Molecular Phylogenetics and Evolution 116 (2017) 78–86

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Unusual loss of chymosin in mammalian lineages parallels neo-natal immune transfer strategies

Mónica Lopes-Marques[a,b], Raquel Ruivo[a], Elza Fonseca[a,c], Ana Teixeira[a], L. Filipe C. Castro[a,c,*]

[a] CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Porto, Portugal
[b] ICBAS – Institute of Biomedical Sciences Abel Salazar, U. Porto – University of Porto, Portugal
[c] Department of Biology, Faculty of Sciences, U. Porto – University of Porto, Portugal

ARTICLE INFO

Keywords:
Chymosin
Pseudogene
Mammals
Immunoglobulin G
Passive transfer

ABSTRACT

Gene duplication and loss are powerful drivers of evolutionary change. The role of loss in phenotypic diversification is notably illustrated by the variable enzymatic repertoire involved in vertebrate protein digestion. Among these we find the pepsin family of aspartic proteinases, including chymosin (*Cmy*). Previous studies demonstrated that *Cmy*, a neo-natal digestive pepsin, is inactivated in some primates, including humans. This pseudogenization event was hypothesized to result from the acquisition of maternal immune immunoglobulin G (IgG) transfer. By investigating 94 mammalian subgenomes we reveal an unprecedented level of *Cmy* erosion in placental mammals, with numerous independent events of gene loss taking place in Primates, Dermoptera, Rodentia, Cetacea and Perissodactyla. Our findings strongly suggest that the recurrent inactivation of *Cmy* correlates with the evolution of the passive transfer of IgG and uncovers a noteworthy case of evolutionary cross-talk between the digestive and the immune system, modulated by gene loss.

1. Introduction

Gene loss has long been considered a secondary driver in adaptive evolution. Yet, the current paradigm is shifting and gene loss emerging as a pivotal player in the sculpting of evolutionary change (Albalat and Canestro, 2016; Opazo et al., 2017). The genetic repertoire of gastric genes across vertebrate lineages, for instance, provides a remarkable example on the decisive role of gene loss in adaptive phenotypic variation: with several cases of gene expansion and gene loss with morpho-functional consequences (Castro et al., 2014, 2012; Kageyama, 2002; Ordoñez et al., 2008). A subset of digestive enzymes include the pepsin family of aspartic proteinases (Pearl and Blundell, 1984). In mammals, the pepsin family consists of 5 members highly expressed in the gastric mucosa, grouped according to phylogenetics and substrate specificity: chymosin (*Cmy*), pepsin A (*PgA*), B (*PgB*), C (*PgC*), and F (*PgF*) (Carginale et al., 2004; Kageyama, 2002; Wu et al., 2009; Yakabe et al., 1991). The pepsin gene family is widely disseminated, yet erratically distributed, within the tetrapod lineage, with cases of gene expansion, pseudogenization and loss (Castro et al., 2014, 2012; Kageyama, 2002). For example, while *Homo sapiens* has 3 copies of *PgA*, a single copy is found in *Anolis carolinesis* and *Xenopus tropicalis*; on the other hand, *PgA* is pseudogenized in *Monodelphis domestica* and *Mus musculus* exhibits no sequence evidence of this gene (Castro et al., 2014, 2012; Narita et al., 2010; Ordoñez et al., 2008). This species-specific distribution has been suggested to result from dietary adaptations; generally, higher levels of pepsinogens are found in the gastric mucosa of animals with an herbivorous diet, in contrast to omnivorous and carnivorous species (Kageyama, 2002).

In contrast to the other members of the pepsin family, little is known about the evolutionary history, distribution and function of *Cmy* in mammalian lineages. Despite exhibiting a conserved quaternary structure and catalytic residues, *Cmy* displays an unusual profile with low general proteolytic activity and high specificity towards milk κ-casein (Kageyama, 2002; Pearl and Blundell, 1984). Milk κ-casein, along with α- and β-caseins, belong to the secretory calcium-binding phosphoprotein gene family and provide nutritional calcium, amino acids, as well as other bioactive peptides, with putative antimicrobial activity (Caroli et al., 2009; Kawasaki and Weiss, 2003). Additionally, caseins form heterogeneous micellar structures with κ-casein coats for increased stability (Mercier et al., 1976). Cleavage by *Cmy* splits κ-casein into an insoluble para-κ-casein and a soluble caseinomacropeptide, leading to the disruption of the micelles, release of the entrapped content, and to the clotting of milk, a feature widely used in the manufacturing of dairy products (Caroli et al., 2009; Langholm Jensen et al., 2013; Mercier et al., 1976; Palmer et al., 2010). In fact, the use of *Cmy* in the manufacturing of cheese is considered to be one of the

* Corresponding author at: CIIMAR/CIMAR, U. Porto, Avenida General Norton de Matos, S/N, 4450-208 Matosinhos, Portugal.
E-mail address: filipe.castro@ciimar.up.pt (L.F.C. Castro).

http://dx.doi.org/10.1016/j.ympev.2017.08.014
Received 31 December 2016; Received in revised form 7 August 2017; Accepted 25 August 2017
Available online 26 August 2017
1055-7903/

earliest biological applications of enzymes, with remains of cheese found in Egyptian pots dating to approximately 3000-2800BCE (Palmer et al., 2010; Szecsi, 1992). Thus, research on *Cmy* has focused mainly on the characterization of biochemical, structural and functional properties for industrial purposes.

Although *Cmy* activity is directly related to milk clotting, it has been indirectly correlated with colostrum-dependent immunoglobulin transfer (Borghesi et al., 2014; Hurley and Theil, 2011; Kageyama, 2002). Several key aspects were suggested to enhance immunoglobulin transfer from colostrum: such as the mild enzymatic landscape of the neonate digestive tract, which protects immunoglobulins from proteolysis (Bela Szecsi and Harboe, 2013; Foltmann, 1992), and the presence of trypsin inhibitors in the colostrum, conferring added protection against enzymatic digestion in the small intestine (Foltmann, 1992). Thus, in colostrum dependent immunoglobulin G (IgG) transfer (passive transfer - PT), *Cmy* would contribute to the mild environment while allowing the release of micellar immunoglobulins and whey proteins (Bela Szecsi and Harboe, 2013; Foltmann, 1992). However, IgG transfer strategies vary across mammals. In humans, for instance, IgG is transferred from mother to fetus during the last stages of gestation (maternal transfer - MT). Curiously, *Cmy* was found to be a pseudogene in humans, due to a shift in the reading frame, and consequent premature stop codon, caused by a deletion in exon 4 (Örd et al., 1990). Despite the pseudogene status in humans, *Cmy* can be found in several tetrapod species (mammals, reptiles and bird (Ordoñez et al., 2008)). A cryptic orthologue was also suggested to be present in teleost genomes, through the analysis of the homologous syntenic region (Castro et al., 2014). However, contrary to the other members of the pepsin family with variable gene distribution in tetrapods, the evolutionary history and distribution of *Cmy* is largely unknown. Thus, the emerging questions are (1) whether the pseudogenized condition detected in humans is unique, or conversely, if it represents a wider genomic trait of mammals, and, if so, (2) does *Cmy* pseudogenization follow the acquisition of maternal immune transfer strategies. Here, we sought to illuminate the evolutionary history of the *Cmy* gene and its correlation with feeding and immune transfer strategies by providing an extensive analysis of available mammalian genomes.

2. Materials and methods

2.1. Sequence analysis

All major mammalian lineages with available genome data in Ensembl and GenBank were searched with blastp and blastn using as query *Bos taurus Cmy* amino acid and/or nucleotide sequence. *Cmy*-like nucleotide sequences were retrieved for the following lineages: Monotremata, Marsupialia, Cingulata, Tubulidentata, Macroscelidea, Afrosoricida, Proboscidea, Sirenia, Eulipotyphla, Chiroptera, Perissodactyla, Carnivora, Cetacea, Artiodactyla, Lagomorpha, Hystricomorpha, Sciuromorpha, Myomorpha, Scandentia, Dermoptera, Strepsirrhini and Haplorrhini (Platyrrhini, Cercopithecoidea, Hominoidea) (Accession numbers available in Supplementary Table 1). In the cases where no *Cmy* gene annotation was found, the homologous syntenic region was investigated for *Cmy*-like sequences (genomic coordinates of the searched regions are provided in Supplementary Table 1). Non-mammalian lineages, namely reptiles and birds, were also searched and the corresponding nucleotide sequences retrieved. A total of 99 *Cmy*-like nucleotide sequences were recovered and corresponding accession files were inspected to determine if the annotated RefSeq transcript was modified relative to its genomic source. If affirmative, the genomic region of the *Cmy-like* gene would be retrieved, and further examined to determine the coding status.

2.2. Gene annotation and mutational validation

Using *Bos taurus* prochymosin nucleotide sequence (NM_180994.2) as reference, each exon was isolated and mapped to the genomic region of the candidate pseudogenes using Geneious V7.1.9 map to reference tool. The aligned regions were individually screened for ORF disrupting mutations (exon deletions, sequence frameshifts and premature stop codons) and then concatenated to obtain a predicted coding sequence. Validation of the identified ORF-abolishing mutations was performed by blastn searches in available sequence read archive (SRA) and Trace Archive in NCBI (when available) using as query the nucleotide sequence of the exon containing the mutation. Blast hits were uploaded to Geneious V7.1.9 and mapped to the corresponding exon. The final alignment with SRA reads was inspected to remove poorly aligned sequences and to confirm mutation status. The validation of at least one abolishing mutation per species by SRA reads and or Trace archives reads was performed.

2.3. Phylogeny selection and correlation analysis

Initial screening and gene annotation identified 30 potential pseudogenes. The remaining 69 coding *Cmy* ORFs were selected for phylogenetic analysis. An initial sequence alignment was performed to identify and purge partial sequences from further analysis. Nucleotide sequence alignment for phylogenetic analysis was performed in MAFFT (Katoh et al., 2005; Katoh and Toh, 2008) with L-INS-I method. The resulting sequence alignment was stripped of all columns containing gaps leaving a total of 1072 positions for phylogenetic analysis. Maximum likelihood phylogenetic analysis was performed in PhyML V3.0 (Guindon et al., 2010) and the evolutionary model was determined using the smart model selection (SMS) option resulting in a GTR + G + I + F. The branch support was calculated using aBayes. The resulting tree was analysed in Fig. Tree V1.3.1 available at http://tree.bio.ed.ac.uk/software/figtree/ and rooted on the bird and reptile clade.

The analysis of the selective regime was performed exclusively in the mammalian sequences. These were aligned by codon translation in Geneious V7.1.9; exon 1 was stripped from all sequences, as well as, columns containing 90% of gaps. The final sequence alignment was submitted to the Datamonkey Webserver Suite (Pond and Frost, 2005; Pond et al., 2005) and selective strength was calculated using RELAX (Wertheim et al., 2014). Data type was set to codon and the genetic code was set to universal. For each clade analysed one analysis was run in RELAX were the target clade was set as test branch, while the remaining clades were defined as reference branches.

To determine *Cmy* distribution and correlation with IgG transfer in a phylogenetic context, the coding *Cmy* status and IgG transfer strategies (PT or MT) were coded into discrete states: 1 - coding *Cmy* and MT and 0 - non-coding *Cmy* and PT. Correlation analysis was run in Mesquite V3.2 (build 801) (Maddison and Maddison, 2017) using Pagel test of correlated (discrete) character evolution with 100 simulations (Pagel, 1994). The phylogenetic tree used in the analysis was obtained from time tree public knowledge-base (Hedges et al., 2006, 2015; Kumar and Hedges, 2011).

3. Results

3.1. Sequence analysis and gene annotation

A total of 94 species covering all major mammalian lineages were examined for the presence of *Cmy-like* sequences. For each retrieved sequence the corresponding GenBank file was inspected to determine the gene coding status. Sequence search and analysis returned a total of 30 candidate pseudogenes and no annotation of a *Cmy-like* sequence was found in 4 species: *Heterocephalus glaber* (naked mole-rat); *Octodon degus* (Degu), *Fukomys damarensis* (Damaraland mole-rat) and *Chinchilla lanigera* (common chinchilla).

All 30-candidate pseudogenes were individually inspected and re-annotated using the corresponding species-specific genomic data. The analysis of the *Homo sapiens* CMYP (OMIM#118943) genomic region,