006) 201–214.

[24] Seemann J., Weber K., Gerke V., Structural requirements for annexin I-S100C complex-formation, Biochem. J. 319 (Pt 1) (1996) 123–129.

[25] Shang X., Sun J., He Y., Zhao W., Li Q., Zhou F., Chen B., Cheng H., Zhou R., Identification and predominant expression of annexin A2 in epithelial-type cells of the rice field eel, J. Cell Biochem. 101 (2007) 600–608.

[26] Sheng J.G., Mrak R.E., Griffin W.S., S100 beta protein expression in Alzheimer disease: potential role in the pathogenesis of neuritic plaques, J. Neurosci. Res. 39 (1994) 398–404.
[27] Sopkova-de Oliveira Santos J., Oling F.K., Rety S., Brisson A., Smith J.C., Lewit-Bentley A., S100 protein-annexin interactions: a model of the (Anx2-p11)(2) heterotetramer complex, Biochim. Biophys. Acta 1498 (2000) 181–191.
[28] Svenningsson P., Chergui K., Rachleff I., Flajolet M., Zhang X., El Yacoubi M., Vaugeois J.M., Nomikos G.G., Greengard P., Alterations in 5-HT1B receptor function by p11 in depression-like states, Science 311 (2006) 77–80.
[29] Tirkos S., Newbigging S., Nguyen V., Keet M., Ackerley C., Kent G., Rozmahel R.F., Expression of S100A8 correlates with inflammatory lung disease in congenic mice deficient of the cystic fibrosis transmembrane conductance regulator, Respir. Res. 7 (2006) 51.
[30] Woods I.G., Wilson C., Friedlander B., Chang P., Reyes D.K., Nix R., Kelly P.D., Chu F., Postlethwait J.H., Talbot W.S., The zebrafish gene map defines ancestral vertebrate chromosomes, Genome Res. 15 (2005) 1307–1314.
[31] Zimmer D.B., Wright Sadosky P., Weber D.J., Molecular mechanisms of S100-target protein interactions, Microsc. Res. Tech. 60 (2003) 552–559.