

**Phylogenetic distribution of CMP-Neu5Ac hydroxylase (CMAH), the
enzyme synthetizing the pro-inflammatory human xeno-antigen
Neu5Gc**

Sateesh Peri¹, Asmita Kulkarni¹, Felix Feyertag¹, Patricia M. Berninsone¹, David Alvarez-
Ponce^{1,*}

¹Department of Biology, University of Nevada, Reno, Nevada, USA.

***Corresponding author:**

David Alvarez-Ponce, Department of Biology, University of Nevada, Reno, 1664 N. Virginia
Street, Reno, NV 89557. Email: dap@unr.edu

© The Author(s) 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License
(<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium,
provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

ABSTRACT

The enzyme CMP-N-acetylneuraminic acid hydroxylase (CMAH) is responsible for the synthesis of N-glycolylneuraminic acid (Neu5Gc), a sialic acid present on the cell surface proteins of most deuterostomes. The *CMAH* gene is thought to be present in most deuterostomes, but it has been inactivated in a number of lineages, including humans. The inability of humans to synthesize Neu5Gc has had several evolutionary and biomedical implications. Remarkably, Neu5Gc is a xenoantigen for humans, and consumption of Neu5Gc-containing foods, such as red meats, may promote inflammation, arthritis and cancer. Likewise, xenotransplantation of organs producing Neu5Gc can result in inflammation and organ rejection. Therefore, knowing what animal species contain a functional *CMAH* gene, and are thus capable of endogenous Neu5Gc synthesis, has potentially far-reaching implications. In addition to humans, other lineages are known, or suspected, to have lost *CMAH*; however, to date reports of absent and pseudogenic *CMAH* genes are restricted to a handful of species. Here, we analyze all available genomic data for non-deuterostomes, and 322 deuterostome genomes, to ascertain the phylogenetic distribution of *CMAH*. Among non-deuterostomes, we found *CMAH* homologs in two green algae and a few prokaryotes. Within deuterostomes, putatively functional *CMAH* homologs are present in 184 of the studied genomes, and a total of 31 independent gene losses/pseudogenization events were inferred. Our work produces a list of animals inferred to be free from endogenous Neu5Gc based on the absence of *CMAH* homologs and are thus potential candidates for human consumption, xenotransplantation research, and model organisms for investigation of human diseases.

Keywords: Neu5Gc, *CMAH*, pseudogene

INTRODUCTION

Sialic acids are a family of more than 50 nine-carbon sugars that are typically found at the terminal ends of N-glycans, O-glycans, and glycosphingolipids that are secreted or attached to the cell membrane. They are involved in recognition processes, frequently serving as ligands for receptor-mediated interactions that enable intercellular or host-pathogen recognition. Sialic acids may also function as a class of “Self-associated molecular patterns” (SAMPs), whose presence/absence in a species serves as a signal to modulate innate immune responses (Varki 2011). Sialic acids are found predominantly in deuterostomes (the group including chordates, hemichordates and echinoderms), being uncommon in other organisms (Warren 1963; Corfield and Schauer 1982; Staudacher 1999; Angata and Varki 2002; Schauer 2004).

The most common sialic acids are N-acetylneuraminic acid (Neu5Ac) and N-glycolylneuraminic acid (Neu5Gc). The enzyme cytidine monophospho-N-acetylneuraminic acid hydroxylase (*CMAH*) catalyzes the synthesis of Neu5Gc by hydroxylation of Neu5Ac (Schauer, et al. 1968; Schoop, et al. 1969; Schauer 1970). *CMAH*^{-/-} mice lack Neu5Gc, indicating that *CMAH* is the only enzyme capable of synthesizing Neu5Gc (Hedlund, et al. 2007; Bergfeld, et al. 2012). This enzyme and/or its encoding gene has been found in many chordates, in echinoderms and in a hemichordate (Kawano, et al. 1995; Martensen, et al. 2001; Varki 2009; Bergfeld, et al. 2012; Ikeda, et al. 2012). In non-deuterostomes, homologous *CMAH* sequences have recently been detected only in two green algae and in a few prokaryotes, with phylogenetic analyses suggesting a horizontal gene transfer event from green algae to deuterostomes (Simakov, et al. 2015).

Consistently, Neu5Gc has been reported in a variety of deuterostomes including echinoderms, fish, amphibians, and the majority of mammals studied so far (see exceptions below) (Muralikrishna, et al. 1992; Klein, et al. 1997; Tangvoranuntakul, et al. 2003; Schauer, et al. 2009; Samraj, et al. 2015). It is generally thought that non-deuterostomes cannot synthesize Neu5Gc, even though some can incorporate it from the environment (Schauer, et al. 1983).

Some deuterostomes, including humans, have undergone inactivation or loss of the gene *CMAH*, and therefore have lost the ability of synthesizing Neu5Gc. In an ancestor of humans, an Alu-mediated deletion removed a region of the genome encompassing a 92-bp exon of *CMAH*

(Irie and Suzuki 1998; Hayakawa, et al. 2001; Chou, et al. 2002). Deletion of this exon resulted in a frameshift mutation, as a result of which the human protein is only 72 amino acids long and non-functional (the full ancestral *CMAH* protein was 590 amino acids long). Therefore, human tissues exhibit very low levels of Neu5Gc (Muchmore, et al. 1998; Tangvoranuntakul, et al. 2003; Diaz, et al. 2009), which are probably the result of incorporation from animal foods. It has been estimated that this inactivation took place 2.5-3 million years ago (Chou, et al. 2002) and was fixed rapidly in the population, probably with the intervention of positive selection (Hayakawa, et al. 2006). Inactivation of *CMAH* may have affected human biology in multiple ways (for a comprehensive review, see Varki 2009; Okerblom and Varki 2017). First, it may have freed our ancestors from pathogens that require attaching to Neu5Gc for infection, such as *Plasmodium reichenowi* (responsible for malaria in chimpanzees and gorillas; Martin, et al. 2005), *E. coli* K99 (Kyogashima, et al. 1989), transmissible gastroenteritis coronavirus (Schwegmann-Wessels and Herrler 2006), and simian virus 40 (Campanero-Rhodes, et al. 2007). Nonetheless, inactivation of *CMAH* probably made humans susceptible to pathogens preferentially recognizing Neu5Ac, such as *Plasmodium falciparum* (Martin, et al. 2005) and *Streptococcus pneumoniae* (Hentrich, et al. 2016). Remarkably, *P. falciparum* emerged from *P. reichenowi* after inactivation of *CMAH* in humans (Rich, et al. 2009; Varki and Gagneux 2009). Second, the *CMAH* pseudogene may have been driven to fixation via sexual selection. In the time in which the presence of a functional *CMAH* gene was polymorphic in the ancestral hominin population, anti-Neu5Gc antibodies in the reproductive tract of Neu5Gc-negative females may have targeted Neu5Gc-containing sperm or fetal tissues, thus reducing reproductive compatibility (Ghaderi, et al. 2011). Third, loss of Neu5Gc may have unchained a series of changes in human sialic acid biology and its controlling genes. Out of the less than 60 genes known to be involved in sialic acid biology, at least 10 have undergone human-specific changes, some of which have been linked directly to Neu5Gc loss (Altheide, et al. 2006; Varki and Varki 2007; Varki 2009). Fourth, the phenotypes of *CMAH*^{-/-} mice suggest that loss of Neu5Gc may have contributed to a number of human-specific diseases (Hedlund, et al. 2007; Chandrasekharan, et al. 2010; Kavalier, et al. 2011).

Inactivation of *CMAH* in humans meant that Neu5Gc became a foreign antigen. Neu5Gc from animal foods (predominantly red meats and milk products; Tangvoranuntakul, et al. 2003;

Samraj, et al. 2015) is incorporated into the glycoproteins of human tissues (Tangvoranuntakul, et al. 2003; Bardor, et al. 2005; Banda, et al. 2012), where it is thought to elicit an immune response that may result in chronic inflammation, rheumatism and cancer (Varki and Varki 2007). This may explain, at least in part, the link between red meat consumption and cancer (Rose, et al. 1986; Giovannucci, et al. 1993; Fraser 1999; Tavani, et al. 2000; Willett 2000; Linseisen, et al. 2002; Bosetti, et al. 2004; Zhang and Kesteloot 2005; Tseng, et al. 2015). In support of this hypothesis, Neu5Gc often concentrates in human tumors and sites of inflammation (Malykh, et al. 2001; Tangvoranuntakul, et al. 2003; Diaz, et al. 2009), and *CMAH*^{-/-} mice develop systemic inflammation and a high frequency of cancer when fed with bioavailable Neu5Gc (Samraj, et al. 2015).

In the context of animal-based xenotransplantation, tissues from animals expressing Neu5Gc have been shown to cause human recipients to develop antibodies against Neu5Gc, triggering inflammation and contributing to delayed tissue rejection (Salama, et al. 2015; Hurh, et al. 2016). Neu5Gc has been found in most tissues of pigs such as heart, kidney, liver, pancreas (Bouhours and Bouhours 1988; Bouhours, et al. 1996; Diswall, et al. 2010) and in adult pig pancreatic islets (Komoda, et al. 2004), raising concerns over their application in xenotransplantation.

In addition to humans, *CMAH* has also been reported or suggested to be inactivated or lost in other animal lineages independently. New World monkeys underwent inactivation of *CMAH* ~30 million years ago due to inversion of exons 4-13 and loss of exons 4-8 and 10-13. This may explain why they are susceptible to certain human pathogens, such as *P. falciparum* (Springer, et al. 2014). In the ferret genome, the first 9 exons of the gene have been lost, and PCR analyses did not detect conserved portions of the gene in a number of pinnipeds and musteloids, indicating that *CMAH* was inactivated in an ancestor of pinnipeds and musteloids (Ng, et al. 2014). Sequence similarity searches against the genomes of chicken and zebra finch did not detect any *CMAH* homolog (Schauer, et al. 2009) and southern blot analysis did not detect expression of the gene in chicken liver (Kawano, et al. 1995). Consistently, Neu5Gc has been shown to be rare in birds and reptiles (Fujii, et al. 1982; Schauer and Kamerling 1997; Ito, et al. 2000; Schauer, et al. 2009). These observations led to the hypothesis that the *CMAH* gene

may have been lost in an ancestor of Sauropsida (reptiles and birds) (Schauer, et al. 2009). According to this hypothesis, the low amounts of Neu5Gc detected in some reptiles and birds may have been incorporated from the diet. Analysis of the platypus genome did not reveal any *CMAH* homolog, and Neu5Gc was not detected in platypus muscles or liver (Schauer, et al. 2009), or in the milk of the Australian spiny anteater echidna (Kamerling, et al. 1982), suggesting that *CMAH* was also lost in an ancestor of extant monotremes.

In cats, blood types A, B and AB are determined by the presence or absence of Neu5Gc on certain erythrocyte glycolipids, and the absence of Neu5Gc in type B cats might be due to mutations in the 5' UTR or the protein-coding regions of *CMAH* (Bighignoli, et al. 2007; Omi, et al. 2016). Depending on their geographical origin, dog breeds express Neu5Gc (Northern China, Korea and Southern Japan) or not (Europe, Hokkaido dog from North of Japan) (Yasue, et al. 1978; Hashimoto, et al. 1984). The molecular basis of why certain dog breeds lack Neu5Gc is yet to be elucidated.

Given the biological and biomedical relevance of Neu5Gc, it is important to know what animals have the ability to synthesize it. Knowing the exact lineages in which *CMAH* has been inactivated or lost will allow scientists to identify lineages that may have experienced changes similar to those experienced by humans, and will have implications for human nutrition and xenotransplantation. However, to date all studies of *CMAH* evolution have been restricted to specific organisms or groups of organisms. Here, we conduct the first comprehensive analysis of the evolution of *CMAH*. For that purpose, we analyzed all available genomic data in the National Center for Biotechnology Information (NCBI) databases, and the genomes of 323 deuterostomes. We found that the gene has been lost or inactivated at least 31 times during deuterostome evolution.

MATERIALS AND METHODS

Genomic dataset and determination of *CMAH* presence/absence

We used all genomic data available from the NCBI Genome database to determine the occurrence of *CMAH* in non-deuterostomes. We then focused on the 321 deuterostome genomes and two echinoderm *CMAH* mRNA sequences available in the NCBI Genome

database as of January 2017 (NCBI Coordinators 2016). An initial screening for the presence or absence of the *CMAH* gene was done using sequence similarity searches (BLASTP and TBLASTN searches) using the chimpanzee *CMAH* protein sequence as query. If BLASTP searches failed for a given organism (which depend on *CMAH* proteins being annotated), TBLASTN searches were conducted (which can detect unannotated sequences). All BLAST searches were performed using an E-value cut-off of 10^{-10} (Altschul, et al. 1990). All genomes, depending on the level of assembly, were either screened in the NCBI RefSeq genomic database (non-redundant, well-annotated reference sequence database) or in the NCBI WGS database (genome assemblies of incomplete genomes with or without annotation).

Gene annotation and curation

Gene annotations, particularly for non-model organisms, are known to be subjected to high rates of error (Devos and Valencia 2001; Tu, et al. 2012). For each *CMAH* homolog, the nucleotide sequence of the coding region (CDS) was translated *in silico* and aligned with the chimpanzee sequence using ProbCons, version 1.12 (Do, et al. 2005). The resulting protein sequence alignment was used to guide the alignment of the CDSs. The CDS alignments were visualized using BioEdit, version 7.0.0 (Hall 1999). Where possible, all gene annotation errors were fixed manually. Erroneous and extra exons (not showing significant similarity to the chimpanzee sequence) were removed. Missing exons (present in the chimpanzee sequence but not in the species of interest) were searched for in the genome using the chimpanzee sequence as query in TBLASTN or BLASTN searches. If the missing exon could not be detected, a careful analysis was conducted to determine whether the exon had been lost or it was part of an unsequenced region. Given the small size of the coding segment of exon 1 (only 8 nucleotides), no attempt was made to annotate this exon.

Where possible, signatures of pseudogenization (premature stop codons, frameshift mutations and exon losses) were verified by visualizing the original chromatograms in the Sequence Read Archive (SRA) database. *CMAH* sequences were only considered pseudogenes if at least one pseudogenization signature was confirmed.

Phylogenetic analysis

The trees represented in Figures 1-4 were derived from the NCBI Taxonomy database (Sayers, et al. 2009) using PhyloT (Letunic and Bork 2007). Relevant polytonies observed in the orders Caniformia and Chiroptera were resolved based on prior phylogenetic analyses (Tsagkogeorga, et al. 2013; van Valkenburgh, et al. 2014; Lei and Dong 2016). The trees were visualized using the TreeGraph software (Stöver and Müller 2010).

Phylogenetic analyses were used to better characterize certain gene duplication events in fish and non-vertebrate deuterostomes and to discard lateral gene transfer. These analyses were conducted using IQ-Tree (Nguyen et al. 2015) with default parameters. Only sequences covering at least 20% of the length of the gene were included in our phylogenetic analyses.

Purifying selection analysis

The *Xenopus tropicalis* and *Xenopus laevis* *CMAH* sequences are annotated as pseudogenes in the NCBI RefSeq database. We aligned the two CDSs and estimated the nonsynonymous to synonymous divergence ratio, d_N/d_S , using the codeml program of the PAML package, version 4.4 (Yang 2007). We tested whether this d_N/d_S ratio was significantly different from 1 using a likelihood ratio test. Twice the difference of the log-likelihoods of both models (M0 with a free d_N/d_S ratio vs. M0 with a fixed d_N/d_S of 1) was assumed to follow a χ^2 distribution with one degree of freedom.

RESULTS AND DISCUSSION

CMAH in non-deuterostomes

We first used the chimpanzee *CMAH* protein sequence (Chou, et al. 1998) as query in a BLASTP search against the NCBI nr database, excluding all sequences from deuterostomes. The first two hits corresponded to the green algae *Ostreococcus tauri* and *Micromonas commoda*, and 60 prokaryotes, including representatives of proteobacteria, firmicutes, cyanobacteria, actinobacteria, nitrospirae, FCB group (Fibrobacteres, Chlorobi, and Bacteroidetes), and archaeobacteria (Supplementary table S1). No hits were detected in any other group, including non-deuterostome animals. The distribution of *CMAH* homologs in non-

deuterostomes is equivalent to that observed by Simakov et al. (2015), who proposed that the *CMAH* gene could have been passed onto the deuterostome lineage by green algae through horizontal gene transfer.

***CMAH* in non-vertebrate deuterostomes**

We next focused our analyses on the 322 deuterostome genomes available from the NCBI Genome database, including those for 8 echinoderms, 2 hemichordates, 3 cephalochordates, 5 urochordates (tunicates), and 304 vertebrates (Table 1). These represent almost completely sequenced genomes. The *CMAH* mRNA sequence for another two echinoderm species were obtained from the NCBI nr database: *Asterias rubens* (ID: AJ308602.1) (Martensen, et al. 2001) and *Hemicentrotus pulcherrimus* (ID: AB699316.1). In each genome, gene similarity searches (BLASTP and/or TBLASTN searches using the chimpanzee *CMAH* protein as query) were used to identify *CMAH* homologs. Where necessary, gene annotations were curated manually to resolve the intron/exon structure, and to determine signatures of inactivation/pseudogenization (missing exons, frameshift mutations and premature stop codons).

CMAH homologs were detected in all of the studied echinoderms, hemichordates and cephalochordates; however, no *CMAH* orthologs (active or inactive) were detected in urochordates. These results indicate that *CMAH* was present in the ancestor of deuterostomes, and it was lost in an ancestor of urochordates. Figure 1 summarizes the distribution of *CMAH* homologs across non-vertebrate deuterostomes, and Supplementary table S2 contains a detailed description of all found genes. Supplementary dataset 1 is a multiple sequence alignment of representative non-vertebrate deuterostome sequences.

We found putatively functional copies in all 10 echinoderm species studied. *Acanthaster planci* (crown-of-thorns starfish) has two copies: one putatively functional and another putatively inactivated due to a premature stop codon in coding exon 10 (Table 2). Our observations are consistent with prior works that have reported the presence of the *CMAH* enzyme, its encoding gene, or Neu5Gc in all echinoderms studied so far. The enzyme was purified from gonads of *Asterias rubens* (common star fish), *Ctenodiscus crispatus* (mud star), *Strongylocentrotus*

pallidus (pale sea urchin), and a species of *Holothuria* (a sea cucumber) (Gollub and Shaw 2003). The *Asterias rubens* *CMAH* cDNA was subsequently cloned and sequenced (Martensen, et al. 2001), revealing a highly conserved *CMAH* coding sequence. Neu5Gc was detected in whole body extracts of *Ophioderma brevispina* (a brittle star), *Nemaster rubiginosa* (sea lily), and *Sclerodactyla briareus* (a sea cucumber) (Warren 1963; Sumi, et al. 2001) and in the egg jelly coat of *Paracentrotus lividus* (a sea urchin) (Yeşilyurt, et al. 2015).

Our searches identified *CMAH* homologs in the two hemichordates studied, *Ptychodera flava* (yellow acorn worm) and *Saccoglossus kowalevskii* (acorn worm). These sequences were also noted by Simakov et al (Simakov, et al. 2015).

We identified *CMAH* homologous sequences in all the studied cephalochordates, including one copy in *Asymmetron lucayanum* (Bahama lancelet), five in *Branchiostoma belcheri* (Chinese amphioxus) and seven in *Branchiostoma floridae* (Floridan amphioxus). One of the *B. belcheri* copies had a premature stop codon in coding exon 6, and one of the *B. floridae* copies had a premature stop codon in coding exon 13. Our results are in agreement with a previous report of Neu5Gc in *B. belcheri* (belcher's lancelet) (Guérardel, et al. 2012).

We did not find *CMAH* homologs in any of the 5 urochordate genomes studied, indicating that *CMAH* was lost in an ancestor of urochordates. These results are in agreement with prior observations that tunicates are devoid from sialic acids (Warren 1963).

***CMAH* in fish**

Several groups have independently reported non-significant levels (< 2% of the sialic acid fraction) of Neu5Gc in the meat of different fish species belonging to the classes Actinopterygii (including tilapia, yellowfin tuna, mahi mahi, swordfish, rainbow trout and sardines, crucian carp, grass carp, golden pomphret and European seabass) and Chondrichthyes (including the thresher shark) (Chen, et al. 2014; Samraj, et al. 2015). The roe of salmon and whitefish, in contrast, exhibit high Neu5Gc concentrations (Samraj, et al. 2015). Our sequence similarity searches against 77 fish genomes belonging to classes Cephalaspidomorphi (lampreys, n = 2), Chondrichthyes (cartilaginous fishes, n = 3), Sarcopterygii (lobe-finned fishes, n = 1), and Actinopterygii (ray-finned fishes, n = 71) show that the gene is present in classes

Sarcopterygii (1 species) and Actinopterygii (71 species). A total of 16 gene loss events were inferred to have occurred in the fish lineages, including complete gene losses at the ancestors of Cephalaspidomorpha and Chondrichthyes (Figure 2).

Gene duplicates were observed in *Sinocyclocheilus anshuiensis*, *Sinocyclocheilus grahami* and *Sinocyclocheilus rhinoceros*, and our phylogenetic analyses indicate a duplication in a common ancestor. In addition, species-specific *CMAH* gene duplicates were observed in the genomes of *Maylandia zebra* (zebra mbuna), *Neolamprologus brichardi* (princess cichlid), *Salmo salar* (Atlantic salmon), *Xiphophorus maculatus* (southern platyfish), *Haplochromis burtoni* (Burton's mouthbrooder) and *Oreochromis niloticus* (Nile tilapia). The last two species having premature stop codon in coding exons 8 and 5, respectively, in their duplicate copies (Table 2).

***CMAH* in amphibians**

There are 4 amphibian genomes available in the NCBI Genome database, including three frogs and one salamander (Table 1). We found putatively functional *CMAH* homologs in all three frogs, *X. tropicalis* (Western clawed frog), *X. laevis* (African clawed frog) and *Nanorana parkeri* (high Himalaya frog). However, no *CMAH* homolog was detected in the salamander *Ambystoma mexicanum* (axolotl). These results suggest that the ancestor of amphibians had a functional *CMAH*, and that it was lost in an ancestor of salamanders (Figure 3). Supplementary dataset 2 is a multiple sequence alignment including these sequences.

Previous studies did not detect Neu5Gc in the liver of the frog *Rana esculata* (Schauer, et al. 1980) or in the brain gangliosides of *X. laevis* (Rizzo, et al. 2002). It should be noted, nonetheless, that Neu5Gc is usually not detected in vertebrate brain, where it appears to have adverse effects (Naito-Matsui, et al. 2017). The red blood cells of the salamander *Amphiuma means* also have been reported to contain only Neu5Ac (Pape, et al. 1975). However, Neu5Gc has been reported in the oviductal mucins of fire-bellied toads (*Bombina bombina* and *Bombina variegata*) and the alpine newt (*Triturus alpestris*) (Schauer, et al. 2009). In addition, *CMAH* has been identified as one of the key genes downregulated during oocyte maturation in *X. laevis* (Gohin, et al. 2010).

The *CMAH* genes of *X. tropicalis* and *X. laevis* have been classified as pseudogenes in the NCBI RefSeq database (Gene IDs: 100216283, 379989). However, this has not been validated by any study so far. To gain insight into the functionality of these genes, we calculated the non-synonymous to synonymous divergence ratio (d_N/d_S). This value was significantly lower than one ($d_N/d_S = 0.142$, $2\Delta l = 109.03$, $P = 1.60 \times 10^{-25}$), indicating strong purifying selection, and thus strongly suggesting functionality of the *CMAH* enzyme in *Xenopus*. Therefore, we suspect that the genes have been erroneously annotated as pseudogenes by the automatic gene annotation pipelines.

Our results suggest that *X. laevis* and other frogs previously reported to possess only Neu5Ac contain a *CMAH* enzyme with all its conserved domains. However, it is inconclusive at this point whether frogs, like fish, exhibit non-significant levels of Neu5Gc in most of their tissues despite having an intact *CMAH* coding sequence.

***CMAH* in reptiles and birds**

Sequence similarity searches against the chicken and zebra finch genomes revealed no *CMAH* homologs (Schauer, et al. 2009). In agreement, Schauer et al (Schauer, et al. 2009) reported the absence of Neu5Gc in several species of birds (chicken, duck, turkey, goose, ostrich, emu, scarlet macaw, budgerigar, swallow and oriental swiftlet) and non-bird reptiles (green iguana, agama, green basilisk hatchling, anaconda, hundred pace viper, Taiwan stings snake, Taiwan beauty snake, crocodile and Amboina box turtle). This led to the suggestion that *CMAH* may have been lost in an ancestor of Sauropsida (reptiles and birds). Neu5Gc was found in the gastrointestinal tract of ducks (Ito, et al. 2000), the eggs of the budgerigar, and in both egg and tissues of an adult green basilisk (Schauer, et al. 2009). However, the source of Neu5Gc in these species is yet to be established, with diet being a likely origin (Schauer, et al. 2009).

We queried the genomes of 73 birds and 19 non-bird reptiles for *CMAH* homologs (Table 1). Surprisingly, we found homologs in the two lizards included in our dataset, *Anolis carolinensis* (green anole lizard), and *Gekko japonicus* (Japanese gecko), but not in any snake, turtle, crocodilian or bird. These results contradict the hypothesis that *CMAH* may have been lost in an ancestor of Sauropsida (Schauer, et al. 2009). Instead, our results indicate that the gene was

present in the most recent common ancestor of Sauropsida, and that it was lost at least twice during the evolution of the group: in the snake lineage, and in an ancestor of turtles, crocodylians and birds (Figure 3). The *A. carolinensis* *CMAH* sequence appears to be functional. The *G. japonicus* sequence, however, contains a premature stop codon at the end of coding exon 4. This indicates a recent inactivation of *CMAH* in an ancestor of *G. japonicus* (Table 2).

Our results indicate that the Neu5Gc observed in the duck gastrointestinal tract and in budgerigar eggs is not endogenous, strongly supporting the hypothesis that it may have been incorporated from the diet. However, the Neu5Gc observed in the eggs and adult tissues of the green basilisk (a lizard) might be endogenous, as our observations indicate the presence of *CMAH* in lizards.

***CMAH* in mammals**

The NCBI Genome database contains the genomes of one monotreme, three marsupials and 126 placental mammals (Table 1). We found putatively functional *CMAH* homologs in the genomes of all three marsupials and in 110 of the studied placental genomes, but none in the platypus (a monotreme). Our analyses suggest that at least ten *CMAH* gene loss or inactivation events occurred during mammalian evolution (mapped in Figure 4). Four of these events have been described previously, including one in the human lineage (Chou, et al. 1998), one in an ancestor of New World monkeys (Springer, et al. 2014), one in an ancestor of pinnipeds and musteloids (Ng, et al. 2014), and one in an ancestor of platypus. The other five events are described here for the first time, including one in an ancestor of the sperm whale (*Physeter catodon*), two events in bats, one in an ancestor of the white-tailed deer (*Odocoileus virginianus*), and another in an ancestor of the European hedgehog (*Erinaceus europaeus*) (Figure 4). Supplementary dataset 3 is a multiple sequence alignment of all mammalian vertebrates.

Neu5Gc is known to be present in most placental mammals studied so far (e.g., Davies, et al. 2012) and has been reported in kangaroo (Schauer, et al. 2009). In contrast, among monotremes Neu5Gc could not be detected in liver and muscle tissues of the platypus (*Ornithorhynchus anatinus*) (Schauer, et al. 2009), nor in the milk of the spiny anteater echidna (*Tachyglossus aculeatus*) (Kamerling, et al. 1982). Schauer et al (2009), did not find any *CMAH*

homolog in the genome of platypus using BLAST searches. However, our analyses revealed a *CMAH* pseudogene in the platypus genome. Using TBLASTN searches we could retrieve eight of the protein-coding exons (the chimpanzee *CMAH* has 15 protein-coding exons). Exon 5 harbors a premature stop codon, which may explain the lack of Neu5Gc in this species (Table 2).

Schauer et al. (2009) identified a *CMAH* homolog in the genome of the marsupial *Monodelphis domestica* (gray short-tailed opossum). Consistent with this observation, we found *CMAH* homologous sequences in the genomes of gray short-tailed opossum, tammar wallaby and Tasmanian devil.

An inactivation of *CMAH* is known to have occurred in a human ancestor, approximately 2.5-3 million years ago (Hayakawa, et al. 2001; Chou, et al. 2002). An Alu insertion resulted in a deletion of a genomic region encompassing coding exon 3, a 92-bp exon that codes for part of the Rieske catalytic domain. Our genomic analyses also found all the protein coding exons of the human gene except exon 3 (relative to chimp coding sequence).

The presence of inactive *CMAH* sequences, and the absence of Neu5Gc, was also reported in a number of New World monkeys, indicating that *CMAH* pseudogenized in an ancestor of New World monkeys (Springer, et al. 2014). In agreement with these findings, we found a deletion spanning coding exons 3 to 15 in all available New World monkey genomes, including those of *Saimiri boliviensis boliviensis* (Bolivian squirrel monkey), *Cebus capucinus* (white-faced sapajou), *Callithrix jacchus* (marmoset) and *Aotus nancymaae* (Nancy Ma's night monkey).

Ng et al (2014), investigated the ferret genome and performed PCR analyses on another 14 species of musteloids and two species of pinnipeds. In all species investigated, they found a large deletion of nine protein-coding exons in the *CMAH* gene, suggesting a pseudogenization event in a common ancestor of musteloids and pinnipeds. In agreement with this hypothesis, we found no putatively functional *CMAH* genes in any of the musteloid and pinniped species investigated: The ferret (*Mustela putorius furo*) genome contains a disrupted *CMAH* gene, and the genomes of the Weddell seal (*Leptonychotes weddellii*) and the walrus (*Odobenus rosmarus divergens*) lack *CMAH* homologs.

We found a putatively functional *CMAH* coding sequence in the dog genome, consistent with prior results from Ng et al. (2014), who showed the presence of conserved *CMAH* coding exons 3, 5, 8, 11 and 12 in dogs. The dog genome assemblies available from the NCBI Genome database correspond to three breeds of European ancestry (boxer, poodle and beagle). Dogs of Western ancestry seem to lack significant levels Neu5Gc, but Western dog breed cells exhibit low levels of canine *CMAH* transcripts and protein (Löfling, et al. 2013), consistent with our observations.

Bats (order Chiroptera) are commonly classified into suborders Yangchiroptera and Yinchiroptera (Teeling, et al. 2005; Tsagkogeorga, et al. 2013; Lei and Dong 2016). While one study showed that the milk of island flying fox, *Pteropus hypomelanus* (Pteropodidae, Yinchiroptera) contains Neu5Gc (Senda, et al. 2011), no studies have reported the levels of Neu5Gc in other bat species. The NCBI Genome database contains the genomes of seven Yinchiroptera and six Yangchiroptera species. We found *CMAH* sequences in the families Pteropodidae and Megadermatidae of Yinchiroptera and in the family Mormoopidae of Yangchiroptera. In contrast, *CMAH* gene could not be detected in two species belonging to Rhinolophidae and Hipposideridae (Yinchiroptera) and in all five species belonging to Vespertilionidae (Yangchiroptera). Thus, the *CMAH* gene has undergone independent gene losses in both Yinchiroptera and Yangchiroptera (Figure 4).

The NCBI Genome database contains the genomes of 14 members of the taxon Ruminantia (10 bovids, 2 giraffids and 2 cervids). All bovid and giraffid species represented in the database contain a putatively functional *CMAH* gene, consistent with the high levels of Neu5Gc reported in beef (Samraj, et al. 2015). Among cervids, *CMAH* was present in the European roe deer (*Capreolus capreolus*), but not in the white-tailed deer (*O. virginianus*), indicating that the gene was lost specifically in the *O. virginianus* lineage. Of note, gangliosides isolated from antlers of sika deer (*Cervus nippon*) have been shown to contain both Neu5Ac and Neu5Gc (Jhon, et al. 1999).

Our study involved 6 cetacean genomes, which exhibited a putatively functional *CMAH* gene, with the exception of sperm whale (*P. catodon*), in which exon 5 (relative to the chimpanzee *CMAH*) was missing. Inspection of the scaffold including exons 4 and 6 did not

reveal any unsequenced region, and BLAST searches against the *P. catodon* genomic data revealed no sequences similar to exon 5. These observations indicate that *CMAH* might have pseudogenized in an ancestor of the sperm whale. Terabayashi et al. (1992) analyzed brain gangliosides of 10 cetacean species, reporting the presence of low levels of Neu5Gc in only three species: sperm whale, Dall's porpoise and killer whale. These results are in contrast with the fact that vertebrates generally do not express Neu5Gc in the brain, where it is believed to have adverse effects (e.g., Naito-Matsui et al. 2017). In addition, these species are carnivorous, thus raising the possibility that the observed Neu5Gc might have a dietary origin.

Our dataset included three genomes of the group Insectivora. Two of these genomes, *Sorex araneus* (common shrew) and *Condylura cristata* (star-nosed mole) contained a putatively functional *CMAH* gene. In contrast the *CMAH* sequence of *E. europaeus* (common hedgehog) contained a premature stop codon in coding exon 12.

Mammalian species in which we found putatively functional *CMAH* genes include a number of species in which Neu5Gc has previously been described, including pig (Malykh, et al. 1998), sheep (Koizumi, et al. 1988), cow, horse, elephant, dolphin, chimpanzee, macaque, mouse, rat, and rabbit (Davies, et al. 2012).

Evolution of the *CMAH* gene structure

The chimpanzee *CMAH* coding sequence (CDS) consists of 15 exons. The Rieske iron-sulphur domain, which contains the active site of the *CMAH* enzyme, is sequenced by exons 2-4 (position 43-345 bp in the chimp CDS). The structure of the gene is generally well conserved across deuterostomes, with the following exceptions. First, echinoderms and hemichordates exhibit an extra intron that interrupts exon 11, at a position that is equivalent to position 1329 of the chimpanzee CDS (see Supplementary dataset 3). Second, all studied fish, except *Latimeria chalumnae* (coelacanth) and *Lepisosteus oculatus* (spotted gar), exhibit an extra intron that interrupts exon 6 (after nucleotide 702 of the chimpanzee coding sequence). This suggests an intron insertion in teleost fishes. Third, the *CMAH* gene of *A. carolinensis* lacks the intron between exons 6 and 7 (after nucleotide 768 of the chimpanzee coding sequence).

Potential implication of our findings

We have characterized the phylogenetic distribution of the *CMAH* gene. Among non-deuterostomes, the gene is present in two green algae and in a handful of bacteria and archaea. Within deuterostomes, potentially functional *CMAH* homologs are present in 184 of the 323 genomes studied. Mapping the presence and absences of putatively functional *CMAH* homologs onto the deuterostome phylogeny allowed us to infer a total of 31 independent gene loss or pseudogenization events (Figures 1-4). Our inferred gene trees (Figures S1-S3) do not contradict the species tree (Figures 1-4), indicating that lateral gene transfer does not account for the observed phylogenetic distribution. A few of these events had already been described, including those in the human (Chou, et al. 1998) and platypus (Schauer, et al. 2009) lineages, an ancestor of New World monkeys (Springer, et al. 2014), and an ancestor of pinnipeds and musteloids (Ng, et al. 2014). The other 27 events represent new discoveries. At one point, our results contradict a prior hypothesis. Based on the observations that Neu5Gc is rare in reptiles and birds, and that *CMAH* is absent in birds, Schauer et al. (2009) suggested that *CMAH* may have been lost in an ancestor of Sauropsida. Our analyses, however, identified a putatively functional *CMAH* sequence in the green anole lizard *A. carolensis*, implying that the most recent common ancestor of Sauropsida had a functional *CMAH* gene, which was then lost both in the snake lineage and in an ancestor of turtles, crocodylians and birds (Figure 3). The fact that *CMAH* was lost so many times during the evolution of deuterostomes strongly suggests that the gene is not essential. However, given the relevance of Neu5Gc (e.g., as part of the ancestral SAMP; Varki 2011), its loss probably needs to be compensated by adjustments in sialic acid biology.

Due to the incompleteness of all available deuterostome genome assemblies (Chain, et al. 2009), 41 of the coding sequences of putatively functional *CMAH* genes identified in our study have some unsequenced fraction (typically, one coding exon is unsequenced; Supplementary table S2; Supplementary datasets 1-3). It is possible that some of these unsequenced regions may contain pseudogenization signatures (premature stop codons or frameshift mutations). In addition, our study has not considered the *CMAH* promoter. Therefore, it is possible that some of the *CMAH* homologs classified as “putatively functional” in our study might actually be

pseudogenes. Finally, for most species only one genome is available, making it impossible to detect polymorphic variants of the *CMAH* gene.

We expect that animals with a putatively functional *CMAH* gene should be able to synthesize endogenous Neu5Gc, whereas those with pseudogenic or absent *CMAH* homologs should not. Given the toxicity of Neu5Gc for humans, determining what animals lack the capability of synthesizing this sialic acid is key from the point of view of human nutrition and xenotransplantation research. Prior studies have quantified Neu5Gc levels in the tissues of many animals. It should be noted, however, that Neu5Gc can be incorporated from the diet (Tangvoranuntakul, et al. 2003), which means that finding Neu5Gc in the tissues of a certain organism does not imply that the organism can synthesize Neu5Gc endogenously. We have produced a list of species that lack a functional *CMAH* gene, and that should be free from Neu5Gc if fed with a Neu5Gc-free diet. These species are thus interesting candidates for human consumption and/or xenotransplantation. Our list includes, for instance, all poultry, 23 species of fish and the white-tailed deer.

Loss of Neu5Gc triggered a cascade of changes in the sialic acid biology of humans, with several evolutionary and biomedical consequences (for review, see Okerblom and Varki 2017). The animals lacking a functional *CMAH* gene identified in this study may have undergone similar changes, making them ideal model organisms for the study of human sialic acid biology and its related diseases. For instance, the altered sialic acid profile of humans makes us susceptible to pathogens using Neu5Ac for binding and recognition of the host, including *P. falciparum* (Martin, et al. 2005), *S. pneumoniae* (Hentrich, et al. 2016), and Influenza type A virus (Rogers and Paulson 1983). Ferrets, which lack a functional *CMAH* gene (Ng, et al. 2014), are used as model organisms to study the transmission mechanisms of human-adapted influenza-A virus strains (Ng, et al. 2014). Likewise, New World monkeys, which also lack a functional *CMAH* gene, have been proposed as model organisms for the study of the effects of anti-Neu5Gc antibodies in xenotransplantation (Salama, et al. 2015). In addition, species lacking a functional *CMAH* gene are potential reservoirs for Neu5Ac-binding human pathogens (Chothe, et al. 2017). For instance, bats are asymptomatic hosts to viruses like Hantaviruses (Guo et al. 2013), which

might lack the antigenic Neu5Gc on their viral envelopes. Our study very significantly expands the list of animals with these characteristics.

ACKNOWLEDGEMENTS

This work was supported by funds from the University of Nevada, Reno awarded to DAP, and by a grant from the National Institute of General Medical Sciences (P20GM103440) from the National Institutes of Health. PB was supported by grant P20GM103554 from the National Institute of General Medical Sciences of the National Institutes of Health.

REFERENCES

- Altheide TK, Hayakawa T, Mikkelsen TS, Diaz S, Varki N, Varki A. 2006. System-wide genomic and biochemical comparisons of sialic acid biology among primates and rodents: Evidence for two modes of rapid evolution. *J Biol Chem* 281:25689-25702.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.
- Angata T, Varki A. 2002. Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. *Chem Rev* 102:439-469.
- Banda K, Gregg CJ, Chow R, Varki NM, Varki A. 2012. Metabolism of vertebrate amino sugars with N-glycolyl groups: mechanisms underlying gastrointestinal incorporation of the non-human sialic acid xeno-autoantigen N-glycolylneuraminic acid. *J Biol Chem* 287:28852-28864.
- Bardor M, Nguyen DH, Diaz S, Varki A. 2005. Mechanism of uptake and incorporation of the non-human sialic acid N-glycolylneuraminic acid into human cells. *J Biol Chem* 280:4228-4237.
- Bergfeld AK, Pearce OM, Diaz SL, Pham T, Varki A. 2012. Metabolism of vertebrate amino sugars with N-glycolyl groups: elucidating the intracellular fate of the non-human sialic acid N-glycolylneuraminic acid. *J Biol Chem* 287:28865-28881.
- Bighignoli B, Niini T, Grahn RA, Pedersen NC, Millon LV, Polli M, Longeri M, Lyons LA. 2007. Cytidine monophospho-N-acetylneuraminic acid hydroxylase (CMAH) mutations associated with the domestic cat AB blood group. *BMC Genet* 8:27.
- Bosetti C, Micelotta S, Dal Maso L, Talamini R, Montella M, Negri E, Conti E, Franceschi S, La Vecchia C. 2004. Food groups and risk of prostate cancer in Italy. *Int J Cancer* 110:424-428.
- Bouhours D, Bouhours JF. 1988. Tissue-specific expression of GM3(NeuGc) and GD3(NeuGc) in epithelial cells of the small intestine of strains of inbred rats. Absence of NeuGc in intestine and presence in kidney gangliosides of brown Norway and spontaneously hypertensive rats. *J Biol Chem* 263:15540-15545.
- Bouhours D, Pourcel C, Bouhours JE. 1996. Simultaneous expression by porcine aorta endothelial cells of glycosphingolipids bearing the major epitope for human xenoreactive antibodies (Gal alpha 1-3Gal), blood group H determinant and N-glycolylneuraminic acid. *Glycoconj J* 13:947-953.
- Campanero-Rhodes MA, Smith A, Chai W, Sonnino S, Mauri L, Childs RA, Zhang Y, Ewers H, Helenius A, Imberty A, et al. 2007. N-glycolyl GM1 ganglioside as a receptor for simian virus 40. *J Virol* 81:12846-12858.
- Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genomics. Genome project standards in a new era of sequencing. *Science* 326:236-237.
- Chandrasekharan K, Yoon JH, Xu Y, deVries S, Camboni M, Janssen PM, Varki A, Martin PT. 2010. A human-specific deletion in mouse Cmah increases disease severity in the mdx model of Duchenne muscular dystrophy. *Sci Transl Med* 2:42ra54.
- Chen Y, Pan L, Liu N, Troy FA, Wang B. 2014. LC-MS/MS quantification of N-acetylneuraminic acid, N-glycolylneuraminic acid and ketodeoxyonulonic acid levels in the urine and potential relationship with dietary sialic acid intake and disease in 3- to 5-year-old children. *Br J Nutr* 111:332-341.
- Chothe SK, Bhushan G, Nissly RH, Yeh YT, Brown J, Turner G, Fisher J, Sewall BJ, Reeder DM, Terrones M, et al. 2017. Avian and human influenza virus compatible sialic acid receptors in little brown bats. *Sci Rep* 7:660.
- Chou HH, Hayakawa T, Diaz S, Krings M, Indriati E, Leakey M, Paabo S, Satta Y, Takahata N, Varki A. 2002. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc Natl Acad Sci U S A* 99:11736-11741.
- Chou HH, Takematsu H, Diaz S, Iber J, Nickerson E, Wright KL, Muchmore EA, Nelson DL, Warren ST, Varki A. 1998. A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc Natl Acad Sci U S A* 95:11751-11756.
- Coordinators NR. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44:D7-19.

- Corfield AP, Schauer R. 1982. Occurrence of sialic acids. In: Sialic acids. Vienna: Springer. p. 5-50.
- Davies LR, Pearce OM, Tessier MB, Assar S, Smutova V, Pajunen M, Sumida M, Sato C, Kitajima K, Finne J, et al. 2012. Metabolism of vertebrate amino sugars with N-glycolyl groups: resistance of α 2-8-linked N-glycolylneuraminic acid to enzymatic cleavage. *J Biol Chem* 287:28917-28931.
- Devos D, Valencia A. 2001. Intrinsic errors in genome annotation. *Trends Genet* 17:429-431.
- Diaz SL, Padler-Karavani V, Ghaderi D, Hurtado-Ziola N, Yu H, Chen X, Brinkman-Van der Linden EC, Varki A, Varki NM. 2009. Sensitive and specific detection of the non-human sialic Acid N-glycolylneuraminic acid in human tissues and biotherapeutic products. *PLoS One* 4:e4241.
- Diswall M, Angström J, Karlsson H, Phelps CJ, Ayares D, Teneberg S, Breimer ME. 2010. Structural characterization of alpha1,3-galactosyltransferase knockout pig heart and kidney glycolipids and their reactivity with human and baboon antibodies. *Xenotransplantation* 17:48-60.
- Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 15:330-340.
- Fraser GE. 1999. Associations between diet and cancer, ischemic heart disease, and all-cause mortality in non-Hispanic white California Seventh-day Adventists. *Am J Clin Nutr* 70:532S-538S.
- Fujii Y, Higashi H, Ikuta K, Kato S, Naiki M. 1982. Specificities of human heterophilic Hanganutziu and Deicher (H-D) antibodies and avian antisera against H-D antigen-active glycosphingolipids. *Mol Immunol* 19:87-94.
- Ghaderi D, Springer SA, Ma F, Cohen M, Secret P, Taylor RE, Varki A, Gagneux P. 2011. Sexual selection by female immunity against paternal antigens can fix loss of function alleles. *Proc Natl Acad Sci U S A* 108:17743-17748.
- Giovannucci E, Rimm EB, Colditz GA, Stampfer MJ, Ascherio A, Chute CG, Chute CC, Willett WC. 1993. A prospective study of dietary fat and risk of prostate cancer. *J Natl Cancer Inst* 85:1571-1579.
- Gohin M, Bobe J, Chesnel F. 2010. Comparative transcriptomic analysis of follicle-enclosed oocyte maturational and developmental competence acquisition in two non-mammalian vertebrates. *BMC Genomics* 11:18.
- Gollub M, Shaw L. 2003. Isolation and characterization of cytidine-5'-monophosphate-N-acetylneuraminase hydroxylase from the starfish *Asterias rubens*. *Comp Biochem Physiol B Biochem Mol Biol* 134:89-101.
- Guérardel Y, Chang LY, Fujita A, Coddeville B, Maes E, Sato C, Harduin-Lepers A, Kubokawa K, Kitajima K. 2012. Sialome analysis of the cephalochordate *Branchiostoma belcheri*, a key organism for vertebrate evolution. *Glycobiology* 22:479-491.
- Guo, W. P., et al. 2013. Phylogeny and origins of hantaviruses harbored by bats, insectivores, and rodents. *PLoS Pathog* 9(2):e1003159.
- Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.
- Hashimoto Y, Yamakawa T, Tanabe Y. 1984. Further studies on the red cell glycolipids of various breeds of dogs. A possible assumption about the origin of Japanese dogs. *J Biochem* 96:1777-1782.
- Hayakawa T, Aki I, Varki A, Satta Y, Takahata N. 2006. Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. *Genetics* 172:1139-1146.
- Hayakawa T, Satta Y, Gagneux P, Varki A, Takahata N. 2001. Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci U S A* 98:11399-11404.
- Hedlund M, Tangvoranuntakul P, Takematsu H, Long JM, Housley GD, Kozutsumi Y, Suzuki A, Wynshaw-Boris A, Ryan AF, Gallo RL, et al. 2007. N-glycolylneuraminic acid deficiency in mice: implications for human biology and evolution. *Mol Cell Biol* 27:4340-4346.
- Hentrich K, Löfling J, Pathak A, Nizet V, Varki A, Henriques-Normark B. 2016. *Streptococcus pneumoniae* Senses a Human-like Sialic Acid Profile via the Response Regulator CiaR. *Cell Host Microbe* 20:307-317.

- Hurh S, Kang B, Choi I, Cho B, Lee EM, Kim H, Kim YJ, Chung YS, Jeong JC, Hwang JI, et al. 2016. Human antibody reactivity against xenogeneic N-glycolylneuraminic acid and galactose- α -1,3-galactose antigen. *Xenotransplantation* 23:279-292.
- Ikeda K, Yamamoto A, Nanjo A, Inuinaka C, Takama Y, Ueno T, Fukuzawa M, Nakano K, Matsunari H, Nagashima H, et al. 2012. A cloning of cytidine monophospho-N-acetylneuraminic acid hydroxylase from porcine endothelial cells. *Transplant Proc* 44:1136-1138.
- Irie A, Suzuki A. 1998. The molecular basis for the absence of N-glycolylneuraminic acid in humans. *Tanpakushitsu Kakusan Koso* 43:2404-2409.
- Ito T, Suzuki Y, Suzuki T, Takada A, Horimoto T, Wells K, Kida H, Otsuki K, Kiso M, Ishida H, et al. 2000. Recognition of N-glycolylneuraminic acid linked to galactose by the α 2,3 linkage is associated with intestinal replication of influenza A virus in ducks. *J Virol* 74:9300-9305.
- Jhon GJ, Park SY, Han SY, Lee S, Kim Y, Chang YS. 1999. Studies of the chemical structure of gangliosides in deer antler, *Cervus nippon*. *Chem Pharm Bull (Tokyo)* 47:123-127.
- Kamerling JP, Dorland L, van Halbeek H, Vliegenthart JF, Messer M, Schauer R. 1982. Structural studies of 4-O-acetyl- α -N-acetylneuraminyl-(2 goes to 3)-lactose, the main oligosaccharide in echidna milk. *Carbohydr Res* 100:331-340.
- Kavaler S, Morinaga H, Jih A, Fan W, Hedlund M, Varki A, Kim JJ. 2011. Pancreatic beta-cell failure in obese mice with human-like CMP-Neu5Ac hydroxylase deficiency. *FASEB J* 25:1887-1893.
- Kawano T, Koyama S, Takematsu H, Kozutsumi Y, Kawasaki H, Kawashima S, Kawasaki T, Suzuki A. 1995. Molecular cloning of cytidine monophospho-N-acetylneuraminic acid hydroxylase. Regulation of species- and tissue-specific expression of N-glycolylneuraminic acid. *J Biol Chem* 270:16458-16463.
- Klein A, Diaz S, Ferreira I, Lamblin G, Roussel P, Manzi AE. 1997. New sialic acids from biological sources identified by a comprehensive and sensitive approach: liquid chromatography-electrospray ionization-mass spectrometry (LC-ESI-MS) of SIA quinoxalinones. *Glycobiology* 7:421-432.
- Koizumi N, Hara A, Uemura K, Taketomi T. 1988. Glycosphingolipids in sheep liver, kidney, and various blood cells. *Jpn J Exp Med* 58:21-31.
- Komoda H, Miyagawa S, Kubo T, Kitano E, Kitamura H, Omori T, Ito T, Matsuda H, Shirakura R. 2004. A study of the xenoantigenicity of adult pig islets cells. *Xenotransplantation* 11:237-246.
- Kyogashima M, Ginsburg V, Krivan HC. 1989. Escherichia coli K99 binds to N-glycolylsialoparagloboside and N-glycolyl-GM3 found in piglet small intestine. *Arch Biochem Biophys* 270:391-397.
- Lei M, Dong D. 2016. Phylogenomic analyses of bat subordinal relationships based on transcriptome data. *Sci Rep* 6:27726.
- Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127-128.
- Linseisen J, Kesse E, Slimani N, Bueno-De-Mesquita HB, Ocké MC, Skeie G, Kumle M, Dorransoro Iraeta M, Morote Gómez P, Janzon L, et al. 2002. Meat consumption in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohorts: results from 24-hour dietary recalls. *Public Health Nutr* 5:1243-1258.
- Löffling J, Lyi SM, Parrish CR, Varki A. 2013. Canine and feline parvoviruses preferentially recognize the non-human cell surface sialic acid N-glycolylneuraminic acid. *Virology* 440:89-96.
- Malykh YN, Schauer R, Shaw L. 2001. N-Glycolylneuraminic acid in human tumours. *Biochimie* 83:623-634.
- Malykh YN, Shaw L, Schauer R. 1998. The role of CMP-N-acetylneuraminic acid hydroxylase in determining the level of N-glycolylneuraminic acid in porcine tissues. *Glycoconj J* 15:885-893.
- Martensen I, Schauer R, Shaw L. 2001. Cloning and expression of a membrane-bound CMP-N-acetylneuraminic acid hydroxylase from the starfish *Asterias rubens*. *Eur J Biochem* 268:5157-5166.

- Martin MJ, Rayner JC, Gagneux P, Barnwell JW, Varki A. 2005. Evolution of human-chimpanzee differences in malaria susceptibility: relationship to human genetic loss of N-glycolylneuraminic acid. *Proc Natl Acad Sci U S A* 102:12819-12824.
- Muchmore EA, Diaz S, Varki A. 1998. A structural difference between the cell surfaces of humans and the great apes. *Am J Phys Anthropol* 107:187-198.
- Muralikrishna G, Reuter G, Peter-Katalinić J, Egge H, Hanisch FG, Siebert HC, Schauer R. 1992. Identification of a new ganglioside from the starfish *Asterias rubens*. *Carbohydr Res* 236:321-326.
- Naito-Matsui Y, Davies LR, Takematsu H, Chou HH, Tangvoranuntakul P, Carlin AF, Verhagen A, Heyser CJ, Yoo SW, Choudhury B, et al. 2017. Physiological Exploration of the Long-term Evolutionary Selection Against Expression of N-glycolylneuraminic Acid in the Brain. *J Biol Chem*.
- Ng PS, Böhm R, Hartley-Tassell LE, Steen JA, Wang H, Lukowski SW, Hawthorne PL, Trezise AE, Coloe PJ, Grimmond SM, et al. 2014. Ferrets exclusively synthesize Neu5Ac and express naturally humanized influenza A virus receptors. *Nat Commun* 5:5750.
- Nguyen, L. T., et al. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268-74.
- Okerblom J, Varki A. 2017. Biochemical, Cellular, Physiological and Pathological Consequences of Human loss of N-glycolylneuraminic Acid. *Chembiochem*.
- Omi T, Nakazawa S, Udagawa C, Tada N, Ochiai K, Chong YH, Kato Y, Mitsui H, Gin A, Oda H, et al. 2016. Molecular Characterization of the Cytidine Monophosphate-N-Acetylneuraminic Acid Hydroxylase (CMAH) Gene Associated with the Feline AB Blood Group System. *PLoS One* 11:e0165000.
- Pape L, Kristensen BI, Bengtson O. 1975. Sialic acid, electrophoretic mobility and transmembrane potentials of the *Amphiuma* red cell. *Biochim Biophys Acta* 406:516-525.
- Rich SM, Leendertz FH, Xu G, LeBreton M, Djoko CF, Aminake MN, Takang EE, Dikko JL, Pike BL, Rosenthal BM, et al. 2009. The origin of malignant malaria. *Proc Natl Acad Sci U S A* 106:14902-14907.
- Rizzo AM, Berra B, Rossi F, Guerra A, Gornati R, Bernardini G, Taki T, Kasama T, Mauri L, Sonnino S. 2002. Structure of the main ganglioside from the brain of *Xenopus laevis*. *Glycoconj J* 19:53-57.
- Rogers GN, Paulson JC. 1983. Receptor determinants of human and animal influenza virus isolates: differences in receptor specificity of the H3 hemagglutinin based on species of origin. *Virology* 127:361-373.
- Rose DP, Boyar AP, Wynder EL. 1986. International comparisons of mortality rates for cancer of the breast, ovary, prostate, and colon, and per capita food consumption. *Cancer* 58:2363-2371.
- Salama A, Evanno G, Harb J, Soulillou JP. 2015. Potential deleterious role of anti-Neu5Gc antibodies in xenotransplantation. *Xenotransplantation* 22:85-94.
- Samraj AN, Pearce OM, Läubli H, Crittenden AN, Bergfeld AK, Banda K, Gregg CJ, Bingman AE, Secrest P, Diaz SL, et al. 2015. A red meat-derived glycan promotes inflammation and cancer progression. *Proc Natl Acad Sci U S A* 112:542-547.
- Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvermin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37:D5-15.
- Schauer R. 1970. Biosynthesis of N-glycolylneuraminic acid by an ascorbic acid- or NADP-dependent N-acetyl hydroxylating "N-acetylneuramate: O²-oxidoreductase" in homogenates of porcine submaxillary gland. *Hoppe Seylers Z Physiol Chem* 351:783-791.
- Schauer R. 2004. Sialic acids: fascinating sugars in higher animals and man. *Zoology (Jena)* 107:49-64.
- Schauer R, Haverkamp J, Ehrlich K. 1980. Isolation and characterization of acylneuramate cytidyltransferase from frog liver. *Hoppe Seylers Z Physiol Chem* 361:641-648.
- Schauer R, Kamerling JP. 1997. Chemistry, biochemistry and biology of sialic acids. *New Comprehensive Biochemistry* 29:243-402.

- Schauer R, Reuter G, Mühlpfordt H, Andrade AF, Pereira ME. 1983. The occurrence of N-acetyl- and N-glycolylneuraminic acid in *Trypanosoma cruzi*. *Hoppe Seylers Z Physiol Chem* 364:1053-1057.
- Schauer R, Schoop HJ, Faillard H. 1968. On biosynthesis of the glycolyl groups of N-glycolylneuraminic acid. Oxidative conversion of N-acetyl groups to glycolyl groups. *Hoppe Seylers Z Physiol Chem* 349:645-652.
- Schauer R, Srinivasan GV, Coddeville B, Zanetta JP, Guérardel Y. 2009. Low incidence of N-glycolylneuraminic acid in birds and reptiles and its absence in the platypus. *Carbohydr Res* 344:1494-1500.
- Schoop HJ, Schauer R, Faillard H. 1969. On the biosynthesis of N-glycolylneuraminic acid. Oxidative formation of N-glycolylneuraminic acid from N-acetylneuraminic acid. *Hoppe Seylers Z Physiol Chem* 350:155-162.
- Schwegmann-Wessels C, Herrler G. 2006. Sialic acids as receptor determinants for coronaviruses. *Glycoconj J* 23:51-58.
- Senda A, Kobayashi R, Fukuda K, Saito T, Hood WR, Kunz TH, Oftedal OT, Urashima T. 2011. Chemical characterization of milk oligosaccharides of the island flying fox (*Pteropus hypomelanus*) (Chiroptera: Pteropodidae). *Anim Sci J* 82:782-786.
- Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, Hisata K, Bredeson J, Shoguchi E, Gyoja F, et al. 2015. Hemichordate genomes and deuterostome origins. *Nature* 527:459-465.
- Springer SA, Diaz SL, Gagneux P. 2014. Parallel evolution of a self-signal: humans and new world monkeys independently lost the cell surface sugar Neu5Gc. *Immunogenetics* 66:671-674.
- Staudacher E, BuÈrgmayr, S., Grabher-Meier, H. & Halama, T. 1999. N-glycans of *Arion lusitanicus* and *Arion rufus* contain sialic acid residues. *Glycoconj. J* 16.
- Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7.
- Sumi T, Sallay I, Asakawa M, Park SS, Miyazaki M, Ohba H. 2001. Purification and identification of N-glycolylneuraminic acid (Neu5Gc) from the holothuroidea *Gumi*, *Cucumaria echinata*. *Prep Biochem Biotechnol* 31:135-146.
- Tangvoranuntakul P, Gagneux P, Diaz S, Bardor M, Varki N, Varki A, Muchmore E. 2003. Human uptake and incorporation of an immunogenic nonhuman dietary sialic acid. *Proc Natl Acad Sci U S A* 100:12045-12050.
- Tavani A, La Vecchia C, Gallus S, Lagiou P, Trichopoulos D, Levi F, Negri E. 2000. Red meat intake and cancer risk: a study in Italy. *Int J Cancer* 86:425-428.
- Teeling EC, Springer MS, Madsen O, Bates P, O'brien SJ, Murphy WJ. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science* 307:580-584.
- Terabayashi T, Ogawa T, Kawanishi Y. 1992. A comparative study on ceramide composition of cetacean brain gangliosides. *Comp Biochem Physiol B* 103:721-726.
- Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. *Curr Biol* 23:2262-2267.
- Tseng M, Wright DJ, Fang CY. 2015. Acculturation and dietary change among Chinese immigrant women in the United States. *J Immigr Minor Health* 17:400-407.
- Tu Q, Cameron RA, Worley KC, Gibbs RA, Davidson EH. 2012. Gene structure in the sea urchin *Strongylocentrotus purpuratus* based on transcriptome analysis. *Genome Res* 22:2079-2087.
- van Valkenburgh B, Pang B, Bird D, Curtis A, Yee K, Wysocki C, Craven BA. 2014. Respiratory and olfactory turbinals in feliform and caniform carnivores: the influence of snout length. *Anat Rec (Hoboken)* 297:2065-2079.
- Varki, A. 2011. Since there are PAMPs and DAMPs, there must be SAMPs? Glycan “self-associated molecular patterns” dampen innate immunity, but pathogens can mimic them. *Glycobiology* 21 (9):1121-4.
- Varki A. 2009. Multiple changes in sialic acid biology during human evolution. *Glycoconj J* 26:231-245.
- Varki A, Gagneux P. 2009. Human-specific evolution of sialic acid targets: explaining the malignant malaria mystery? *Proc Natl Acad Sci U S A* 106:14739-14740.

- Varki NM, Varki A. 2007. Diversity in cell surface sialic acid presentations: implications for biology and disease. *Lab Invest* 87:851-857.
- Warren L. 1963. The distribution of sialic acids in nature. *Comp Biochem Physiol* 10:153-171.
- Willett WC. 2000. Diet and cancer. *Oncologist* 5:393-404.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-1591.
- Yasue S, Handa S, Miyagawa S, Inoue J, Hasegawa A, Yamakawa T. 1978. Difference in form of sialic acid in red blood cell glycolipids of different breeds of dogs. *J Biochem* 83:1101-1107.
- Yeşilyurt B, Şahar U, Deveci R. 2015. Determination of the type and quantity of sialic acid in the egg jelly coat of the sea urchin *Paracentrotus lividus* using capillary LC-ESI-MS/MS. *Mol Reprod Dev* 82:115-122.
- Zhang J, Kesteloot H. 2005. Milk consumption in relation to incidence of prostate, breast, colon, and rectal cancers: is there an independent effect? *Nutr Cancer* 53:65-72.

TABLES

Table 1. Number of deuterostome genomes covered in this study

Group	Number of genomes available	Number of genomes with putatively functional <i>CMAH</i>	Number of genomes with <i>CMAH</i> absent or pseudogene
Echinoderms	10*	10*	0
Cephalochordates	3	3	0
Hemichordates	2	2	0
Urochordates	5	0	5
Fish	78	55	23
Amphibia	4	3	1
Reptiles	19	1	18
Birds	73	0	73
Mammals	129	110	19
Total	323	184	139

*Includes two unsequenced genomes with available *CMAH* mRNA sequences.

Table 2. CMAH pseudogenes identified in the current study

Species	Common name	Status	Type	Position*
<i>Homo sapiens</i>	Human	Known	Deletion of coding exon 3	74-103
<i>Aotus nancymaae</i>	Ma's night monkey [New World monkey]	Known	Deletion of coding exon 3-15	74-605
<i>Cebus capucinus</i>	White-faced sapajou [New World monkey]	Known	Deletion of coding exon 3-15	74-605
<i>Saimiri boliviensis boliviensis</i>	Bolivian squirrel monkey [New World monkey]	Known	Deletion of coding exon 3-15	74-564
<i>Callithrix jacchus</i>	White-tufted-ear marmoset [New World monkey]	Known	Deletion of coding exon 3-15	74-564
<i>Ornithorhynchus anatinus</i>	Platypus	Known	PSC in coding exon 5	154
<i>Mustela putorius furo</i>	Domestic ferret	Known	PSC in coding exon 11	444
<i>Physeter catodon</i>	Sperm whale	Unknown	Deletion of coding exon 5	143-205
<i>Erinaceus europaeus</i>	Western European hedgehog	Unknown	PSC in coding exon 13	530, 556
<i>Gekko japonicus</i>	Japanese gecko	Unknown	PSC in coding exon 4	141
<i>Oreochromis niloticus</i> (copy 2)	Nile tilapia	Unknown	FSM in coding exon 4 & 5	122, 152
<i>Oreochromis niloticus</i> (copy 2)	Nile tilapia	Unknown	PSC in coding exon 5	177
<i>Haplochromis burtoni</i> (copy 2)	Burton's mouthbrooder	Unknown	PSC in coding exon 8	305
<i>Branchiostoma belcheri</i> (copy 2)	Chinese amphioxus	Unknown	PSC in coding exon 6	213
<i>Branchiostoma floridae</i> (copy 3)	Floridan amphioxus	Unknown	PSC in coding exon 13	539, 545
<i>Acanthaster planci</i> (copy 2)	Crown-of-thorns starfish	Unknown	PSC in coding exon 10	425

PSC, premature stop codon; FSM, frameshift mutation. *Position relative to chimpanzee protein.

FIGURE LEGENDS

Fig. 1. Presence/absence of the gene *CMAH* in non-vertebrates. Species in which gene is absent or inactivated are marked in grey. “X” symbols represent inferred gene loss events. *, one of the duplicates is a pseudogene.

Fig. 2. Presence/absence of the gene *CMAH* in fish. Species in which gene is absent or inactivated are marked in grey. “X” symbols represent inferred gene loss events. *, one of the duplicates is a pseudogene.

Fig. 3. Presence/absence of the gene *CMAH* in amphibians, birds and reptiles. Species in which gene is absent or inactivated are marked in grey. “X” symbols represent inferred gene loss events.

Fig. 4. Presence/absence of the gene *CMAH* in mammals. Species in which gene is absent or inactivated are marked in grey. “X” symbols represent inferred gene loss events.

Fig. 1.

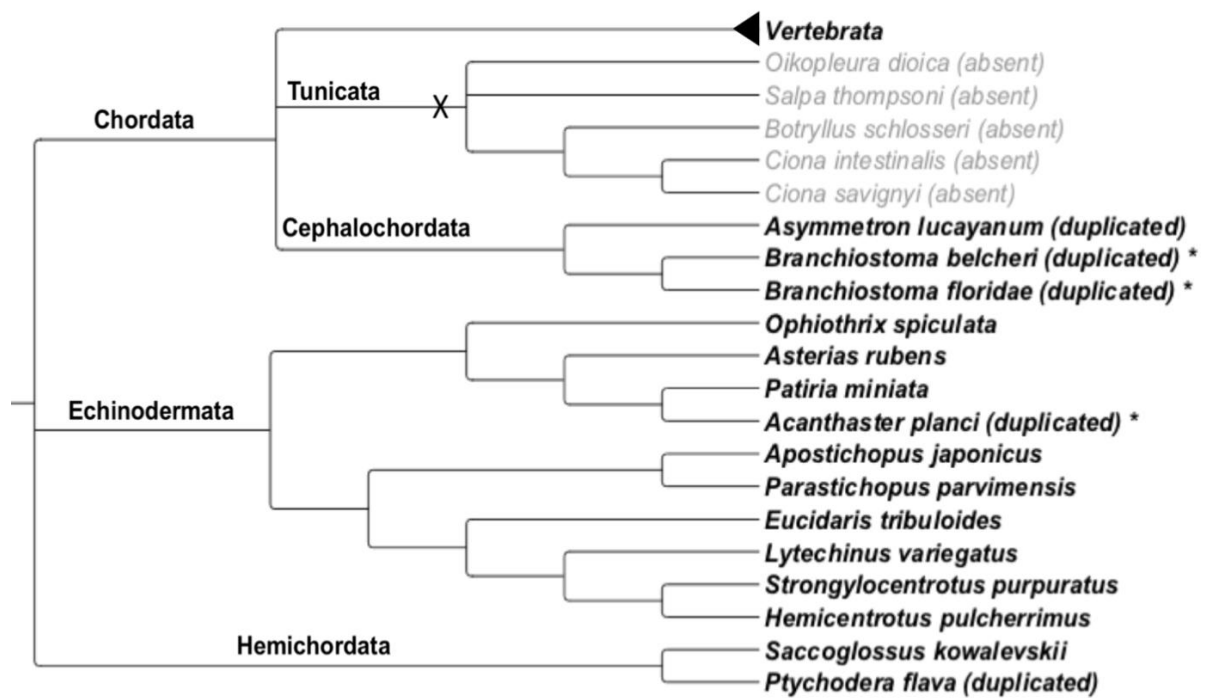


Fig. 2.

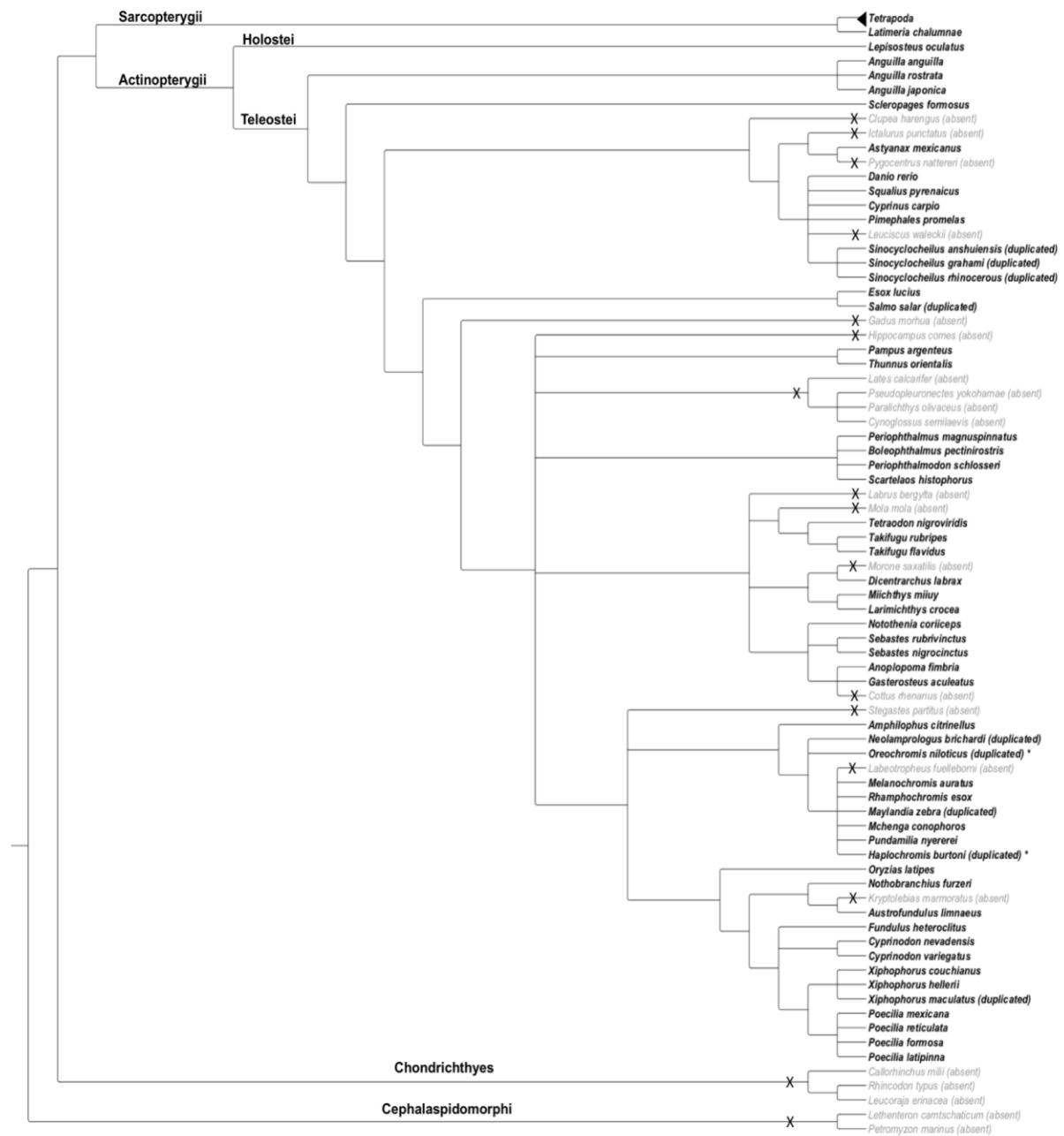


Fig. 3.

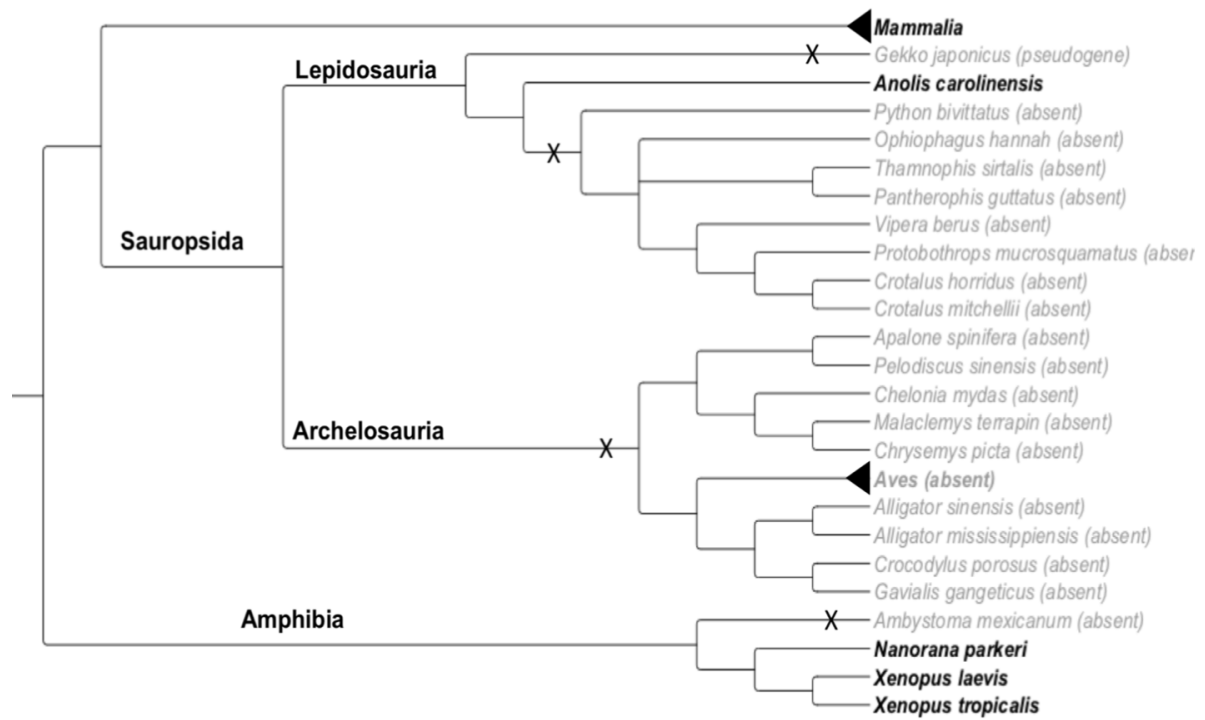


Fig. 4.

