Phylogenetic Distribution of CMP-Neu5Ac Hydroxylase (CMAH), the Enzyme Synthetizing the Proinflammatory Human Xenoantigen Neu5Gc

Sateesh Peri, Asmita Kulkarni, Felix Feyertag, Patricia M. Berninsone, and David Alvarez-Ponce*
Department of Biology, University of Nevada, Reno

*Corresponding author: E-mail: dap@unr.edu.
Accepted: November 29, 2017

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 nondeuterostomes, and 322 deuterostome genomes, to ascertain the phylogenetic distribution of CMAH. Among nondeuterostomes, 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.

Key words: 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, echinoderms, and echinoderms), being uncommon in other organisms (Warren 1963; Corfield and Schauer 1982; Staudacher et al. 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 synthetizing 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 nondeuterostomes, 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...

CMAH gens preferentially recognizing Neu5Ac, such as Varki 2007; Varki 2009). Fourth, the phenotypes of Campanero-Rhodes et al. 2007). Nonetheless, inactiva-

tion of Plasmodium reichenowi (Schwegmann-Wessels and Herrler 2006), and simian virus 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 <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 2009; Varki and 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; Kavaler 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 (Bouhous and Bouhous 1988; Bouhous et al. 1996; Diswall et al. 2010), and in adult pig pancreatic islets (Komoda et al. 2004), raising concerns over their use 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 Ma 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 nine 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 (Fuji 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.
euchidna (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 experimented 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 Data Set and Determination of CMAH Presence/Absence

We used all genomic data available from the NCBI Genome database to determine the occurrence of CMAH in nondeuterostomes. We then focused on the 321 deuterostome genomes and 2 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 nonmodel 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 eight 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 pseudogene 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 polytomy 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 nonvertebrate
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

---

**Fig. 2.**—Presence/absence of the gene CMAH in fish. Species in which gene is absent or inactivated are marked in gray. “X” symbols represent inferred gene loss events. *One of the duplicates is a pseudogene.*
Results and Discussion

CMAH in Nondeuterostomes

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, FCGB group (Fibrobacteres, Chlorobii, and Bacteroidetes), and archaea bacteria (supplementary table S1, Supplementary Material online). No hits were detected in any other group, including nondeuterostome animals. The distribution of CMAH homologs in nondeuterostomes 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 Nonvertebrate 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 chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates (table 1). These represent almost completely sequenced chordates, 5 urochordates (tunicates), and 304 vertebrates.
FIG. 4.—Presence/absence of the gene CMAH in mammals. Species in which gene is absent or inactivated are marked in gray. “X” symbols represent inferred gene loss events.
an ancestor of urochordates. These results are in agreement with prior observations that tunicates are devoid of sialic acids (Warren 1963).

**CMAH in Fish**

Several groups have independently reported nonsignificant 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 Cephalaspidomorphi and Chondrichthyes (fig. 2).

**CMAH in Amphibians**

There are four 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

### Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Genomes Available</th>
<th>Number of Genomes with Putatively Functional CMAH</th>
<th>Number of Genomes with CMAH Absent or Pseudogene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Echinoderms</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Cephalochordates</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Hemichordates</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Urochordates</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Fish</td>
<td>78</td>
<td>55</td>
<td>23</td>
</tr>
<tr>
<td>Amphibia</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reptiles</td>
<td>19</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Birds</td>
<td>73</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>Mammals</td>
<td>129</td>
<td>110</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>323</td>
<td>184</td>
<td>139</td>
</tr>
</tbody>
</table>

*Includes two unsequenced genomes with available CMAH mRNA sequences.

### Table 2

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

PSC, premature stop codon; FSM, frameshift mutation.

*Position relative to chimpanzee protein.
salamander \textit{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 (fig. 3). Supplementary data set S2, Supplementary Material online, is a multiple sequence alignment including these sequences.

Previous studies did not detect Neu5Gc in the liver of the frog \textit{Rana esculata} (Schauer et al. 1980) or in the brain gangliosides of \textit{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 \textit{Amphiuma means} also have been reported to contain only Neu5Ac (Pape et al. 1975). However, Neu5Gc has been reported in the oviducal mucins of fire-bellied toads (\textit{Bombina bombina} and \textit{Bombina variegata}) and the alpine newt (\textit{Triturus alpestris}) (Schauer et al. 2009). In addition, CMAH has been identified as one of the key genes downregulated during oocyte maturation in \textit{X. laevis} (Gohin et al. 2010).

The CMAH genes of \textit{X. tropicalis} and \textit{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 nonsynonymous divergence ratio (\(d_{N}/d_{S}\)). This value was significantly <1 (\(d_{N}/d_{S} = 0.142, 2\alpha = 109.03, P = 1.60 \times 10^{-25}\)), indicating strong purifying selection, and thus strongly suggesting functionality of the CMAH enzyme in \textit{Xenopus}. Therefore, we suspect that the genes have been erroneously annotated as pseudogenes by the automatic gene annotation pipelines.

Our results suggest that \textit{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 nonsignificant levels of Neu5Gc in most of their tissues despite having an intact CMAH coding sequence.

\section*{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. (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 nonbird reptiles (green iguana, agama, green basilisk hatchling, anaconda, hundred pace viper, Taiwan stings snake, Taiwan beauty snake, crocodile, and Ambonya 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 nonbird reptiles for CMAH homologs (table 1). Surprisingly, we found homologs in the two lizards included in our data set, \textit{Anolis carolinensis} (green anole lizard), and \textit{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, crocodilians, and birds (fig. 3). The \textit{A. carolinensis} CMAH sequence appears to be functional. The \textit{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 \textit{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.

\section*{CMAH in Mammals}

The NCBI Genome database contains the genomes of 1 monotreme, 3 marsupials, and 126 placental mammals (table 1). We found putatively functional CMAH homologs in the genomes of all 3 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 fig. 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 (\textit{ Physeter catodon}), two events in bats, one in an ancestor of the white-tailed deer (\textit{Odocoileus virginianus}), and another in an ancestor of the European hedgehog (\textit{Erinaceus europaeus}) (fig. 4).

Supplementary data set S3, Supplementary Material online, is a multiple sequence alignment of all mammalian vertebrates.

Neu5Gc is known to be present in most placental mammals studied so far (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 (\textit{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, ∼2.5–3 Ma (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 3–15 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 nancymaeae (Nancy Ma’s night monkey).

Ng et al. (2014) investigated the ferret genome and performed PCR analyses on another 14 species of musteloids and 2 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). Although 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 Vesper（tillonidae (Yangchiroptera). Thus, the CMAH gene has undergone independent gene losses in both Yinchiroptera and Yangchiroptera (fig. 4).

The NCBI Genome database contains the genomes of 14 members of the taxon Ruminantia (ten bovids, two giraffids, and two 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 six 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 ten 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 (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 data set included three genomes of the group Insectivora. Two of these genomes, Sorex araneus (common shrew) and Condyalla 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 data set S3, Supplementary Material online). 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. carolinesis 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 nondeuterostomes, 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 (figs. 1–4). Our inferred gene trees (supplementary figs. S1–S3, Supplementary Material online) do not contradict the species tree (figs. 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. carolinesis, 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, crocodilians, and birds (fig. 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 un sequenced fraction (typically, one coding exon is un sequenced; supplementary table S2 and data sets S1–S3, Supplementary Material online). It is possible that some of these un sequenced 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.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

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

**Literature Cited**


Varki A. 2011. Since there are PAMPs and DAMPs, there must be SAMPs? Glycan “self-associated molecular patterns” dappen innate immunity, but pathogens can mimic them. Glycobiology 21(9):1121–1124.


Associate editor: Bill Martin