Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales

Robert W. Meredith, John Gatesy, Joyce Cheng and Mark S. Springer*

Department of Biology, University of California, Riverside, CA 92521, USA

Whales in the suborder Mysticeti are filter feeders that use baleen to sift zooplankton and small fish from ocean waters. Adult mysticetes lack teeth, although tooth buds are present in foetal stages. Cladistic analyses suggest that functional teeth were lost in the common ancestor of crown-group Mysticeti. DNA sequences for the tooth-specific genes, ameloblastin (AMBN), enamelin (ENAM) and amelogenin (AMEL), have frameshift mutations and/or stop codons in this taxon, but none of these molecular cavities are shared by all extant mysticetes. Here, we provide the first evidence for pseudogenization of a tooth gene, enamelysin (MMP20), in the common ancestor of living baleen whales. Specifically, pseudogenization resulted from the insertion of a CHR-2 SINE retroposon in exon 2 of MMP20. Genomic and palaeontological data now provide congruent support for the loss of enamel-capped teeth on the common ancestral branch of crown-group mysticetes. The new data for MMP20 also document a polymorphic stop codon in exon 2 of the pygmy sperm whale (Kogia breviceps), which has enamel-less teeth. These results, in conjunction with the evidence for pseudogenization of MMP20 in Hoffmann’s two-toed sloth (Choloepus hoffmanni), another enamel-less species, support the hypothesis that the only unique, non-overlapping function of the MMP20 gene is in enamel formation.

Keywords: baleen whale; enamel; enamelysin; macroevolution; pseudogene; teeth

1. INTRODUCTION

The evolution of Cetacea from terrestrial ancestors is one of the best-documented macroevolutionary transitions in the fossil record [1]. Stem cetaceans (pakicetids, ambulocetids, remingtonocetids, protocetids and basilosaurids) document progressive limb reduction, posterior migration of the external nares and other specializations for a fully aquatic lifestyle. Archaeocetes retained teeth, as do living cetaceans in the suborder Odontoceti. Cetaceans in the suborder Mysticeti, by contrast, have lost their adult teeth and instead use racks of baleen to filter zooplankton and small fish from ocean waters. Baleen is a key innovation that facilitated the exploitation of an unexploited ecological niche, bulk filter-feeding and laid the foundation for the evolution of the largest animals on the Earth [2]. In addition to including the largest extant mammal (blue whale), Mysticeti also includes the putatively longest living extant mammal (bowhead whale) [3–5]. The fossil record of stem mysticetes includes primitive forms that had teeth but not baleen (e.g. *Janjucetus*, *Mammalodon*), intermediate forms that had teeth, and by inference baleen, based on the presence of lateral nutrient foramina and sulci on the palate (e.g. *Aetiocetus*) and more derived forms with baleen but not teeth (e.g. *Eomysticetus*, *Micromysticetus*) [2,6–10]. Ontogenetic observations provide additional evidence for the occurrence of teeth in ancestral baleen whales; tooth buds develop in mysticete foetuses, but are subsequently aborted and resorbed prior to enamel formation [11–14]. The presence of teeth in foetal whales was even known to Darwin [15], who discussed the significance of this rudiment in his long argument for evolution.

Cladistic analyses of living and extinct mysticetes support the hypothesis that mineralized teeth were lost in the common ancestor of crown Mysticeti [2,8,9]. Molecular sequences of three enamel-specific genes, ameloblastin (AMBN), enamelin (ENAM) and amelogenin (AMEL) contain stop codons and/or frameshift mutations in various mysticete species ([2,16]; J. Gatesy 2010, unpublished data), but none of the inactivating mutations are common to all extant mysticetes even though a total of approximately 3650 bp have been sequenced from exonic regions of AMBN, ENAM and AMEL. Thus, the current body of molecular evidence agrees with the phylogenetic studies of fossils in documenting the pseudogenization of enamel-specific genes in living mysticetes, but falls short of supporting the hypothesis that the genetic toolkit for manufacturing enamel was knocked out on the same branch on which functional, mineralized teeth were apparently lost, i.e. the common ancestor of crown Mysticeti.

Several explanations can account for this inconsistency between fossil and molecular evidence. First, given that only partial protein-coding sequences were generated for AMBN, ENAM and AMEL, it is possible that frameshift mutations and/or stop codons will be discovered in the unsequenced protein-coding regions of one or more of these extracellular matrix protein (EMP) genes.

* Author for correspondence (mark.springer@ucr.edu).

Electronic supplementary material is available at [http://dx.doi.org/10.1098/rspb.2010.1280](http://dx.doi.org/10.1098/rspb.2010.1280) or via [http://rspb.royalsocietypublishing.org](http://rspb.royalsocietypublishing.org).
A second possibility is that one or more of these genes were initially silenced by mutations in a regulatory gene region on the ancestral mysticete branch, and that mutations in protein-coding regions accumulated subsequently on descendant branches within crown-group Mysticeti. A third possibility is that a different enamel- or tooth-specific gene was knocked out in the common ancestor of mysticetes, and that AMBN, ENAM and AMEL acquired molecular cavities on descendant branches within crown-group Mysticeti. Indeed, mysticetes have the slowest rates of nuclear gene evolution among mammals \[16–18\], which reduces the likelihood that all enamel-specific genes acquired their first inactivating mutation on the common mysticete branch. Finally, enamel may have been lost independently in several mysticete lineages, rather than once in the common ancestor of crown mysticetes. Edentulous stem mysticetes (e.g. *Eomysticetus*) and early crown mysticetes (‘cetotheres’) may have retained rudimentary, enamel-capped teeth that were embedded in soft tissue, rather than set in bony alveoli, as is the case for maxillary teeth in Ziphiidae (beaked whales) and Physeteroidea (sperm whales) \[19–21\]. In summary, the occurrence of teeth in Ziphiidae (beaked whales) and Physeteroidea (‘cetotheres’) may have retained rudimentary, enamel-capped teeth that were embedded in soft tissue, rather than set in bony alveoli, as is the case for maxillary teeth in Ziphiidae (beaked whales) and Physeteroidea (sperm whales) \[19–21\].

In an attempt to discriminate among competing hypotheses, we amplified and sequenced three of the longer exons (2, 3, 4) of the enamelysin gene of representative mysticetes, and searched for shared frameshift mutations and/or stop codons. Enamelysin belongs to the matrix metalloproteinase gene family and is otherwise known as matrix metalloproteinase 20 (MMP20). The MMP20 gene is located in a cluster of matrix metalloproteinase genes at human chromosome 11q22 [22,23]. MMP20 diverged from other matrix metalloproteinase loci prior to the common ancestry of tetrapods [24], and plays a key role in processing structural proteins (amelogenin, ameloblastin, enamelin) that are sequestered into the extracellular matrix by ameloblasts during the secretory stage of enamel formation [23,25–28]. MMP20 may also be necessary to activate kallikrein-related peptidase 4 (KLK4; [29,30]), which cleaves and degrades remnants of enamel matrix proteins during the maturation stage of amelogenesis. MMP20-deficient mice have an amelogenesis imperfecta phenotype that is characterized by thin, hypomineralized enamel that easily chips away from the underlying dentin [31,32]. There are also mutations in the human MMP20 gene that cause non-syndromic amelogenesis imperfecta [30]. Other evidence suggests a role for MMP20, along with matrix metalloproteinase 2 (MMP2), in cleaving dentin sialophosphoprotein (DSP), which consists of three parts: dentin sialoprotein (DSP), dentin glycoprotein (DGP) and dentin phosphoprotein (DPP) [33]. Specifically, MMP20 cleaves DSP–DGP to generate DSP and DGP, and also cleaves DSP at multiple sites to yield smaller DSP products [33]. However, the absence of conspicuous dentin phenotypes in humans and mice that lack a functional copy of MMP20 suggests that this enzyme and MMP2 are functionally redundant [33]. Given that the only unique, non-overlapping functions of MMP20 are enamel- or tooth-specific, we hypothesized that the MMP20 gene should show evidence of pseudogenization in crown mysticetes as it occurs for AMBN, ENAM and AMEL ([2,16]; J. Gatesy 2010, unpublished data).

2. MATERIAL AND METHODS

(a) Laboratory procedures

PCR amplifications for three different exons of MMP20 (2, 3, 4) were performed with primers from flanking introns to negate the possibility of amplifying processed pseudogenes. PCR primers were designed based on aligned sequences for *Bos taurus*, *Sus scrofa*, *Tursiops truncatus* and *Vicugna pacos* that were obtained from Ensembl 56. Exons 3 and 4 were each amplified with a single set of primers. Exon 2 was amplified with a nested set of primers after initial amplification with an outer set of primers. We used 1 μl of the first PCR reaction product as the template DNA in nested reactions. Primer sequences are provided in electronic supplementary material, table S1. Amplifications were performed with Denville Scientific Inc. Ramp-Taq DNA polymerase in 50 μl reactions with the following thermal cycling profile: pre-activation step at 95 °C for 7 min; initial denaturation at 95 °C for 2 min; 45 cycles of 1 min at 95 °C (denaturation), 1 min at 50 °C (annealing) and 2 min at 72 °C (extension); and a final extension at 72 °C for 10 min. In our initial screen, we attempted to amplify exons 2–4 from *Eubalaena australis* (southern right whale), *Caperea marginata* (pygmy right whale), *Esoxichthys robustus* (grey whale) and *Megaptera novaeangliae* (humpback whale), which are representative of the four extant mysticete families. After discovering a SINE insertion in exon 2 of MMP20 in each of these mysticetes, we performed additional amplifications with the mysticete taxa *Balaena mysticetus* (bowhead whale), *Balaenoptera acutorostrata* (common minke whale), *Balaenoptera physalus* (fin whale) and *Balaenoptera musculus* (blue whale). We also amplified exon 2 of MMP20 from representatives of most odontocete families, and additional cetartiodactyl outgroups, as follows: Monodontidae (*Monodon monoceros* (narwhal), *Delphinapterus leucas* (beluga)); Phocoenidae (*Phocoena phocoena* (harbour porpoise)); Iniidae (*Inia geoffrensis* (Amazon River dolphin)); Pontoporiidae (*Pontoporia blainvillei* (La Plata dolphin)); Platanistidae (*Platanista minor* (Indus River dolphin)); Physeteridae (*Physeter macrocephalus* (giant sperm whale)); Kogiidae (*Kogia sima* (dwarf sperm whale), *Kogia breviceps* (pygmy sperm whale)); Ziphiidae (*Mesoplodon bidens* (Sowerby’s beaked whale)); Hippopotomidae (*Hippopotamus amphibius* (hippopotamus)); Cervidae (*Cervus nippon* (sika deer)); Giraffidae (*Okapia johnstoni* (okapi)); Moschidae (*Moschus sp.* (musk deer)); Antilopacridae (*Antilocapra americana* (pronghorn)); Tayassuidae (*Peccari tajacu* (collared peccary)) and Camelidae (*Lama glama* (llama)).

Specimen numbers for genomic DNA samples are given in electronic supplementary material, table S2. Successfully amplified PCR products were electrophoresed on 1 per cent agarose gels, excised and cleaned with the Bioneer AccuPrep Gel Purification Kit. Cleaned PCR products were sequenced at the UCR Core Instrumentation Facility with an ABI 3730xl automated DNA sequencer. *Sequencher* 4.8 was used to assemble contigs. GenBank accession numbers for the new MMP20 sequences are HQ171778–HQ171814. Sequences for *B. taurus*, *S. scrofa* and *T. truncatus* were obtained from
Ensembl 56. The identification of the insertion in mysticete MMP20 exon 2 as a CHR-2 short interspersed nuclear element (SINE) retroposon resulted from BLASTing the Megaptera novaeangliae insertion against nucleotide sequences in GenBank. Accession numbers for additional CHR-2 SINEs that were employed in alignments and/or phylogenetic analyses are as follows: AB054403, AB054436, AB054471, AB054480, AB071537, AB071542, AB071567, AB071586, AB195475, AB195478, AB195479, AB195481–AB195486, AB195488, AB195492, AB195495.

(b) Alignments and phylogenetic analyses

Sequences were aligned manually with Se-Al [34]. The MMP20 exon 2 alignment (578 base pairs (bp)) consisted of complete exonic sequences, including the CHR-2 SINE retroposon in mysticetes, and the 5′ end of intron 2 for 28 species (electronic supplementary material, alignment 1). We also constructed a CHR-2 SINE alignment (263 bp) that consisted of sequences from the MMP20 locus, additional sequences from the CD (cetacean deletion) subfamily of CHR-2 SINEs and an outgroup sequence (Megaptera novaeangliae) from the Hump14 locus of the CHR-2 SINE DT (deletion type) subfamily (electronic supplementary material, alignment 2). Separate phylogenetic analyses were performed on: (i) the exonic region of the MMP20 exon 2 alignment, which comprised 248 bp after excluding frameshift insertions and intronic sequences, and (ii) the CHR-2 SINE alignment. The Akaike Information Criterion (AIC) of jModeltest [35] was used to select the model of molecular evolution that was implemented in maximum-likelihood (ML) analyses (MMP20 exon 2 alignment 2, CHR-2 SINE = HKY + F). ML searches were performed with PAUP* 4.0b10 [36] and employed stepwise addition with 100 randomized input orders and tree bisection and reconnection branch swapping. ML bootstrap analyses were performed with neighbour-joining starting trees and 500 pseudoreplicates.

(c) dNdS analyses

The codeml program of PAML 4.2 [37] was used to estimate the ratio (ω) of the non-synonymous substitution rate (dN) to the synonymous substitution rate (dS) for functional and pseudogenic branches of exon 2 of MMP20 after removing frameshift insertions and recoding the stop codon in B. acutorostrata as missing data. Given that none of the highest supported nodes (>70%) on the MMP20 tree conflicted with cetartiodactyl species trees (see below), and that differences pertain to nodes that are weakly supported by MMP20 alone and typically require larger datasets to achieve improved resolution, we employed a composite species tree based on McGowen et al. [38] for relationships within Mysticeti, and Gatesy [39] for relationships among other cetartiodactyl taxa. The branch model of PAML [37] was used to estimate ω values for functional and pseudogenic branches following Meredith et al.’s [16] branch-coding method. Functional branches lead to external nodes (i.e. extant taxa) having enamel or to internal nodes having enamel based on parsimony reconstructions, and are expected to evolve under purifying selection with ω < 1. Pseudogenic branches post-date the first detected occurrence of a frameshift mutation or stop codon on an earlier branch and are expected to evolve at the neutral rate with ω = 0. We used χ²-tests to compare the observed numbers of non-synonymous and synonymous substitutions, which were estimated using PAML [37], with the expected numbers of non-synonymous and synonymous substitutions according to a neutral model of evolution with ω = 1. PAML estimates for the number of non-synonymous and synonymous sites in the MMP20 alignment were 170.8 and 75.2, respectively. Estimated numbers of non-synonymous and synonymous substitutions on functional branches were 30.4 and 70.7, respectively, whereas expected numbers of non-synonymous and synonymous changes were 70.2 and 30.9 for ω = 1. Estimated numbers of non-synonymous and synonymous substitutions on pseudogenic branches were 17.3 and 4.2, respectively, whereas expected numbers of non-synonymous and synonymous changes were 14.9 and 6.6 for ω = 1. dNdS analyses were run with the CodonFreq = 3 option in PAML.

(d) Ancestral sequence reconstructions

The baseprogram of PAML 4.2 [37] was used to estimate ancestral DNA and amino acid sequences.

3. RESULTS AND DISCUSSION

Representative mysticete sequences from exons 3 and 4 of MMP20 lack frameshift mutations and stop codons, but we discovered a CHR-2 SINE insertion in exon 2 that is shared by eight mysticete species that are representative of all extant mysticete genera. CHR SINES were originally described by Shimamura et al. [40] and given the name CHR based on their exclusive occurrence in the genomes of cetaceans (C), hippopotamids (H) and ruminants (R). CHR SINES are divided into two families, CHR-1 and CHR-2, with the latter derived from the former [41]. In the CHR-2 group, Nikaido et al. [42] defined FL (full-length), MDI (middle deletion I), MDII (middle deletion II), DT (deletion type), CD (cetacean deletion) and CDO (cetacean deletion odonto-cete) subfamilies. The CHR-2 SINE in MMP20 belongs to the CD subfamily (figure 1). MMP20 SINE sequences share diagnostic features with members of the CD and CDO subfamilies [42], including a centrally located deletion in the tRNA unrelated region and several diagnostic nucleotide substitutions (figure 1), but lack the terminal deletion that defines the CDO subfamily. The length of the MMP20 SINE ranges from 302 bp (B. musculus) to 318 bp (B. physalus) and includes a tRNA-related region, tRNA-unrelated region, poly-AT region and 14-nucleotide target site duplication at the 3′ end of the SINE (electronic supplementary material, alignment 1). All of the length variation occurs in the poly-AT region.

The preferential insertion of SINEs into introns, rather than exons, reflects selection against the deleterious effects of SINE insertions in protein-coding regions [43]. There are numerous examples of disease-causing SINE insertions in exons [44–46]. The CHR-2 SINE in the MMP20 gene is located in the propeptide-coding region of exon 2, and would result in premature truncation of the MMP20 protein owing to stop codons in all possible reading frames of the CHR-2 SINE. Three mutations in the human MMP20 gene that cause amelogenesis imperfecta have been characterized, and in every case the inheritance pattern is autosomal recessive. One of these mutations encodes a stop signal in the propeptide-coding region of exon 1 [30]. Enamel in the afflicted individual is thin, hypomineralized and chips

Proc. R. Soc. B (2011)
away from the underlying dentin [30]. Disease-causing SINE insertions in exons are sometimes associated with multiple transcripts, including mRNAs in which the SINE-afflicted exon has been spliced out [46]. However, exon 2 of \textit{MMP20} encodes the carboxyl-terminal region of the propeptide, as well as 18 amino-terminal residues of the catalytic (Zn$^{2+}$, Ca$^{2+}$) subdomain, and removal of this exon from the mRNA would presumably render MMP20 non-functional. Indeed, the homologous 18-amino acid region of the catalytic (Zn$^{2+}$, Ca$^{2+}$) subdomain encoded by exon 2 is completely conserved in human, cow, pig and spectacled caiman, and shows only one amino acid difference in mouse [24]. This pattern of evolutionary conservation validates the critical importance of the catalytic subdomain that is partially encoded by exon 2.

The SINE insertion in exon 2 of \textit{MMP20} provides the first molecular evidence for pseudogenization of the genetic toolkit for enamel production in the common ancestor of extant mysticetes (figure 2). Given that MMP20 is required for proper processing of amelogenin, ameloblastin and enamelin, and may also be required for KLK4 activation; this insert provides compelling evidence that normal enamel formation was abrogated no later than the date of this SINE insertion. There is, therefore, congruent genomic and fossil evidence for the loss of enamel-capped teeth prior to the last common ancestor of crown-group mysticetes (figure 2). Approximately 3.9 kb from among the exons of four tooth-specific genes (\textit{AMBN}, \textit{AMEL}, \textit{ENAM} and \textit{MMP20}) have been sequenced for representative mysticetes, but the SINE insertion in \textit{MMP20} is the only example of a shared frameshift mutation (figure 2). However, mysticete pseudogenes have low rates of frameshift accumulation. Previously, Meredith et al. [16] calculated a rate of 0.0081 frameshifts kb$^{-1}$ myr$^{-1}$ for neutrally evolving mysticete DNA. Assuming this rate, and a stem mysticete branch that comprises 7.6 myr of evolutionary history [38], then we should expect only 0.24 shared frameshifts per 3.9 kb for exonic segments that have evolved under neutral evolution.

In addition to the CHR-2 SINE insertion that is shared by all mysticete genera, there is additional evidence for the inactivation of \textit{MMP20} in toothless and enamel-less cetaceans (figure 2). Within Mysticeti, a 1 bp insertion at position 552 is present in the pygmy right whale, \textit{Caperea marginata}, and a G to T point mutation at position 532 in the common minke whale, \textit{Balaenoptera acutorostrata}, results in an ochre (‘TAA’) stop codon (electronic supplementary material, alignment 1).
We also discovered an opal stop codon (‘TGA’) in the propeptide-coding region of MMP20 exon 2 in a single individual of the pygmy sperm whale, *Kogia breviceps* (electronic supplementary material, alignment 1). This species and its congener, *Kogia sima* (dwarf sperm whale), have enamel-less teeth [47]. Previously, three frameshift mutations were reported in the enamelin (*ENAM*) gene of *Kogia*, two in the common ancestor of the two extant species, and the other in *K. sima* [16]. Although available data suggest that the MMP20 gene was incapacitated...
before ENAM in baleen whales, current evidence suggests that ENAM was pseudogenized before MMP20 in Kogia (figure 2). Indeed, the stop codon in K. breviceps was only present in one of three individuals surveyed here, and also was absent in five individuals of K. sima (electronic supplementary material, alignment 1).

An ML phylogram based on exon 2 sequences from MMP20 (figure 3) is broadly congruent with cetartiodactyl trees based on large supermatrices [38,48,49] even though the exon 2 alignment is only 248 bp. Only eight clades were supported above the 70 per cent bootstrap level (figure 3), but in every case these clades were concordant with the analyses based on supermatrices. Within Cetacea, visual inspection of the MMP20 tree confirms that branches leading exclusively to taxa without enamel (mysticetes, Kogia) are longer than branches leading to taxa that retain enamel. dN/dS values were calculated for functional versus pseudogenic branches of Cetartiodactyla following Meredith et al. [16] and in each case tested against the null hypothesis of no selection ($\omega = 1$) using a $\chi^2$-test. Functional branches have a low dN/dS, consistent with strong purifying selection ($\omega = 0.191$, $\chi^2 = 73.82$, $p < 0.001$), whereas pseudogenic branches had dN/dS nearly an order of magnitude higher and are compatible with the absence of selective constraints ($\omega = 1.84$, $\chi^2 = 1.26$, $0.25 < p < 0.50$).

ML analysis of CD subfamily CHR-2 SINE sequences (figure 4) groups the mysticete MMP20 SINEs to the exclusion of several other loci (BRY28, Hump20, IW A31, Sei23) that have been sequenced for multiple mysticetes [50], albeit with weak bootstrap support. Relationships among mysticete MMP20 SINE sequences are generally congruent with trees supported by large concatenations of molecular data (e.g. [38]).

MMP20 is primarily expressed in developing teeth, but expression has been reported in human lung [51,52] and in mouse large intestine [53]. There are also SNPs in MMP20 that are significantly associated with kidney ageing [54]. Nevertheless, MMP20 generally has been considered a tooth-specific gene owing to its primary expression pattern, and the occurrence of non-syndromic amelogenesis imperfecta in mice and humans that lack a functional copy of this gene [53]. Pseudogenization of MMP20 in two enamel-less cetacean lineages, Mysticeti and Kogia breviceps, provides additional evidence for the tooth-specific function of MMP20. Likewise, the genomic assembly for Hoffmann’s two-toed sloth (Choloepus hoffmanni), another species with enamel-less teeth, revealed...
frameshift mutations in exon 1 (8 bp deletion), exon 4 (1 bp deletion), exon 5 (7 bp insertion), exon 6 (1 bp deletion) and exon 9 (2 bp deletion) of \( M\)MP20 (Ensembl 57; electronic supplementary material, figure S1). A polymorphic stop codon in \( K.\) breviceps and multiple frameshift indels in \( C.\) hoffmanni further suggest that \( M\)MP20 is not only tooth-specific, but also enamel-specific; \( K.\) breviceps and \( C.\) hoffmanni both retain dentin in their mineralized, enamel-less teeth. These findings do not invalidate reports of \( M\)MP20 expression in other tissues, but imply that the only critical, unique and non-overlapping role of this gene is in enamel formation.

In other instances, \( M\)MP20 expression may be incidental and entirely overlapping with other components of the transcriptome. Along these lines, the molecular redundancy of \( M\)MP2 and \( M\)MP20 in cleaving DSPP [33] may have permitted pseudogenization of \( M\)MP20 once it was released from performing its unique function in enamel formation.

Mammalian diversity provides a natural laboratory, complete with replicated experiments, for testing hypotheses of tooth-specific gene function. Multiple lineages of enamel-less and edentulous mammals have descended from ancestors with enamel-capped teeth, and we...
Tooth gene evolution in baleen whales

expect to find degraded remnants of enamel-specific genes in these taxa owing to their evolutionary history. Previous work has documented pseudogenization of three genes that code for structural EMPs (enamelin, ameloblastin, amelogenin) in one or more lineages that lack enamel ([2,16]; J. Gatesy 2010, unpublished data). Molecular cavities in the MMP20 gene in three different lineages of enamel-less mammals (Mysticeti, K. breviceps, C. hoffmanni) provide the first evidence for pseudogenization of an enzymatic EMP gene. Further, the insertion of a CHR-2 SINE retroposon in MMP20 shows that the genetic toolkit for enamel production was knocked out in the common ancestor of living mysticetes. The combination of palaeontological and molecular data now provide support for the gain and loss of two complex adaptations, baleen and enamel-capped teeth, respectively, on the stem mysticete branch. Pseudogenization may occur through neutral evolution when changes in the genetic background or environment render a formerly useful gene worthless, or through positive selection when a previously useful gene becomes harmful to an organism [55,56]. It remains unclear if the SINE insertion in MMP20 was favoured by natural selection because it was advantageous to stop enamel production or was simply a consequence of neutral evolution owing to relaxed functional constraints on tooth-specific genes subsequent to the origin of baleen. Beyond the SINE insertion, the only other reconstructed change in exon 2 of MMP20 on the stem mysticete branch is a non-synonymous transition from A to G at nucleotide position 38 that replaced histidine with arginine. This change has a smaller Grantham matrix distance (29) than the reconstruction of an enzymatic EMP gene. Further, the insertion of a CHR-2 SINE retroposon in MMP20 shows that the genetic toolkit for enamel production was knocked out in the common ancestor of living mysticetes. The combination of palaeontological and molecular data now provide support for the gain and loss of two complex adaptations, baleen and enamel-capped teeth, respectively, on the stem mysticete branch. Pseudogenization may occur through neutral evolution when changes in the genetic background or environment render a formerly useful gene worthless, or through positive selection when a previously useful gene becomes harmful to an organism [55,56]. It remains unclear if the SINE insertion in MMP20 was favoured by natural selection because it was advantageous to stop enamel production or was simply a consequence of neutral evolution owing to relaxed functional constraints on tooth-specific genes subsequent to the origin of baleen. Beyond the SINE insertion, the only other reconstructed change in exon 2 of MMP20 on the stem mysticete branch is a non-synonymous transition from A to G at nucleotide position 38 that replaced histidine with arginine. This change has a smaller Grantham matrix distance (29) than the reconstructed change from histidine to proline (77) at the same amino acid position in ruminants, and is an unlikely candidate for adaptive loss of protein function. Whether adaptive or neutral, the SINE insertion in MMP20 fills an important gap in our understanding of the macroevolutionary transition leading from the last common ancestor of crown Cetacea to the last common ancestor of crown Mysticeti.

This work was supported by NSF (EF0629860 to M.S.S. and J.G.; DEB0743724 to J.G.). For providing DNA samples, we thank Southwest Fisheries Science Center, South Australian Museum, North Slope Borough (Barrow, Alaska), The Marine Mammal Center (Sausalito), Smithsonian Institution, New York Zoological Society, World Wildlife Fund, Greenland Institute of Natural Resources, Alaska Department of Fish and Game, P. Morin, K. Robertson, S. Chivers, A. Dizon, M. Milinkovitch, G. Amato, G. Schaller, M. Cronin, W. Murphy, M. P. Heide-Jørgensen, U. Arnason, H. Rosenbaum and G. Braulik. C. Buell painted living and extinct mammals. Two anonymous reviewers provided helpful comments on an earlier version of this manuscript.

REFERENCES


13 Karlsen, K. 1962 Development of tooth germs and adjacent structures in the whalebone whale (Balaenoptera physalus (L.)). Hvalrædets Skrifter 45, 5–56.


Tooth gene evolution in baleen whales  R. W. Meredith et al. 1001


Nikaido, M., Piskurek, O. & Okada, N. 2006 Toothed whale monophyly reassessed by SINE insertion analysis: the


